SCIENCE CHINA
Information Sciences

• LETTER •

August 2024, Vol. 67, Iss. 8, 189103:1–189103:2
https://doi.org/10.1007/s11432-024-4100-7

Multiple types of disease-associated RNAs identification for disease prognosis and therapy using heterogeneous graph learning

Wenxiang ZHANG[1†], Hang WEI[2†], Wenjing ZHANG[3,4], Hao WU[1*] & Bin LIU[1,5*]

[1]*School of Computer Science and Technology, Beijing Institute of Technology, Beijing* 100081*, China;*
[2]*School of Computer Science and Technology, Xidian University, Xi'an* 710071*, China;*
[3]*Department of Teaching and Research, Shenzhen University General Hospital, Shenzhen* 518055*, China;*
[4]*Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging,*
National-Regional Key Technology Engineering Laboratory for Medical Ultrasound,
School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen 518055*, China;*
[5]*Advanced Research Institute of Multidisciplinary Science, Beijing Institute of Technology, Beijing* 100081*, China*

Received 16 February 2024/Revised 15 April 2024/Accepted 3 June 2024/Published online 24 July 2024

Identifying disease-associated RNAs is crucial in revealing the pathogenic mechanisms of diseases [1], and biologists have made notable progress in this field [2]. However, more effective computational methods are needed to provide reference disease-associated RNAs, reducing the manpower and material resources required for biological experiments.

With the powerful ability of graph neural networks in detecting association patterns [3], several graph-learning-based methods are proposed to identify disease-associated RNAs. For example, HGC-GAN integrated the strengths of heterogeneous graph convolutional neural network and generative adversarial network to predict candidate disease-associated lncRNAs [4]. DMFNet designed a graph autoencoder with variational gate mechanisms to extract multilevel representations of various bio-entities, enabling the inference of potential disease-associated miRNAs [5].

Despite the advancements in the identification of disease-associated RNAs, these computational methods mainly suffer from two limitations: (i) They tend to focus on identifying a single type of disease-associated RNA. For example, methods designed for identifying disease-associated miRNAs may not be suitable for identifying other types of disease-associated RNAs. (ii) While the primary objective of identifying disease-associated RNAs is to explore underlying pathogenic mechanisms and facilitate treatment, the identified disease-associated RNAs are often not subjected to further analysis for downstream disease prognosis and therapy considerations.

Methods. To address these limitations, we propose iRNADis-PT, a novel computational framework that not only identifies multiple types of disease-associated RNAs via heterogeneous graph neural network, but also applies these predicted RNAs for disease prognosis and therapy analysis. As illustrated in Figure 1, iRNADis-PT mainly consists of three important components: heterogeneous network construction, identification of multiple types of disease-associated RNAs, and application of predicted disease-associated RNAs for disease prognosis and therapy analysis.

There are three steps to construct a heterogeneous network: (i) RNA features are extracted via copy number variation analysis and differential expression analysis, and then similarities are calculated to construct four different RNA similarity networks. (ii) Disease similarity network is constructed based on their ontology semantics. (iii) Multiple RNA interactions and experimentally validated RNA-disease associations (Table A1) are collected to link the different bio-entities within the above RNA and disease similarity networks for constructing a heterogeneous network.

For the heterogeneous graph learning component, GraphSAGE is employed as a base graph neural network due to its effective representation learning capability for large-scale graph data. GraphSAGE is utilized to extract node embeddings for different types of sub-networks within the constructed heterogeneous network. Node features are then updated and aggregated by heterogeneous graph learning. The features of RNA-disease pairs can be obtained by concatenating node features of RNAs and diseases, which are fed into the multi-layer perceptron (MLP) to predict their association scores. The top candidate disease-associated RNAs are identified by ranking predicted association scores in descending order.

Considering the alarming mortality rate about hepatocellular carcinoma (HCC), we select HCC as an example, and design a pipeline for downstream applications. The candidate disease-associated RNAs identified by the heterogeneous graph learning are utilized for subsequent disease prognosis and therapy analysis. This analysis includes tasks such as differential expression analysis, risk group classification, and drug sensitivity analysis. It is noteworthy that the above pipeline can also be applied to explore the pathogenic mechanism of other important diseases. Detailed information about our method is provided in Appendix A.

* Corresponding author (email: Hwu@bliulab.net, bliu@bliulab.net)
† Zhang W X and Wei H have the same contribution to this work.

Results. Various RNAs interact and participate in the onset and progression of diseases. We construct four sub-networks involving associations between diseases and different single types of RNAs to analyze the importance of RNA interactions. As illustrated in Figures C1(a) and (b), iRNADis-PT, utilizing a heterogeneous network with RNA interactions, consistently outperforms other sub-network-based methods in terms of NDCG, AUC, and AUPR (detailed information on training process and metrics is provided in Appendix B). Additionally, we compare the performance of iRNADis-PT with the other three baseline methods on four identification tasks (including identification of disease-associated lncRNAs, snoRNAs, miRNAs, and mRNAs). Figure C1(c) illustrates the superior performance of iRNADis-PT in all identification tasks, highlighting its versatility and effectiveness in identifying multiple types of disease-associated RNAs. Further details can be found in Appendix C.

