RESEARCH ARTICLE

A deep forest model integrating multi-source medical data for the prediction of the association between lncRNAs and diabetes

Xiangxiang Mei*, Fang Wu, Xiaodan Cai

School of Yonyou Digital Intelligence, Nantong Institute of Technology, Nantong, Jiangsu, China.

Received: December 9, 2025; **accepted:** April 3, 2026.

Long non-coding RNAs (lncRNAs) are RNA transcripts longer than 200 nucleotides that lack protein-coding potential but regulate gene expression and metabolic pathways. Growing evidence indicates that dysregulated lncRNAs contribute to the development of diabetes mellitus and its complications. However, predicting lncRNA-diabetes associations remains challenging due to heterogeneous data sources and complex molecular interactions. This research developed a deep forest model integrating multi-source medical data including Genome-Wide Association Study (GWAS) data, gene expression profiles, and protein interaction networks. Weighted Gene Co-expression Network Analysis (WGCNA) and DeepWalk were employed for feature extraction followed by machine learning classifiers to construct the predictive framework. The model was applied to differential lncRNA screening in hypertriglyceridemia patients with type 2 diabetes, comparative analysis of obesity-associated GWAS loci, and exosomal lncRNA association studies in diabetic retinopathy (DR). The FPKM values ranged from 0 to 9, and the correlation coefficients exceeded 0.9, indicating acceptable sequencing consistency. Twenty-three single nucleotide polymorphisms (SNPs) with both sequence and mutation-site conservation were identified, and three functional gene modules were detected in the obesity GWAS interaction network. Additionally, three exosomal lncRNAs were significantly associated with DR ($P < 0.05$). These results demonstrated that the proposed deep forest framework provided an effective computational strategy for identifying diabetes-associated lncRNAs.

Keywords: GWAS data; WGCNA algorithm; DeepWalk; lncRNA-diabetes; Deep forest modeling.

***Corresponding author:** Xiangxiang Mei, School of Yonyou Digital Intelligence, Nantong Institute of Technology, Nantong, Jiangsu 226000, China. Email: xx_mandy@sina.com.

Introduction

Long non-coding RNAs (lncRNAs) are RNA transcripts longer than 200 nucleotides that lack protein-coding potential but play important regulatory roles in gene expression and cellular processes [1]. Although initially considered transcriptional by-products, lncRNAs are now recognized as key regulators in diverse biological pathways including epigenetic modification, transcriptional control, and signal transduction. Increasing evidence has demonstrated that dysregulated lncRNAs are closely associated with complex diseases such as cancer, diabetes, and metabolic disorders [2, 3]. Previous research found that LIMIT as an immunogenic lncRNA involved in cancer immunity [4]. lncRNA DRAIR could modulate inflammatory phenotypes in diabetic monocytes [3]. In addition, lncRNA-mediated regulation of the NLRP3 inflammasome has been implicated in diabetes complications [5]. These findings indicate that lncRNAs play

critical roles in the development and progression of diabetes mellitus.

With advances in high-throughput sequencing technologies, experimentally validated lncRNA-disease associations have accumulated rapidly [6]. Several curated databases have been established to organize these data including LncRNADisease and Lnc2Cancer [7, 8]. However, the currently known lncRNA-disease associations represent only a small fraction of the true biological landscape. Experimental validation remains time-consuming and costly, which limits large-scale identification of novel disease-related lncRNAs [9]. Therefore, computational prediction methods have emerged as effective complementary strategies. Early approaches employed network-based inference methods and matrix factorization techniques [10, 11]. More recently, deep learning models such as LDICDL [12], GANLDA [13], BiGAN [14], and DMFLDA [15] have improved nonlinear feature representation and predictive performance. Random Forest-based models have also demonstrated effectiveness in identifying potential lncRNA-disease associations [16]. Despite these advances, several limitations remain. Many existing models rely on single data sources and do not fully integrate heterogeneous biological information such as genomic variants, gene expression patterns, and interaction networks [17]. In addition, deep neural network models often require large labeled datasets and may lack interpretability and robustness when applied to limited biomedical samples [18]. Given the complex genetic architecture of diabetes and its complications, there remains a need for integrative computational frameworks capable of leveraging multi-source medical data while maintaining stable predictive performance.

