BMC Bioinformatics

SOFTWARE

Open Access

GeCoNet-Tool: a software package for gene co-expression network construction and analysis

Junyao Kuang[1*], Kristin Michel[2] and Caterina Scoglio[1]

*Correspondence:
kuang@ksu.edu

[1] Department of Electrical and Computer Engineering, Kansas State University, Manhattan, KS 66506, USA
[2] Division of Biology, Kansas State University, Manhattan, KS 66506, USA

Abstract

Background: Network analysis is a powerful tool for studying gene regulation and identifying biological processes associated with gene function. However, constructing gene co-expression networks can be a challenging task, particularly when dealing with a large number of missing values.

Results: We introduce GeCoNet-Tool, an integrated gene co-expression network construction and analysis tool. The tool comprises two main parts: network construction and network analysis. In the network construction part, GeCoNet-Tool offers users various options for processing gene co-expression data derived from diverse technologies. The output of the tool is an edge list with the option of weights associated with each link. In network analysis part, the user can produce a table that includes several network properties such as communities, cores, and centrality measures. With GeCoNet-Tool, users can explore and gain insights into the complex interactions between genes.

Keywords: co-expression network, Missing value, Pearson correlation

Background

A gene co-expression network is helpful for analyzing and predicting gene functions and regulations [1, 2]. A gene co-expression network is composed of nodes and edges, in which nodes represent genes and edges represent co-expressed gene pairs [3, 4]. In the past decade, high throughput technologies (such as single-cell RNA sequencing) have enabled biologists to measure gene expression levels under various conditions [5, 6]. To study the relationships between genes, some researchers employ dimension-reduction algorithms such as PCA, UMAP, and t-SNE [7–10] to visualize genes in 2D or 3D space. However, analyzing co-expression data that encompasses diverse conditions can be challenging, particularly when missing values are present [6, 11–40].

In our previous work [1], we developed a data processing scheme to construct a gene co-expression network for *Anopheles gambiae*. The experimental results demonstrated that the proposed approach is effective in studying gene functions and patterns, even when dealing with different experimental technologies and many missing values.

Building upon this prior work, we developed an integrated tool–GeCoNet-Tool for gene co-expression network construction and analysis.

Implementation

GeCoNet-Tool is an open-source package that combines gene co-expression network construction and network analysis. This tool is an implementation and improvement of our previous work [1], which presented a scheme for studying gene expression data across a large number of conditions. GeCoNet-Tool is an executable file written in Python with a user-friendly graphical interface that facilitates configuration for different data types (as shown in Fig. 1). The process of using GeCoNet-Tool can be split into two independent parts: network construction and network analysis.

Data processing and network generation

To construct a gene co-expression network, the user needs to input a gene co-expression matrix (.csv format), in which rows represent N genes and columns represent M experimental conditions. GeCoNet-Tool allows users to process the input data with different options, depending on the data type. For example, users can choose to remove zeros, re-scale expression values by log2, or normalize columns by z-score if the input data are obtained through RNA-seq [41]. Additionally, users can choose to save the processed data in table format.

GeCoNet-Tool calculates the Pearson Correlation Coefficient (PCC) between each pair of genes based on the processed data. The PCC matrix is saved as an upper triangular matrix if the user chooses to save the PCC matrix [42, 43]. In our previous work [1], we observed that the number of experimental conditions could significantly affect the PCC between two genes. Therefore, GeCoNet-Tool determines the number of paired elements between every pair of genes, and the user can save this data as an additional upper triangular matrix.

