**Contrastive self-supervised graph convolutional network for detecting the relationship among lncRNAs, miRNAs, and diseases**

Nan Sheng1, Lan Huang1,\*, Yan Wang1,2,\*, Ling Gao1, Yangkun Cao2, Xuping Xie1

*1* *Key laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun 130012, China.*

*2* *School of Artificial Intelligence, Jilin University, Changchun, 130012, China*

\* Corresponding authors.

E-mail: huanglan@jlu.edu.cn (Huang L), wy6868@jlu.edu.cn (Wang Y)

## 1 Integrated lncRNA, miRNA and disease similarity calculations

(1)

(2)

(3)

Firstly, lncRNA functional similarity , miRNA functional similarity , and disease semantic similarity were calculated according to the literature [1, 2], respectively. Then, based on the assumption that functionally similar lncRNA (miRNA) may be correlated with phenotypically similar diseases, and vice versa. The first types of lncRNA GIPK similarity , miRNA GIPK similarity , and two types of disease GIPK similarity , can be measured based on the lncRNA-disease, miRNA-disease network. Finally, two lncRNAs are more likely to be similar if they interact with more identical miRNAs, and vice versa. The second types of lncRNA and miRNA GIPK similarities , are calculated from the lncRNA-miRNA network. Considering that all disease semantic similarity values are greater than 0, and the default semantic features are most significant, the integrated disease similarity is designed as in equation 3.

## 2 Ablation study

In this section, we design three variants of CSGLMD to better understand the contribution of each component to the performance. These variants only change one component, the other components are the same as original approach. The differences between the three variants and CSGLMD are briefly described as follows:

**CSGLMD without label initialization (CSGLMD\_noLI):** The integration similarities of lncRNA, miRNA, and disease are represented using original similarity values without label initialization, which are utilized to construct the LMDHG with weights.

**CSGLMD without self-supervised learning task (CSGLMD\_noSSL):** The graph contrastive learning framework is removed, and only using supervised learning task to train the model.

**CSGLMD without supervised learning task (CSGLMD\_noSL):** The supervised learning task is removed and only training based on contrastive self-supervised learning objective is available. Then, the obtained node representations will be exploited to predict lncRNA-disease pairs, miRNA-disease pairs, and lncRNA-miRNA pair scores.

We conduct ablation study for three tasks on dataset 1, and **Supplementary Table 1 (ST1)** summarizes the experimental results for all variants and CSGLMD. It is clear that CSGLMD with similarity-valued label instantiation and self-supervised learning tasks can achieve excellent performance, while removing of either them would weaken the prediction power. A comparison of CSGLMD, CSGLMD\_noLI, and CSGLMD\_noSL shows that the removal of similarity-valued label instantiation and supervised learning task have a significant impact on AUC and AUPR values. Furthermore, it can be seen that CSGLMD outperforms the variant CSGLMD\_noSSL, demonstrating the usefulness and necessity of the graph contrastive learning module for building self-supervised task components. Overall, both similarity-valued label instantiation, and self-supervised learning, supervised learning are efficient in improving the performance of CSGLMD.

**Supplementary Table 1 (ST1). Performance of different variants of CSGLMD on dataset 1.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | LDA | | | | MDA | | | | LMI | | | |
| 5-cv1 | | 5-cv2 | | 5-cv1 | | 5-cv2 | | 5-cv1 | | 5-cv2 | |
| AUC | AUPR | AUC | AUPR | AUC | AUPR | AUC | AUPR | AUC | AUPR | AUC | AUPR |
| CSGLMD\_noLI | 0.942 | 0.944 | 0.937 | 0.298 | 0.952 | 0.950 | 0.951 | 0.435 | 0.963 | 0.956 | 0.966 | 0.105 |
| CSGLMD\_noSSL | 0.950 | 0.945 | 0.941 | 0.400 | 0.956 | 0.950 | 0.955 | 0.451 | 0.968 | 0.964 | 0.973 | 0.169 |
| CSGLMD\_noSL | 0.941 | 0.940 | 0.944 | 0.355 | 0.951 | 0.947 | 0.950 | 0.439 | 0.952 | 0.950 | 0.953 | 0.112 |
| CSGLMD | **0.958** | **0.957** | **0.958** | **0.407** | **0.960** | **0.957** | **0.960** | **0.481** | **0.967** | **0.971** | **0.975** | **0.177** |