To address these challenges, a deep forest-based model was developed to predict lncRNA-diabetes associations in this research. Deep forest models, also known as gcForest, have demonstrated strong performance in biological sequence classification tasks and predictive modeling [19, 20], offering advantages in feature representation and robustness without requiring extensive parameter tuning. Genome-wide association study (GWAS) data, gene expression profiles, and protein interaction networks were integrated. Weighted gene co-expression network analysis (WGCNA) was applied to identify biologically relevant gene modules, and DeepWalk was used to capture topological features of heterogeneous networks. These features were incorporated into a deep forest ensemble framework to construct the predictive model. By integrating multi-source biological data and ensemble learning strategies, the proposed framework provided a systematic computational approach for identifying lncRNAs associated with diabetes and its complications. The proposed framework could improve the accuracy and robustness of lncRNA-disease prediction and offer methodological support for future precision medicine research in metabolic diseases.

Materials and methods

Data sources

Genome-wide association data were obtained from the GWAS catalog maintained by the European Bioinformatics Institute (https://www.ebi.ac.uk/gwas/). Data retrieval was performed in March 2025. Studies associated with diabetes mellitus, obesity, hypertriglyceridemia, and diabetic retinopathy were identified by using phenotype-based keyword searches. The downloaded records included single nucleotide polymorphism (SNP) identifiers (rsID), associated traits, mapped genes, reported risk alleles, *P*-values, and effect size information when available. Only associations reaching genome-wide significance ($P < 5 \times 10^{-8}$) were retained. After filtering and removing duplicated entries, 23 SNP loci showing both sequence conservation and mutation-site conservation were included for subsequent analyses. All data were exported directly from the GWAS portal in TSV format. Gene annotation and molecular validation data were retrieved from databases hosted by the National Center for

Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/) in March 2025. Gene structural and positional information was obtained from the Gene database (https://www.ncbi.nlm.nih.gov/gene/), while SNP validation and conservation analysis were performed by using dbSNP (https://www.ncbi.nlm.nih.gov/snp/). RNA sequencing datasets related to type 2 diabetes and hypertriglyceridemia were screened from the Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/). Corresponding expression matrices in the fragments per kilobase of transcript per million mapped reads (FPKM) format were downloaded and used for differential lncRNA analysis and downstream co-expression network construction.

Network construction and predictive modeling

Gene co-expression networks were constructed by using the WGCNA package (https://cran.r-project.org/package=WGCNA) implemented in R (version 4.0.4) (https://www.r-project.org/). A soft-thresholding power was selected based on the scale-free topology criterion to ensure network stability. Modules significantly associated with diabetes-related phenotypes were identified through module–trait correlation analysis. To explore the biological relevance of candidate modules, functional enrichment analysis was performed by using the Gene ontology (GO) (http://geneontology.org/) and Kyoto encyclopedia of genes and genomes (KEGG) (https://www.genome.jp/kegg/) databases. Enrichment results were visualized by using the ggplot2 package (https://ggplot2.tidyverse.org/) in R. To capture structural information within the heterogeneous lncRNA-gene interaction network, node embeddings were generated by using the DeepWalk algorithm (https://github.com/phanein/deepwalk), which applied a Skip-Gram model to random walk sequences for representation learning. Random walks were performed on the constructed interaction graph, and low-dimensional feature vectors were learned for each node. The resulting network embeddings were combined with GWAS-derived SNP features and WGCNA module information to construct an integrated feature matrix. Classification models including support vector machine (SVM), random forest (RF), and logistic regression (LR) were implemented by using the scikit-learn library (https://scikit-learn.org/) in Python (version 3.8) (https://www.python.org/). A deep forest framework based on cascade ensemble architecture was subsequently constructed to integrate multi-source biological features. The ensemble model consisted of multi-layer random forest classifiers, enabling hierarchical feature transformation and improved predictive robustness. The final model was used to predict potential associations between lncRNAs and diabetes.