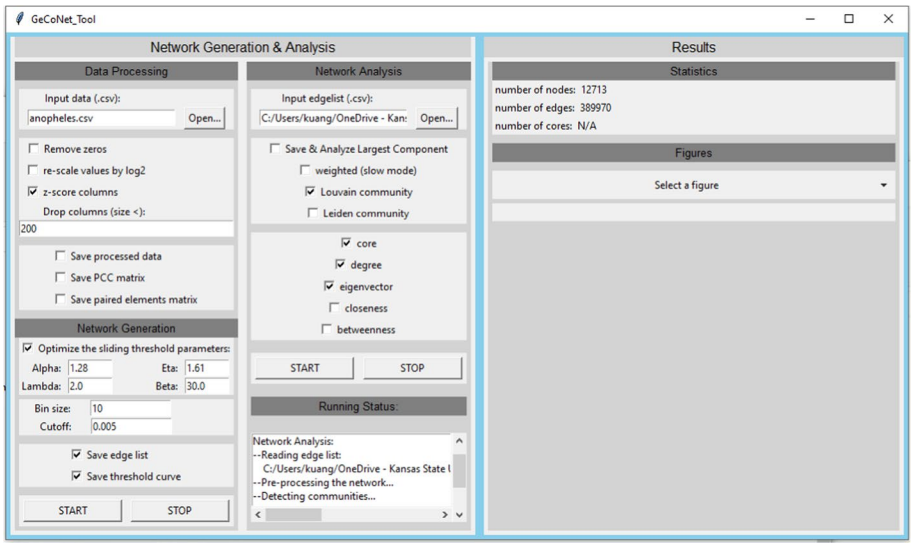

Fig. 1 The GeCoNet_Tool user interface

GeCoNet-Tool classifies the PCCs into different intervals based on the number of paired conditions of gene pairs. The user can specify the size of intervals (*Bin size*) in the GeCoNet-Tool interface. To select edges based on PCC value, the user also needs to input the cutoff value expressed as the chosen top percentage of all PCCs in a given interval (e.g., 0.005, 0.01, or 0.02), which is used to determine the sliding threshold by fitting the following curve:

$$f^{thres}(x) = \alpha - \frac{1}{\eta + \lambda e^{-\frac{x}{\beta}}}, \tag{1}$$

where α, η, λ, and β are the four parameters that were fitted, and x is the number of paired elements. This equation provided a good trade-off between the accuracy of the fitting and the number of parameters to estimate [1]. Once the curve is fitted, the optimal parameters will be updated to the input boxes α, η, λ, and β.

In order to obtain optimized parameters, GeCoNet-Tool automates the optimization of the four parameters instead of manual optimization as in our previous work [1]. The coefficient of determination (R-squared) of the fitted curve is shown in the *Running Status* box. The edges of the co-expression network are selected through the fitted curve based on the number of paired elements [1]. The user can construct networks with different edge densities by tuning the cutoff value.

We recommend using a cutoff value that can maintain the majority of the nodes connected while minimizing the number of edges. In general, increasing the cutoff value, decreases the edge density, which can impact the connectivity of nodes in a network. It is therefore advisable to choose a cutoff value that maintains the majority of connected nodes. To achieve this, the package should be executed multiple times using different cutoff values. This allows the user to observe the number of nodes and edges in each resulting network and select the cut-off value that contains the majority of nodes, while minimizing the number of edges. This approach ensures that the network retains its integrity while avoiding an excessive number of edges, which can impact downstream analyses.

Finally, users can save the list of edges in the co-expression network, along with the fitted threshold curve, for further analysis. The edge list and threshold curve are saved in the same folder as the input data.

Network analysis

Once a network is constructed, various properties of the network can be analyzed through the second part of GeCoNet-Tool. GeCoNet-Tool allows users to produce the following network properties [44]:

- community: The community is defined as a subgraph that is highly connected internally and loosely connected to other subgraphs. GeCoNet-Tool allows users to detect communities through the Louvain and Leiden algorithms [45–47]. Users can customize the settings for community analysis by editing the original code (network_analysis.py), while the tool provides default settings for users who prefer to use them.
- core: The core of the network is obtained by repeatedly removing nodes with a degree less than k by starting with k = 1 and increasing k until no nodes are left in

the network. The core genes are those with a degree = k (i.e., those removed during the last iteration) [48].

- degree: GeCoNet-Tool calculates node degree if the network is unweighted and calculates node strength if the network is weighted [44].
- eigenvector: GeCoNet-Tool calculates the eigenvector centrality, which is determined by the entry of the eigenvector corresponding to the largest eigenvalue of the adjacency matrix of the network [49].
- betweenness: GeCoNet-Tool calculates the betweenness centrality, determined by the number of shortest paths that pass through the node itself [50].
- closeness: GeCoNet-Tool calculates the betweenness centrality, which is based on the distances between nodes. Closeness centrality is the sum of the shortest path distance reciprocals of a node to all other nodes [51].