## 3 Parameter sensitivity analysis

**Influence of similarity threshold** :The performance of CSGLMD under different similarity thresholds is first analyzed, which are searched from 0.5, 0.6, 0.7, 0.8, 0.9. The experimental results are reported in **Supplementary Table 2 (ST2)**. We can observe that the similarity threshold , as the determinant of the number of intra-layer edges in LMDHG, has a significant impact on the performance of the model. When , the model achieves the best performance in the LDA and LMI prediction tasks. Then, as the threshold increases, the number of intra-layer edges decreases, resulting in a decrease in CSGLMD performance. Interestingly, however, CSGLMD exhibit an inverse trend in the MDA prediction task, the performance stabilizing first and then start to increase. The possible reason is that dataset 1 contains a large number of known MDAs, thereby more intra-layer similarity edges can be computed. When the threshold is small, many unnecessary and redundant intra-layer edges are generated, which can add noise to the embedding feature learning process. Therefore, when running the proposed model on MDA prediction task, is set to 0.8. Overall, it is crucial to control the connectivity of intra-layer edges in LMDHG by means of similarity thresholds when predicting different tasks.

**Supplementary Table 2 (ST2). The performance of CSGLMD under different similarity thresholds on dataset 1.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| θ | LDA | | | | MDA | | | | LMI | | | |
| 5-cv1 | | 5-cv2 | | 5-cv1 | | 5-cv2 | | 5-cv1 | | 5-cv2 | |
| AUC | AUPR | AUC | AUPR | AUC | AUPR | AUC | AUPR | AUC | AUPR | AUC | AUPR |
| 0.5 | **0.958** | **0.957** | **0.958** | **0.407** | 0.954 | 0.951 | 0.952 | 0.448 | **0.967** | **0.971** | **0.975** | **0.177** |
| 0.6 | 0.949 | 0.946 | 0.953 | 0.302 | 0.953 | 0.951 | 0.952 | 0.449 | 0.966 | 0.960 | 0.971 | 0.129 |
| 0.7 | 0.946 | 0.936 | 0.948 | 0.194 | 0.954 | 0.950 | 0.954 | 0.448 | **0.967** | 0.960 | 0.973 | 0.131 |
| 0.8 | 0.948 | 0.946 | 0.951 | 0.179 | **0.960** | 0.956 | **0.960** | **0.481** | 0.963 | 0.956 | 0.969 | 0.117 |
| 0.9 | 0.947 | 0.943 | 0.946 | 0.170 | **0.960** | **0.957** | 0.960 | 0.467 | 0.962 | 0.953 | 0.969 | 0.071 |

**Influence of initializing node embedding dimension** : To investigate the sensitivity of the initial node embedding dimension , we perform comparison experiments using different embedding dimension settings ranging from 64, 128, 256, 512. We chose two layers of GCN as encoders, with each embedding dimension set to , in turn. **Supplementary Table 3 (ST3)** shown that, when , CSGLMD yields the best performance for LDA and MDA prediction task. A possible explanation is that too high a dimension may produce redundant features, while too small dimension can generate sparse features that are insufficient to represent the node embedding. For LMI prediction task, CSGLMD obtain the best performance when the embedding dimension is 512. This suggests that increasing embedding dimension is beneficial for learning node features from sparse lncRNA-miRNA networks.