Identification of key modules and genes by weighted gene co-expression network analysis

WGCNA was used to construct weighted gene co-expression networks for exploring the correlation between gene networks and biological traits by identifying co-expression gene modules. The gene expression values for 20,958 genes were ranked according to median absolute deviation (MAD). The top 5,005 most variable genes were selected for co-expression analysis. The included samples were analyzed by the hclust function in the WGCNA package and clustered according to the expression level of the genes in each sample. The correlation matrices were subsequently constructed, and the Pearson correlation coefficients between genes were calculated. The filtering threshold was applied to determine whether the genes had similar expression profiles. The weighted adjacency matrix was converted to a topological overlap matrix (TOM) to assess network connectivity, and a hierarchical clustering method was applied to construct a clustering dendrogram. Genes were clustered into multiple modules based on weighted correlation coefficients according to different expression patterns, and similar modules were clustered into one module by setting the similarity threshold to 0.2. Genes with similar expression patterns were grouped into the same modules. After constructing the co-expression

network and obtaining several different modules, the heat map of module and node gene expression were drawn by calculating the expression value matrix of the genes in the network and the module information corresponding to the node genes. By evaluating the relationship between the calculated module ME and the clinical traits DKD and ESRD, the significant correlation between gene expression and traits (GS) and the correlation coefficient between the expression of a certain gene and the expression of the main component of genes in the module (MM) were obtained, and the modules associated with the development of DKD into ESRD were identified based on GS and MM.

Enrichment analysis

The GO and KEGG pathway enrichment analysis of differential genes and key modules was performed by using the "clusterprofiler" package in R 4.0.4 software. The *P* value less than 0.05 indicated that the difference in enrichment results was statistically significant. The "gg-plot2" package was used to display GO and KEGG entries, and the significance threshold was set at $P < 0.05$.

DeepWalk-based feature vector extraction

DeepWalk was used to compute vectorized representations of all nodes including IncRNAs and proteins. DeepWalk consisted of two major steps. For each vertex v_i, a randomized tour was performed with v_i as the starting vertex and γ with t as the length. The representation of the vertices passed by each walk was then updated using the Skip-Gram algorithm to maximize the likelihood of coexistence of the vertices appearing in the window w based on the assumption of independence as follows.

$$\Pr\left(\{v_{i-w},\mathrm{L}\ ,v_{i+w}\}\setminus v_i \mid \Phi\left(v_i\right)\right)=\prod_{j=i-w,j\neq i}^{i+w}\Pr\left(v_j \mid \Phi\left(v_i\right)\right) \quad (1)$$

where Φ was the potential topological representation associated with each vertex v_i and the matrix for $|V|\times d$, where $|V|$ was the base of the vertex set V, and d was the dimension of the vertex vector. To speed up the training time, the vertices were decomposed and assigned to the leaf nodes of the binary tree by $Pr\left(v_j|\Phi(v_i)\right)$ using Hierarchical Softmax. $Pr\left(v_j|\Phi(v_i)\right)$ was computed as below.

$$\Pr\left(v_j \mid \Phi\left(v_i\right)\right)=\prod_{i=1}^{\lceil \log|V| \rceil} 1\Big/\left(1+e^{-\Phi(v_i)\cdot\Psi(b_l)}\right) \quad (2)$$

where $\psi\left(b_l\right)$ was the parent node of tree node $b_l \cdot \left(b_0, b_1 \cdots b_{[log|V|]}\right)$ and was used to identify v_j, the sequence of nodes in the tree, where b_0 was the root node and $b + [log|V|] += v_j$. After completing the feature training, the output of DeepWalk was a potential topological representation of the nodes in the network, i.e., the word vector was applied to the network space and the feature vector was output, which represented the sequence of nodes that randomly wandered through the network as a "sentence", a feature vector representation. Therefore, the similarity of two nodes *u* and *v* could be calculated by cosine similarity as follows.

$$sim(u,v)=\frac{\sum_{k=1}^{d}u_k v_k}{\sqrt{\sum_{k=1}^{d}u_k^2}\sqrt{\sum_{k=1}^{d}v_k^2}} \quad (3)$$

where d was the dimension. u_i and v_i were the components of vectors *u* and *v*, respectively. To represent the feature vectors of the nodes more conveniently, an index table was built to number all the nodes in the heterogeneous network, keep the name information and node type, and deeply mine the hidden topology information in the heterogeneous network based on the network representation learning algorithm. The nodes were embedded to be represented as feature vectors based on the

more attention to the external connections between the nodes.