In the package, users can choose to analyze either the entire network or only the largest connected component and use either unweighted or weighted edges. GeCoNet-Tool generates a table in the output that contains all the selected properties. In addition, the package creates figures of the node degree distribution, community distribution, and core distribution.

Results

The user can observe the running status of the GeCoNet-Tool through the *Running Status* window. At the same time, network statistics (such as the number of nodes, edges, and core nodes will be shown in the *Results* window as soon as the network is generated or analyzed.

In the experiment, we provide the *Anopheles gambiae* gene expression data and generate a network with default settings (the data is publicly available through VectorBase (www.vectorbase.org) at the following URL: https://tinyurl.com/mr38a7hj). Figure 2a shows the sliding threshold with a cutoff value of 0.005, and the R-squared is 0.908, which suggests that the curve fits the raw data well. In practical applications, we suggest using smaller bin sizes and testing various cutoff values to generate a network that connects the majority of nodes while minimizing the number of edges.

In the second part, users can analyze the generated network and produce a table to store the properties of the network. In the *Anopheles gambiae* gene co-expression network, there are 12660 nodes, 389991 edges, and 164 core nodes. GeCoNet-Tool employs the force-directed algorithm Fruchterman-Reingold layout to visualize the community distribution (Fig. 2b) and core nodes (Fig. 2c). However, the layout shown in the results window is deterministic. Interested users are recommended to use interactive network visualization algorithms in Gephi [53] to show the generated network and properties. GeCoNet-Tool also generates node degree distribution as shown in Fig. 2d.

Conclusion and future works

GeCoNet-Tool is a free and user-friendly research tool that offers a straightforward approach to network construction and analysis, without the need for coding expertise. The package is composed of two parts: (1) network construction and (2) network analysis. In the first part, pairwise relationships between nodes are evaluated using

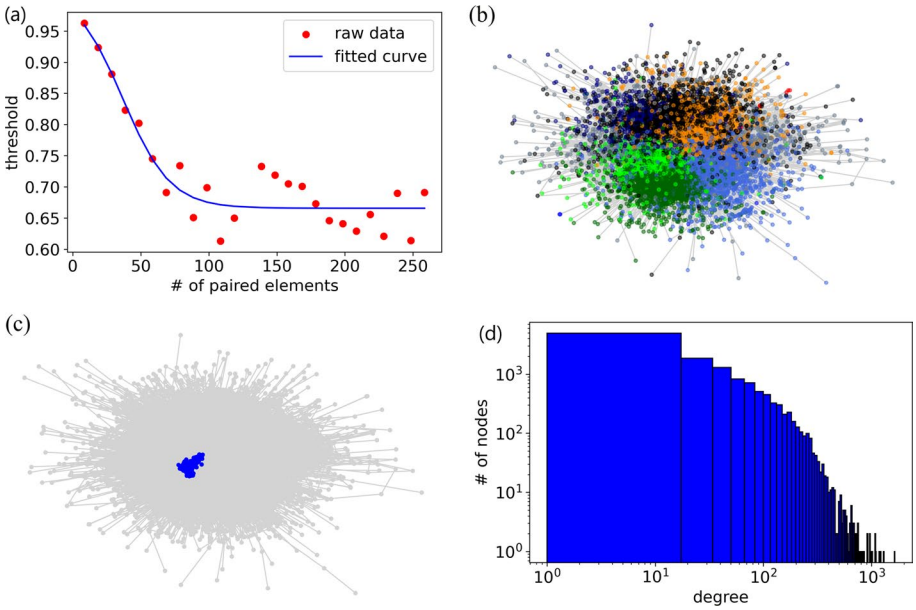

Fig. 2 Examples of figure outputs from GeCoNet-Tool using *Anopheles gambiae* expression data. **a** The fitted threshold curve, **b** communities, **c** core nodes, and **d** degree distribution of the *An. gambiae* gene co-expression network

the PCC and the number of paired conditions. Users can choose from various expression data types and data processing options, such as removing zeros and log2-rescaling. In the second part, GeCoNet-Tool provides multiple tools for network analysis. For example, the community analysis will classify nodes into different communities to identify genes with similar biological functions.