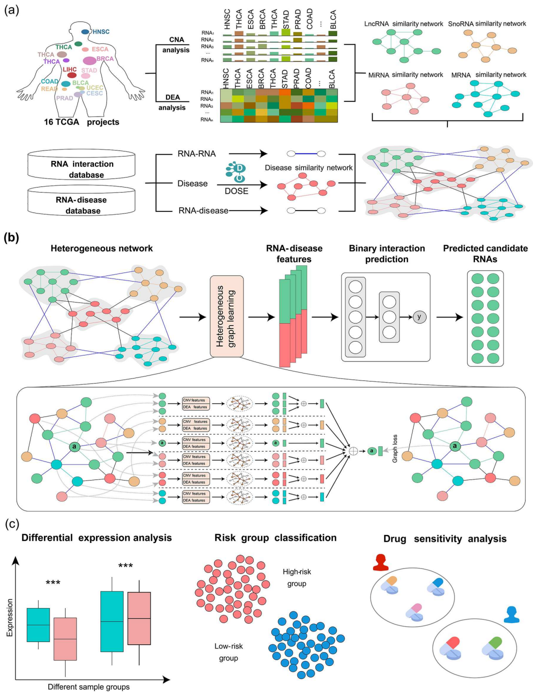

Figure 1 (Color online) Overall framework of the iRNADis-PT. (a) Workflow of data collection and heterogeneous network construction; (b) identification of multiple types of disease-associated RNAs based on heterogeneous graph learning; (c) downstream applications for disease prognosis and therapy.

HCC is selected to investigate how the identified disease-associated RNAs contribute to the analysis of specific disease prognosis and therapy. We reveal 220 significantly differentially expressed RNAs, along with their copy number aberrations and 4 gene ontology (GO) terms related to biological processes, cellular components, and molecular functions. Furthermore, a protein-protein interaction sub-network is deduced, and its significant correlation with HCC has been confirmed through literature mining. Further details can be found in Appendix C.

Based on the expression of identified differentially expressed RNAs, we divide the HCC patients into low- and high-risk groups. The Kaplan-Meier curve and genetic mutation analysis indicate the effectiveness of risk grouping, where patients in the low-risk group have a better survival advantage than those in the high-risk group. Furthermore, we establish a nomogram to provide clinical prognosis prediction for HCC patients, incorporating several clinicopathological features (including age, stage, and gender). Detailed information on this analysis can be found in Appendix C.

To provide potential prognostic and clinical therapy guidance for HCC, immune checkpoints, drug sensitivity, and tumor micro-environment of low- and high-risk groups are explored. We identify 12 immune checkpoint RNAs with significantly differential expression and 4 drugs with noticeable variations in sensitivity between the two risk groups. Among these drugs, three have been proven effective in the treatment of HCC. Additionally, we discover higher immune cell content and stromal component abundance, but lower tumor purity in the low-risk group compared to the high-risk group. Detailed information on this analysis can be found in Appendix C.

Discussion. In this study, we present a computational framework named iRNADis-PT for identifying multiple types of disease-associated RNAs via heterogeneous graph learning. Compared to other methods, iRNADis-PT makes two major contributions: (i) It can work well for identifying multiple types of disease-associated RNAs. (ii) The disease-associated RNAs identified by iRNADis-PT are applied for downstream disease prognosis and therapy analysis tasks. Using HCC as an example, we conduct differential expression analysis to extract high-quality HCC-associated RNAs, and perform risk group classification and drug sensitivity analysis to provide references for the clinical prognosis and therapy of HCC. Overall, iRNADis-PT shows promising predictive performance in identifying multiple types of disease-associated RNAs, and can also serve as a valuable tool for disease prognosis and therapy analysis.

There is still room for improvement in the study. For instance, integrating other advanced graph representation methods like graph attention network (GAT) into the heterogeneous graph learning framework can enhance predictive performance. Additionally, expanding the exploration of downstream prognosis and treatment strategies to encompass various important diseases is warranted.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62325202, 62372041, U22A2039).

Supporting information Appendixes A–C, Table A1, and Figure C1. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Li Q, Gloudemans M J, Geisinger J M, et al. RNA editing underlies genetic risk of common inflammatory diseases. Nature, 2022, 608: 569–577

2 Nemeth K, Bayraktar R, Ferracin M, et al. Non-coding RNAs in disease: from mechanisms to therapeutics. Nat Rev Genet, 2024, 25: 211–232

3 Wu Z, Pan S, Chen F, et al. A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst, 2020, 32: 4–24

4 Lu Z, Zhong H, Tang L, et al. Predicting lncRNA-disease associations based on heterogeneous graph convolutional generative adversarial network. PLoS Comput Biol, 2023, 19: e1011634

5 Guo Y, Zhou D, Ruan X, et al. Variational gated autoencoder-based feature extraction model for inferring disease-miRNA associations based on multiview features. Neural Networks, 2023, 165: 491–505