Classifier construction

Node embeddings generated by DeepWalk from the heterogeneous network were used as input features to construct classifiers including support vector machines (SVM), random forest (RF), and logistic regression (LR). The classifiers were implemented by using the scikit-learn library in Python. For SVM, the radial basis function kernel was applied. The RF model consisted of multiple decision trees constructed *via* bootstrap aggregation. Logistic regression was performed with L2 regularization. Model hyperparameters were tuned by using grid search combined with cross-validation.

Results and discussion

Screening of differential lncRNAs for type 2 diabetes in hypertriglyceridemic populations

Genetic susceptibility to obesity and chronic hyperglycemia is associated with an increased risk of hypertriglyceridemia that is one of the important risk factors for type 2 diabetes, and its pathogenesis is closely related to the regulation of gene expression. This research analyzed the peripheral blood gene (RNA) expression value FPKM of hypertriglyceridemia population using deep forest model based on medical data. Data quality was assessed through correlation analysis of gene and transcript levels among samples. The research successfully screened the feature vectors extracted from DeepWalk to identify the hypertriglyceridemia population with type 2 diabetes mellitus that possessed the differential lncRNAs.

(1) Quality assessment of peripheral blood RNA extraction and sequencing data

In this study, the correlation coefficients were calculated and plotted as heatmaps for intra- and inter-group samples based on the expression values FPKM of all genes and transcripts in each sample. Pearson correlation heatmaps at the gene and transcript levels demonstrated that the Pearson correlation coefficients of the gene levels among samples were above 0.9, and most of them were close to 1 (Figure 1A). The Pearson correlation coefficients at transcript level exceeded 0.8 across all samples and were generally close to 0.9 (Figure 1B). The correlation coefficient results indicated that the gene expressions within the group were well correlated, indicating adequate data quality for downstream analysis.

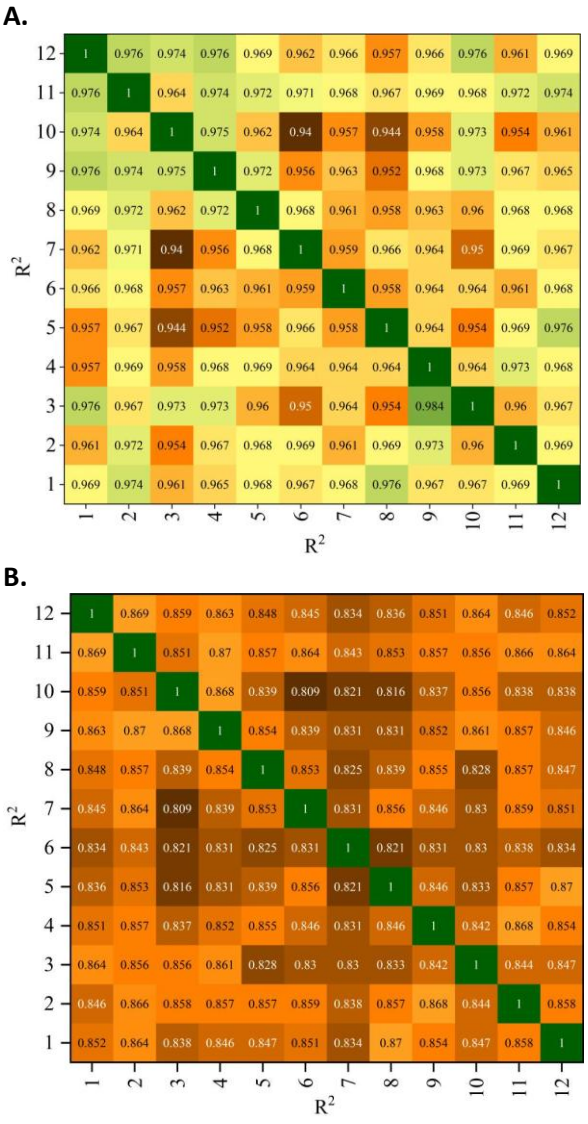
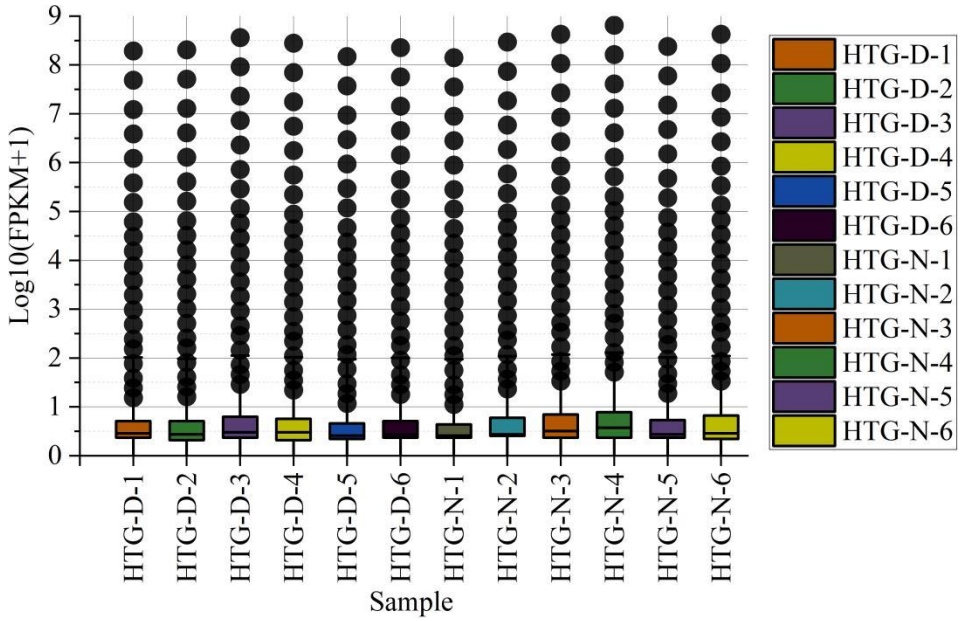