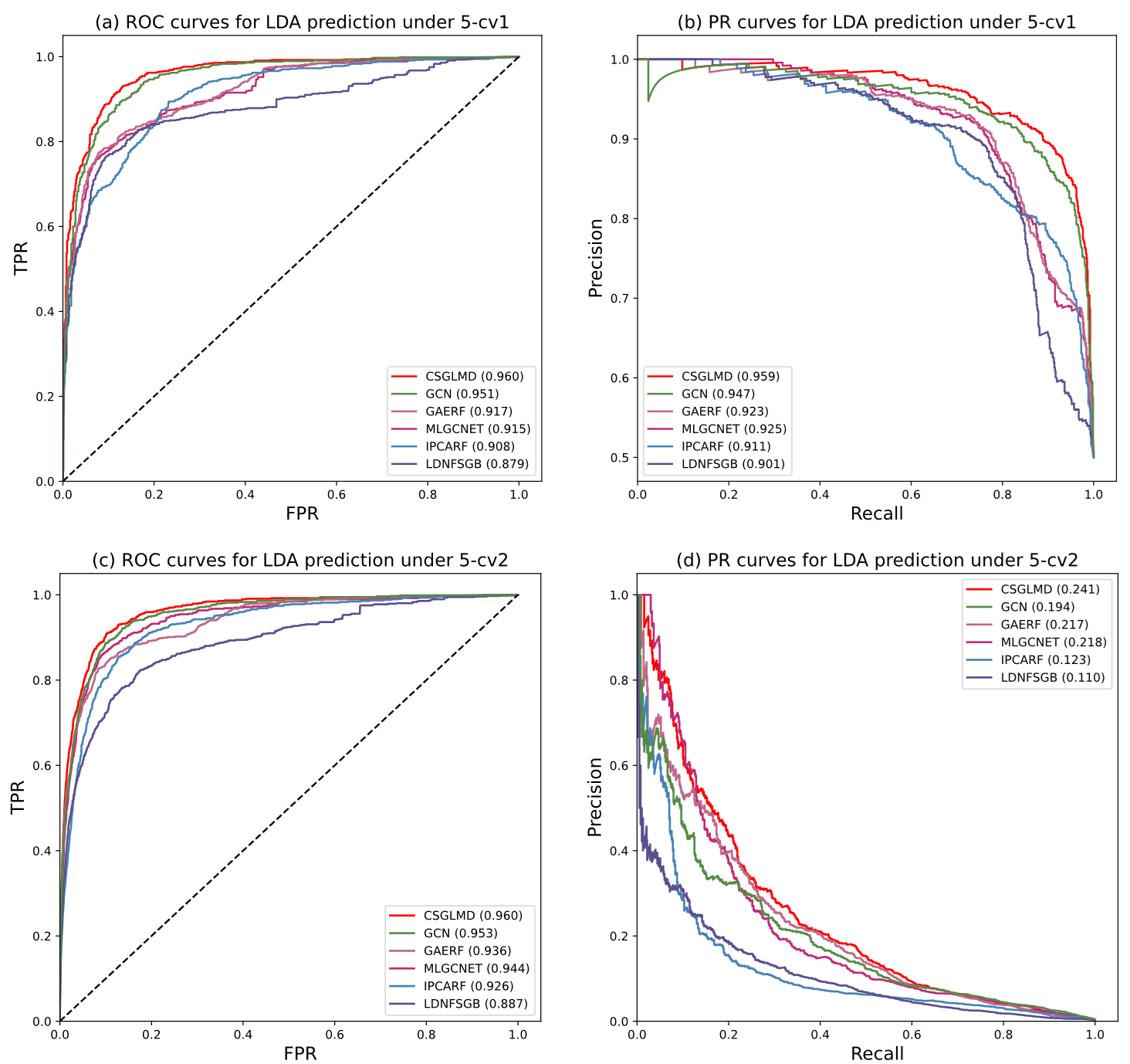
**Supplementary Table 3 (ST3). Performance of CSGLMD under different initialized the node embedding dimension on dataset 1.**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *d* | LDA | | | | MDA | | | | LMI | | | |
| 5-cv1 | | 5-cv2 | | 5-cv1 | | 5-cv2 | | 5-cv1 | | 5-cv2 | |
| AUC | AUPR | AUC | AUPR | AUC | AUPR | AUC | AUPR | AUC | AUPR | AUC | AUPR |
| 64 | 0.945 | 0.948 | 0.950 | 0.268 | 0.952 | 0.950 | 0.952 | 0.453 | 0.962 | 0.960 | 0.963 | 0.095 |
| 128 | 0.949 | 0.951 | 0.953 | 0.362 | 0.955 | 0.952 | 0.953 | 0.452 | 0.964 | 0.963 | 0.965 | 0.099 |
| 256 | **0.958** | **0.957** | **0.958** | **0.407** | **0.960** | **0.957** | **0.960** | **0.481** | 0.968 | 0.960 | 0.970 | 0.152 |
| 512 | 0.953 | 0.951 | 0.957 | 0.337 | 0.955 | 0.950 | 0.953 | 0.439 | **0.971** | **0.967** | **0.975** | **0.177** |