In the *GeCoNet-Tool*, networks are currently constructed with the Pearson correlation coefficient. However, future updates may be expanded to other methods to assess gene co-expression, e.g., signed distance correlation, Spearman correlation, and mutual information, as suggested by recent studies [54]. This approach would offer users more flexibility in constructing co-expression networks that are tailored to their specific research requirements.

Acknowledgements
We thank Dr. Robert MacCallum, Imperial College London, UK for initial advice and sharing of the data set Anopheles-gambiae EXPR-STATS VB-2019-02, which is be publicly available through VectorBase (www.vectorbase.org) at the following URL: https://tinyurl.com/mr38a7hj.

Author contributions
J.K., K.M. and C.S. conceived the tool. J.K. wrote the main manuscript text and code. K.M. and C.S. revised the manuscript. All authors defined research directions, proposed problem solutions, and reviewed the manuscript.

Funding
This work has been supported by the National Institutes of Health under Grant No. R01AI140760 (to K. M.). The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the funding agencies.

Availibility of data and materials
Project name: GeCoNet-Tool. Project home page:https://github.com/KSUNetSE/GeCoNet-Tool. Operating system(s): Windows 10. Programming language: Python. License: Redistribution and use in source and binary forms, with or without modification, are permitted. The resources and data used during the current study are available in the GitHub repository, https://github.com/KSUNetSE/GeCoNet-Tool

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 7 December 2022 Accepted: 9 June 2023
Published online: 11 July 2023

References

1. Kuang J, Buchon N, Michel K, Scoglio C. A global Anopheles gambiae gene co-expression network constructed from hundreds of experimental conditions with missing values. BMC Bioinf. 2022;23:170.
2. Song L, Langfelder P, Horvath S. Comparison of co-expression measures: mutual information, correlation, and model-based indices. BMC Bioinf. 2012;13:328.
3. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Gene Mole Biol. 2005;4:1.
4. Oldham MC, Horvath S, Geschwind DH. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proc Natl Acad Sci. 2006;103(47):17973–8.
5. Horvath S, Dong J. Geometric interpretation of gene coexpression network analysis. PLoS Comput Biol. 2008;4(8): e1000117.
6. MacCallum RM, Redmond SN, Christophides GK. An expression map for Anopheles gambiae. BMC Genomics. 2011;12:1–16.
7. Abdi H, Williams LJ. Principal component analysis. Wiley interdiscip Rev Comput Stat. 2010;2(4):433–59.
8. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426. 2018.
9. Van Der Maaten L. Learning a parametric embedding by preserving local structure. In: Artificial intelligence and statistics, 2009:384-391, PMLR.
10. Kuang J, Scoglio C, Michel K. Feature learning and network structure from noisy node activity data. Physical Rev E. 2022;106: 064301.
11. Seaman JA, Alout H, Meyers JI, Stenglein MD, Dabiré RK, Lozano-Fuentes S, Burton TA, Kuklinski WS, Black WC, Foy BD. Age and prior blood feeding of Anopheles gambiae influences their susceptibility and gene expression patterns to ivermectin-containing blood meals. BMC Genomics. 2015;16:1–18.