Figure 1. Pearson correlation analysis across samples. **A.** gene-level correlations. **B.** transcript-level correlations. Horizontal coordinates 1 - 12 represented HTG-N-6, HTG-N-5, HTG-N-4, HTG-N-3, HTG-N-2, HTG-N-1, HTG-D-6, HTG-D-5, HTG-D-4, HTG-D-3, HTG-D-2, HTG-D-1, respectively. Vertical coordinates 1 - 12 represented HTG-D-1, respectively, HTG-D-2, HTG-D-3, HTG-D-4, HTG-D-5, HTG-D-6, HTG-N-1, HTG-N-2, HTG-N-3, HTG-N-4, HTG-N-5, HTG-N-6, respectively.

A.



B.

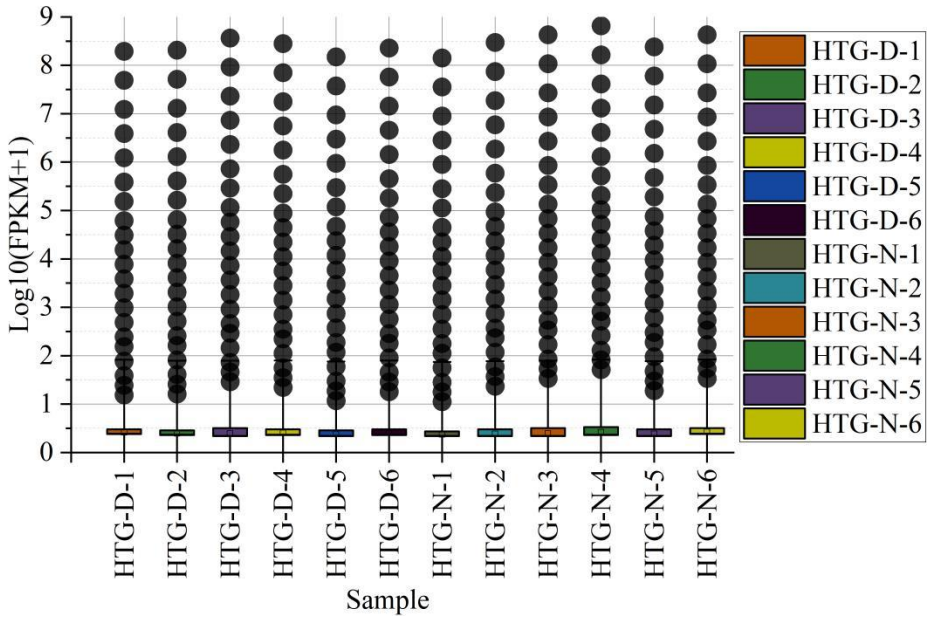


Figure 2. Distribution of gene and transcript expression levels across samples. **A.** gene-level expression distribution. **B.** transcript-level expression distribution.