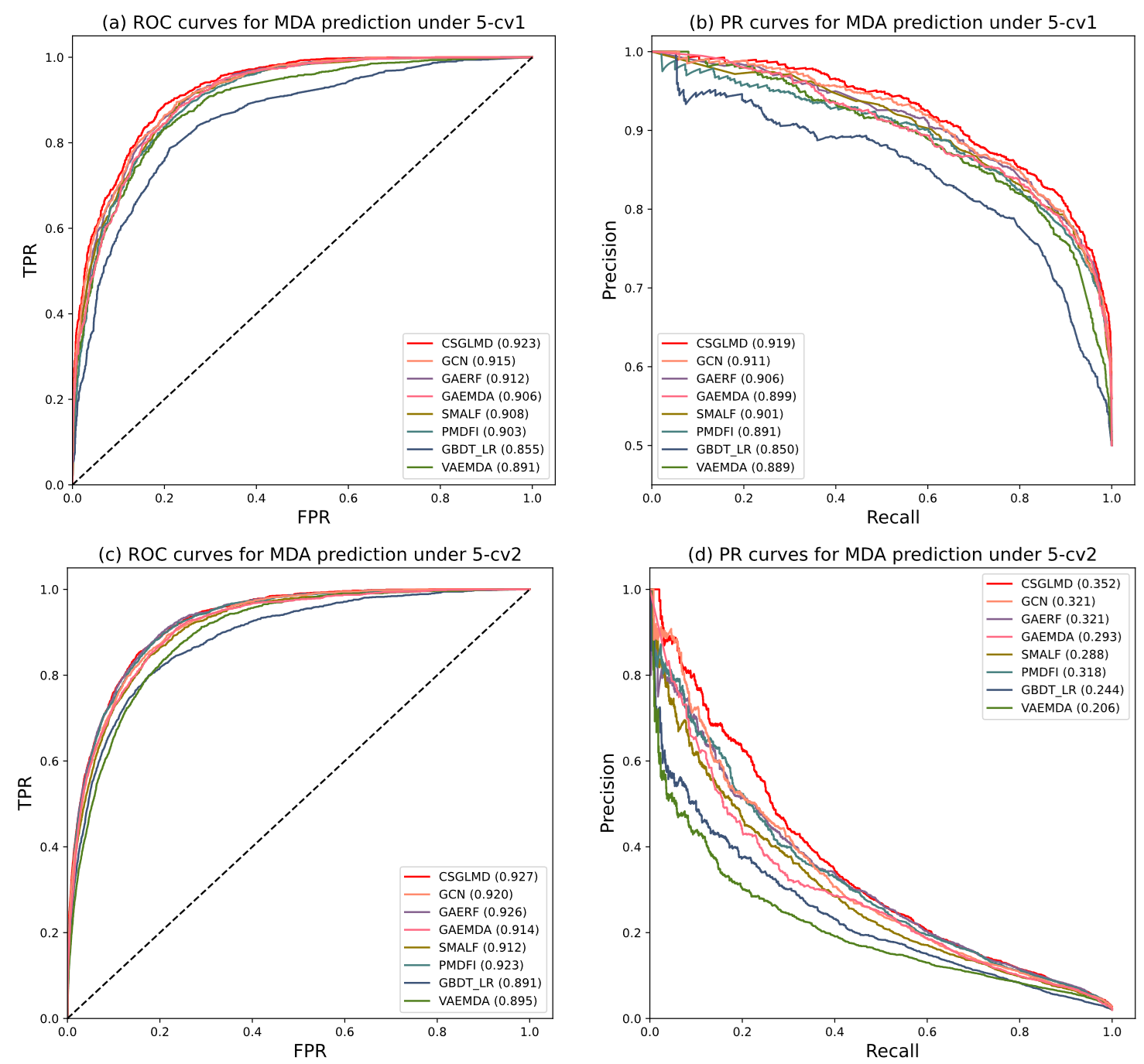
## 4 Comparison with state-of-the-art on dataset 2

**Supplementary Table 4 (ST4): Experimental results of CSGLMD and other baseline methods in LDA and MDA predictions on dataset 2.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Methods | 5-cv1 | 5-cv1 | 5-cv2 | 5-cv2 |
| AUC | AUPR | AUC | AUPR |
| LDA | GCN | 0.951 | 0.947 | 0.953 | 0.194 |
| GAERF | 0.917 | 0.923 | 0.936 | 0.217 |
| MLGCNET | 0.915 | 0.925 | 0.944 | 0.218 |
| IPCARF | 0.908 | 0.911 | 0.926 | 0.127 |
| LDNFSGB | 0.879 | 0.901 | 0.887 | 0.239 |
| CSGLMD | **0.960** | **0.959** | **0.960** | **0.241** |
| MDA | GCN | 0.915 | 0.911 | 0.920 | 0.321 |
| GAERF | 0.912 | 0.906 | 0.926 | 0.321 |
| GAEMDA | 0.906 | 0.899 | 0.914 | 0.293 |
| SMALF | 0.908 | 0.901 | 0.912 | 0.288 |
| PMDFI | 0.903 | 0.891 | 0.923 | 0.318 |
| GBDT\_LR | 0.855 | 0.850 | 0.891 | 0.244 |
| VAEMDA | 0.891 | 0.889 | 0.895 | 0.206 |
| CSGLMD | **0.923** | **0.919** | **0.927** | **0.352** |