12. Koutsos AC, Blass C, Meister S, Schmidt S, MacCallum RM, Soares MB, Collins FH, Benes V, Zdobnov E, Kafatos FC, Christophides GK. Life cycle transcriptome of the malaria mosquito Anopheles gambiae and comparison with the fruitfly Drosophila melanogaster. Proc Natl Acad Sci. 2007;104:11304–9.
13. Marinotti O, Calvo E, Nguyen QK, Dissanayake S, Ribeiro JMC, James AA. Genome-wide analysis of gene expression in adult Anopheles gambiae. Insect Mol Biol. 2007;15:1–12.
14. Cassone BJ, Mouline K, Hahn MW, White BJ, Pombi M, Simard F, Costantini C, Besansky NJ. Differential gene expression in incipient species of Anopheles gambiae. Mol Ecol. 2008;17:2491–504.
15. Goltsev Y, Rezende GL, Vranizan K, Lanzaro G, Valle D, Levine M. Developmental and evolutionary basis for drought tolerance of the Anopheles gambiae embryo. Dev Biol. 2009;330:462–70.
16. Mendes AM, Awono-Ambene PH, Nsango SE, Cohuet A, Fontenille D, Kafatos FC, Christophides GK, Morlais I, Vlachou D. Infection intensity-dependent responses of Anopheles gambiae to the African malaria parasite Plasmodium falciparum. Infect Immun. 2011;79:4708–15.
17. Cassone BJ, Molloy MJ, Cheng C, Tan JC, Hahn MW, Besansky NJ. Divergent transcriptional response to thermal stress by Anopheles gambiae larvae carrying alternative arrangements of inversion 2La. Mol Ecol. 2011;20:2567–80.
18. Baker DA, Nolan T, Fischer B, Pinder A, Crisanti A, Russell S. A comprehensive gene expression atlas of sex-and tissue-specificity in the malaria vector, Anopheles gambiae. BMC Genomics. 2011;12:296.
19. Rund SSC, Hou TY, Ward SM, Collins FH, Duffield GE. Genome-wide profiling of diel and circadian gene expression in the malaria vector Anopheles gambiae. Proc Natl Acad Sci. 2011;108:E421–30.
20. Cook PE, Sinkins SP. Transcriptional profiling of Anopheles gambiae mosquitoes for adult age estimation. Insect Mol Biol. 2010;19:745–51.
21. Wang MH, Marinotti O, Vardo-Zalik A, Boparai R, Yan G. Genome-wide transcriptional analysis of genes associated with acute desiccation stress in Anopheles gambiae. PLoS ONE. 2011;6: e26011.
22. Vlachou D, Schlegelmilch T, Christophides GK, Kafatos FC. Functional genomic analysis of midgut epithelial responses in Anopheles during Plasmodium invasion. Curr Biol. 2005;15:1185–95.
23. Abrantes P, Dimopoulos G, Grosso AR, Do Rosário VE, Silveira H. Chloroquine mediated modulation of Anopheles gambiae gene expression. PLoS ONE. 2008;3: e2587.
24. Oviedo MN, Ribeiro JMC, Heyland A, VanEkeris L, Moroz T, Linser PJ. The salivary transcriptome of Anopheles gambiae (Diptera: Culicidae) larvae: a microarray-based analysis. Insect Biochem Mol Biol. 2009;39:382–94.
25. Oviedo MN, Vanekeris L, Corena-Mcleod MDP, Linser PJ. A microarray-based analysis of transcriptional compartmentalization in the alimentary canal of Anopheles gambiae (Diptera: Culicidae) larvae. Insect Mol Biol. 2008;17:61–72.
26. Rogers DW, Whitten MM, Thailayil J, Soichot J, Levashina EA, Catteruccia F. Molecular and cellular components of the mating machinery in Anopheles gambiae females. Proc Natl Acad Sci. 2008;105:19390–5.