(2) Quantitative sequencing analysis and differential lncRNA screening

Quantitative analysis is the basis of difference analysis. The study analyzed the distribution of FPKM at the gene level or transcript expression level of different samples. The results showed that the log-transformed FPKM values at both gene and transcript levels were primarily distributed within the range of 0 - 9 with the majority concentrated between 1 and 5, showing a right-skewed distribution pattern (Figure 2).

Analysis of the relationship between obesity genes and diabetes

Table 1. Comparative genomic analysis results of obesity GWAS lead SNPs.

SNPs	Chr+	Gene	VISTA		
			Mouse (P)	Chimp(P)	UCSC
rs543874	1	SEC16B	1.90E-18	-	6/9
rs11208659	1	LEPR	0.35	2.20E-06	6/9
rs1514175	1	TNNI3K	1.00E-04	3.90E-05	6/9
rs887912	2	FANCL	3.00E-82	0.22	7/9
rs7599312	2	ERBB4	9.90E-07	-	6/9
rs12617233	2	LINC01122	1.80E-25	-	5/9
rs17203016	2	CREBIKLFT	6.50E-10	-	6/9
rs1460676	2	FIGN	1.50E-66	2.7E-17	6/9
rs12617233	2	LINC01122	1.80E-25	-	5/9
rs1 1126666	2	KCNK3	-	1.00E-07	5/9
rs3849570	3	GBE1	3.00E-12	-	6/9
rs2272903	6	TFAP2B	1.1E-11	-	6/9
rs987237	6	TFAP2B	3.2E-10	-	6/9
rs9400239	6	FOXO3	7.9E-07	-	6/9
rs2033529	6	TDRG1	-	3.60E-03	5/9
rs2245368	7	DTX2P1/UPK3B, PI/PMS2P1	5.50E-04	1.8E-18	8/9
rs10767664	11	BDNF	8.30E-05	-	7/9
rs7132908	12	FAIM2	1.70E-04	-	6/9
rs997295	15	MAP2KS	7.40E-63	3.00E-03	8/9
rs9925964	16	KAT8	-	7.80E-06	5/9
rs1558902	16	FTO	-	1.30E-03	5/9
rs1421085	16	FTO	3.2E-09	1.30E-03	7/9
rs3751812	16	FTO	2.20E-04	-	6/9

Individuals carrying obesity-associated genetic variants are at increased risk of hypertriglyceridemia. Obesity is strongly linked to the development of type 2 diabetes, and its genetic determinants may contribute to the shared pathogenic mechanisms underlying both conditions. The genomics and protein interaction networks and their functional modules were analyzed and compared in this study by using WGCNA algorithm in the deep forest model to identify several key obesity “hub” and “bottomleneck” genes. Their roles in the diabetes-related signaling pathways were then investigated. By studying the roles of these genes in the diabetes-related signaling pathways, this study provided a new perspective to understand the co-morbidity mechanism between obesity and diabetes.

(1) Comparative genomic analysis of obesity GWAS lead SNPs

To further investigate the potential genetic linkage between obesity and diabetes, lead SNPs from obesity-related GWAS were subjected to comparative genomic analysis. Visualization tools for alignments (VISTA) analysis showed that a total of 34 SNPs were sequence conserved, and 55 SNPs were sequence conserved among obese GWAS lead SNPs. University of California Santa Cruz (UCSC) Genome Browser analysis showed that a total of 61 SNPs were conserved > 5/9. The results of the comparative genomics analysis of obesity GWAS lead SNPs showed 23 SNPs with both sequence and mutant site conservation. The SNPs that did not require this test were listed as missing data. Most of the 23 SNPs with both sequence and mutant site conservation showed a conservatism of 6/9 with a maximum of 8/9 and a minimum of 5/9 (Table 1).