**Supplementary Figure 1 (SF1):** Performance CSGLMD and other baseline methods on dataset 2 for LDA prediction. **(a)-(b)** Comparison of ROC curves, AUC values, PR curves, and AUPR values under 5-cv1. (c)-(d) Comparison of ROC curves, AUC values, PR curves, and AUPR values under 5-cv2.



**Supplementary Figure 2 (SF2):** Performance CSGLMD and other baseline methods on dataset 2 for MDA prediction. **(a)-(b)** Comparison of ROC curves, AUC values, PR curves, and AUPR values under 5-cv1. (c)-(d) Comparison of ROC curves, AUC values, PR curves, and AUPR values under 5-cv2.

## 5 Case studies on dataset 1 (based on the old version databases)

**Supplementary Table 5 (ST5). The top 20 predicted SC-related miRNA candidates on dataset 1.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Rank | MiRNA name | Evidence | Rank | MiRNA name | Evidence |
| 1 | hsa-mir-19b | dbDEMC v3.0, HMDD v3.2 | 11 | hsa-mir-30a | HMDD v3.2 |
| 2 | hsa-mir-15a | dbDEMC v3.0, HMDD v3.2 | 12 | hsa-mir-205 | dbDEMC v3.0, HMDD v3.2 |
| 3 | hsa-mir-29b | dbDEMC v3.0, HMDD v3.2 | 13 | hsa-let-7i | dbDEMC v3.0, HMDD v3.2 |
| 4 | hsa-mir-92a | dbDEMC v3.0, HMDD v3.2 | 14 | hsa-let-7e | dbDEMC v3.0 |
| 5 | hsa-mir-133a | dbDEMC v3.0, HMDD v3.2 | 15 | hsa-mir-26a | dbDEMC v3.0, HMDD v3.2 |
| 6 | hsa-let-7d | dbDEMC v3.0 | 16 | hsa-mir-96 | HMDD v3.2 |
| 7 | hsa-let-7b | dbDEMC v3.0, HMDD v3.2 | 17 | hsa-mir-23b | dbDEMC v3.0, HMDD v3.2 |
| 8 | hsa-let-7c | dbDEMC v3.0 | 18 | hsa-mir-708 | dbDEMC v3.0 |
| 9 | hsa-mir-210 | HMDD v3.2 | 19 | hsa-mir-99a | dbDEMC v3.0, HMDD v3.2 |
| 10 | hsa-mir-203 | dbDEMC v3.0, HMDD v3.2 | 20 | hsa-mir-10a | dbDEMC v3.0, HMDD v3.2 |

## 6 Case studies on dataset 2 (based on the new version databases)

PC is the second most common type of male cancer after LC and has a serious impact on males health worldwide [3]. **Supplementary Table 6 (ST6)** reports the top 5 LC-associated candidate lncRNAs and miRNAs. Last year, the top 1 ranked PC-associated lncRNA TP73-AS1 was confirmed by Arslan et al [4]. Their experimental results showed that TP73-AS1 was upregulated in PC cells compared with normal prostate cells. Correspondingly, the results of Zhang et al. and Wo et al. indicated that the lncRNAs OIP5-AS1 and SOX2-OT can promote PC proliferation [5, 6]. However, lncRNA DLEU1 and HAND2-AS1 have not been proved to be associated with PC by biological experiments. Among the top 5 predicted candidate miRNAs associated with PC, 3 out of 5 could be confirmed by the dbDEMC v3.0 database and 2 out of 5 could be verified by the literature. For instance, Zhao et al. applied laboratory experiments to suggest that miRNA has-mir-140 can target the YEAS gene to inhibit the invasion and migration of PC cells [7].