27. Pinto SB, Lombardo F, Koutsos AC, Waterhouse RM, McKay K, An C, Ramakrishnan C, Kafatos FC, Michel K. Discovery of Plasmodium modulators by genome-wide analysis of circulating hemocytes in Anopheles gambiae. Proc Natl Acad Sci. 2009;106:21270–5.
28. Zhao YO, Kurscheid S, Zhang Y, Liu L, Zhang L, Loeliger K, Fikrig E. Enhanced survival of Plasmodium-infected mosquitoes during starvation. PLoS ONE. 2012;7: e40556.
29. Shaw WR, Teodori E, Mitchell SN, Baldini F, Gabrieli P, Rogers DW, Catteruccia F. Mating activates the heme peroxidase HPX15 in the sperm storage organ to ensure fertility in Anopheles gambiae. Proc Natl Acad Sci. 2014;111:5854–9.
30. Gabrieli P, Kakani EG, Mitchell SN, Mameli E, Want EJ, Anton AM, Serrao A, Baldini F, Catteruccia F. Sexual transfer of the steroid hormone 20E induces the postmating switch in Anopheles gambiae. Proceed Natl Acad Sci. 2014;11:16353–8.
31. Kwiatkowska RM, Platt N, Poupardin R, Irving H, Dabire RK, Mitchell S, Jones CM, Diabaté A, Ranson H, Wondji CS. Dissecting the mechanisms responsible for the multiple insecticide resistance phenotype in Anopheles gambiae ss, M form, from Vallee du Kou, Burkina Faso. Gene. 2013;519:98–106.
32. Tene BF, Poupardin R, Costantini C, Awono-Ambene P, Wondji CS, Ranson H, Antonio-Nkondjio C. Resistance to DDT in an urban setting: common mechanisms implicated in both M and S forms of Anopheles gambiae in the city of Yaoundé Cameroon. PLoS ONE. 2013;8: e61408.
33. Wilding CS, Weetman D, Rippon EJ, Steen K, Mawejje HD, Barsukov I, Donnelly MJ. Parallel evolution or purifying selection, not introgression, explains similarity in the pyrethroid detoxification linked GSTE4 of Anopheles gambiae and An, Arabiensis. Molecular genetics and genomics. 2015;290:201–15.
34. Magnusson K, Mendes AM, Windbichler N, Papathanos PA, Nolan T, Dottorini T, Rizzi E, Christophides GK, Crisanti A. Transcription regulation of sex-biased genes during ontogeny in the malaria vector Anopheles gambiae. PLoS ONE. 2011;6: e21572.
35. Isaacs AT, Mawejje HD, Tomlinson S, Rigden DJ, Donnelly MJ. Genome-wide transcriptional analyses in Anopheles mosquitoes reveal an unexpected association between salivary gland gene expression and insecticide resistance. BMC Genomics. 2018;19:1–12.
36. Vannini L, Dunn WA, Reed TW, Willis JH. Changes in transcript abundance for cuticular proteins and other genes three hours after a blood meal in Anopheles gambiae. Insect Biochem Mol Biol. 2014;44:33–43.
37. Mead EA, Li M, Tu Z, Zhu J. Translational regulation of Anopheles gambiae mRNAs in the midgut during Plasmodium falciparum infection. BMC Genomics. 2012;13:1–10.
38. Papa F, Windbichler N, Waterhouse RM, Cagnetti A, D'Amato R, Persampieri T, Lawniczak MK, Nolan T, Papathanos PA. Rapid evolution of female-biased genes among four species of Anopheles malaria mosquitoes. Genome Res. 2017;27:1536–48.
39. Emami SN, Lindberg BG, Hua S, Hill SR, Mozuraitis R, Lehmann P, Birgersson G, Borg-Karlson AK, Ignell R, Faye I. A key malaria metabolite modulates vector blood seeking, feeding, and susceptibility to infection. Science. 2017;355:1076–80.
40. AVCL consortium NCBI BioProject ID 238805. Broad Institute: Umbrella Comparative genomics project (Subtype:Comparative genomics). https://www.ncbi.nlm.nih.gov/bioproject/238805, 2014.
41. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17(1):1–19.
42. Gu Y, Zu J, Li Y. A novel evolutionary model for constructing gene coexpression networks with comprehensive features. BMC Bioinf. 2019;20:1–20.
43. de Anda-Jáuregui G, Alcalá-Corona SA, Espinal-Enríquez J, Hernández-Lemus E. Functional and transcriptional connectivity of communities in breast cancer co-expression networks. Appl Network Sci. 2019;4:1–13.
44. Newman M. Networks. Oxford: Oxford University Press; 2018.
45. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008;10:P10008.
46. Traag VA, Waltman L, Van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019;9(1):5233.
47. Reichardt J, Bornholdt S. Statistical mechanics of community detection. Phys Rev E. 2006;74(1): 016110.
48. Eidsaa M, Almaas E. s-core network decompo-sition: a generalization of k-core analysis toweighted networks. Phys Rev E. 2013;88: 062819.
49. Bonacich P. Some unique properties of eigenvector centrality. Soc Networks. 2007;29(4):555–64.
50. Brandes U. A faster algorithm for betweenness centrality. J Math Sociol. 2001;25(2):163–77.
51. Rochat Y. Closeness centrality extended to unconnected graphs: The harmonic centrality index. CONF: No; 2009.
52. Fruchterman TM, Reingold EM. Graph drawing by force-directed placement. Software: Practice and experience, 1991;21:11.
53. Jacomy M, Venturini T, Heymann S, Bastian M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS ONE. 2014;9: e98679.
54. Pardo-Diaz J, Poole PS, Beguerisse-Díaz M, Deane CM, Reinert G. Generating weighted and thresholded gene coexpression networks using signed distance correlation. Netw Sci. 2022;10(2):131–45.

Publisher's Note