(2) Protein interaction networks and functional modules of obesity GWAS Genes

A protein-protein interaction (PPI) network of obesity-associated GWAS genes was constructed within the deep forest model to characterize functional relationships among encoded proteins. Network clustering analysis was subsequently performed to identify functional modules and key genes potentially involved in diabetes-related signaling pathways. The results showed the partial interaction network of the obese GWAS gene-encoded protein, which had a total of 289 nodes and 382 edges (Figure 3). The maximum node degree, average node degree, and minimum node degree of this network were 22.0, 5.2, and 2.0, respectively. Genes within the top 21% of node degree distribution (> 11 connections) were defined as hub genes including CREB1 (21), MC4R (18), PPARG (18), FTO (17), TMEM18 (17), BCL2 (16), TCF7L2 (14), IRS1 (13), SH2B1 (12), KCNJ11 (11), SEC16B (11), SLC30A8 (11), BTRC (10), CDKAL1 (10), BCDIN3D (10), IGF2BP2 (10), KCTD15 (10), RIT2 (10), MTCH2 (10), FOXO3 (10), and GNPDA2 (10). The maximum betweenness value, average betweenness value, and minimum betweenness value of this network were 1,469.14, 174.25, and 0, respectively. The top 21% of protein-coding genes with a betweenness value greater than 1,005 were defined as “bottleneck” genes including CREB1, BCL2, TMEM18, PPARG, FOXO3, TCF7L2, BDNF, BTRC, TLR4, IRS1, MC4R, FTO, MAP2K5, NRXN3, CTSS, PAX6, TNRC6B, KCNQ1, RPTOR, BMP2 BDNF, BTRC, TLR4, IRS1, MC4R, FTO, MAP2K5, NRXN3, CTSS, PAX6, TNRC6B, KCNQ1, RPTOR, BMP2, PARK2, suggesting potential central regulatory roles within the network. Genes identified as both hubs and bottlenecks were enriched in several diabetes-related pathways including the Neurotrophin signaling pathway (BCL2, BDNF, MAP2K5, FOXO3, IRS1, SH2B1), AMPK signaling pathway (CREB1, FOXO3, IRS1, PPARG, RPTOR), PI3K-Akt signaling pathway (BCL2, CREB1, FOXO3, RPTOR, TLR4, IRS1), type 2 diabetes (IRS1, KCNJ11), as well as several cancer-related pathways, all of which were statistically significant with false discovery rate (FDR) less than 0.05, indicating that the expected proportion of false positives among significant results was less than 5%, suggesting reliable pathway enrichment. Three functional modules with scores higher than 2.5 were obtained by cluster analysis. Module 1 included 19 genes of ADCY9, BTRCLHCGR, HHEX, RBBP6, PARK2, HUWEI, KCNJ, CDKAL1, IGF2BP2, GIPR, ASB4, PPARG, CALCR, TCF7L2, ADCY3, HNFIB, SLC30A8, and KCNQI. Module 2 included 9 genes of MC4R, FTO, SEC16B, TMEM18, BCDIN3D, KCTDI5, MTCH2, GNPDA2, and FAIM2. Module 3 included 5 genes of HLA-DRB1, KLCI, HLA-DOA1, HLA-DRA, and HLA-DRB5. These modules represented distinct biological processes contributing to obesity-diabetes comorbidity.

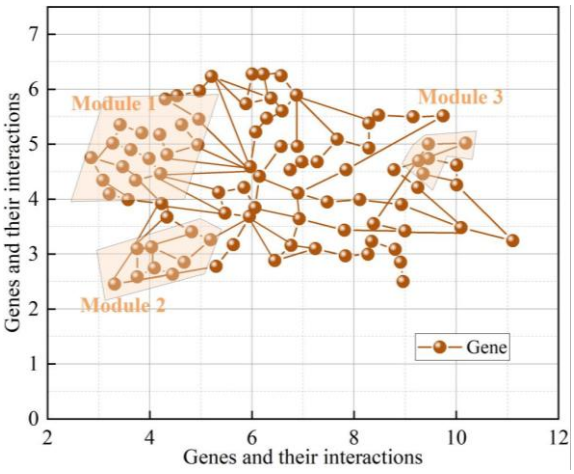


Figure 3. Obesity GWAS gene coding protein partial interaction network.