**Supplementary Table 6 (ST6).** **The top 5 predicted PC-related lncRNAs and miRNA candidates on dataset 2.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Rank | LncRNA name | Evidence | Rank | MiRNA name | Evidence |
| 1 | TP73-AS1 | PMID: 35138524 | 1 | hsa-mir-372 | dbDEMC v3.0 |
| 2 | OIP5-AS1 | PMID: 34051661 | 2 | hsa-mir-10b | dbDEMC v3.0 |
| 3 | DLEU1 | Unknown | 3 | hsa-mir-140 | PMID: 31310382 |
| 4 | SOX2-OT | PMID: 31623830 | 4 | hsa-mir-142 | PMID: 35510208 |
| 5 | HAND2-AS1 | Unknown | 5 | hsa-mir-30b | dbDEMC v3.0 |

CC is a malignant cancer worldwide and the fourth leading cause of cancer-related death in women [3]. **Supplementary Table 7 (ST7)** lists the top 5 predicted lncRNAs and miRNAs associated with CC. Among them, Liu et al. showed that the top 1 ranked lncRNA KCNQ1OT1 can contribute CC tumor growth by regulating miR-296-5p/HYOU1 axis [8]. LncRNA BANCR and SNHG6 have been demonstrated to target miR-582-5p and miR-485-3p, respectively, to regulate CC cell growth [9, 10]. The SOX2-OT and HOTTIP have not been confirmed by laboratory experiments. 4 out of 5 candidate miRNAs are identified by literature and remaining 1 miRNA is confirmed by dbDEMC v3.0. The top 1 ranked miRNA, hsa-mir-183, can be controlled by the lncRNA CRNDE to regulate the gene expression of CCNB and thus the growth of CC cells [11]. Overall, these three case studies further substantiate the capability of CSGLMD to identify potentially novel disease-related lncRNAs and miRNAs.

**Supplementary Table 7 (ST7).** **The top 5 predicted CC-related lncRNAs and miRNA candidates on dataset 2.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Rank | LncRNA name | Evidence | Rank | MiRNA name | Evidence |
| 1 | KCNQ1OT1 | PMID: 34704918 | 1 | hsa-mir-183 | PMID 31605132 |
| 2 | SOX2-OT | Unknown | 2 | hsa-mir-126 | PMID: 31007650 |
| 3 | BANCR | PMID: 35253437 | 3 | hsa-mir-221 | dbDEMC v3.0 |
| 4 | SNHG6 | PMID: 32884447 | 4 | hsa-mir-372 | PMID: 21646351 |
| 5 | HOTTIP | Unknown | 5 | hsa-mir-182 | PMID 26191165 |

## Reference

1. Chen X, Clarence Yan C, Luo C et al. Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity, Scientific Reports 2015;5(1):11338.

2. Wang D, Wang J, Lu M et al. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases, Bioinformatics 2010;26(13):1644-1650.

3. Bray F, Ferlay J, Soerjomataram I et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: A Cancer Journal for Clinicians 2018;68(6):394-424.

4. Arslan A, Batar B, Temiz E et al. Silencing of TP73-AS1 impairs prostate cancer cell proliferation and induces apoptosis via regulation of TP73, Molecular Biology Reports 2022;49(7):6859-6869.

5. Zhang Y, Guo S, Wang S et al. LncRNA OIP5-AS1 inhibits ferroptosis in prostate cancer with long-term cadmium exposure through miR-128-3p/SLC7A11 signaling, Ecotoxicology and Environmental Safety 2021;220:112376.

6. Wo Q, Zhang D, Hu L et al. Long noncoding RNA SOX2-OT facilitates prostate cancer cell proliferation and migration via miR-369-3p/CFL2 axis, Biochemical and Biophysical Research Communications 2019;520(3):586-593.

7. Zhao S, Jie C, Xu P et al. MicroRNA‐140 inhibit prostate cancer cell invasion and migration by targeting YES proto‐oncogene 1, Journal of Cellular Biochemistry 2020;121(1):482-488.

8. Liu J, Wang Y. Long non-coding RNA KCNQ1OT1 facilitates the progression of cervical cancer and tumor growth through modulating miR-296-5p/HYOU1 axis, Bioengineered 2021;12(1):8753-8767.

9. Wang B, Wang M, Jia S et al. Systematic Survey of the Regulatory Networks of the Long Noncoding RNA BANCR in Cervical Cancer Cells, Journal of Proteome Research 2022;21(4):1137-1152.

10. Liu J, Liu X, Li R. LncRNA SNHG6 enhances the radioresistance and promotes the growth of cervical cancer cells by sponging miR-485-3p, Cancer cell international 2020;20(1):424.

11. Bai X, Wang W, Zhao P et al. LncRNA CRNDE acts as an oncogene in cervical cancer through sponging miR-183 to regulate CCNB1 expression, Carcinogenesis 2019;41(1):111-121.