Association study of exosomal lncRNA and diabetic retinopathy

Diabetic retinopathy (DR) is one of the serious microvascular complications of diabetes mellitus, and DR that has progressed to the proliferative stage is one of the major causes of vision loss. Its pathogenesis is complex and involves multiple genes and molecular regulation. Existing studies have concluded that diabetic patients' long-term physiological states such as obesity and high glucose induce obstacles in the insulin pathway and oxidative stress in tissue cells, which in turn cause the development of some serious diabetic complications, resulting in a substantial disease burden for the patients. This study analyzed the association between exosomal lncRNA and DR using deep forest model. Three lncRNAs

Table 2. Factor analysis of the association between exosome LncRNA and DR.

	Single factor		Multi-factor*	
	OR (95% CI)	***P***	**OR (95% CI)**	***P***
DLX6-AS1	3.215 (1.520 - 6.801)	0.001	3.506 (1.506 - 8.164)	0.003
PRINS	0.168 (0.078 - 0.364)	< 0.001	0.131 (0.052 - 0.330)	< 0.001
FAM190A-3	0.320 (0.015 - 0.730)	0.006	0.318 (0.126 - 0.803)	0.014
ACY1	0.617 (0.284 - 1.338)	0.221	0.747 (0.316 - 1.764)	0.505
ARHGAP	1.449 (0.671 - 3.132)	0.347	1.231 (0.517 - 2.938)	0.641

* Adjusted for confounding variables including age, sex, BMI, WHR, smoking status, alcohol consumption, history of fatty liver disease, duration of diabetes, total bilirubin, and blood sugar.

associated with DR were identified by logistic regression analysis of multiple lncRNAs including lncRNA DLX-6AS1, lncRNARNA PRINS, and lncRNA FAM190A-3. The associations were further validated in different subgroups. These three lncRNAs might serve as potential biomarkers for improving early detection and clinical management of diabetic retinopathy, thereby contributing to reduced disease burden.

(1) Correlation analysis of exosomal lncRNAs and DRs

Studying the association between different exosomal lncRNAs and factors related to DR in hypertriglyceridemic population and determining which lncRNAs would be affected by factors such as lifestyle habits may support lifestyle management and assist clinical decision-making. The results of single-factor and multifactor logistic regression analysis of the association between exosomal lncRNA and DR showed that in the unifactorial logistic regression analysis, lncRNA DLX6-AS1, lncRNA PRINS, and lncRNA FAM190A-3 were associated with DR with *P* values of 0.001, < 0.001, and 0.006, respectively, and ORs and 95% CIs of 3.215 (1.520 - 6.801), 0.168 (0.078 - 0.364), and 0.320 (0.015 - 0.730), respectively. After adjusting for possible confounders such as age, sex, BMI, WHR, smoking status, drinking status, history of fatty liver, duration of diabetes mellitus, total bilirubin, and blood glucose, lncRNA DLX6-AS1, lncRNA PRINS, and lncRNA FAM190A-3 remained associated with DR with *P* values of 0.003, < 0.001, 0.014, respectively, and OR and 95% CI of 3.506 (1.506 - 8.164), 0.131 (0.052 - 0.330), and 0.318 (0.126 - 0.803), respectively (Table 2). These lncRNAs might play a role in the latent phase of DR by regulating exosome-mediated intercellular communication. Further identification of the specific influences of these lncRNAs could provide new targets for early diagnosis and treatment of DR and improve clinical management of diabetic complications.

(2) Multifactorial logistic regression analysis of exosomal lncRNA and DR association in different subgroups

Further multifactorial logistic regression analysis of the association of the lncRNA DLX6-AS1, lncRNA PRINS, and lncRNA FAM190A-3 with DR was performed to investigate what the specific influencing factors were. The results of the multifactorial logistic regression analysis in different subgroups showed that exosomal lncRNA DLX6-AS1 was associated with DR only in males whose age > 60, smoking group, alcohol group, fatty liver group with OR and 95% CI of 6.689 (2.136 - 20.953), 10.062 (1.474 - 68.720), 8.049 (1.260 - 51.458), 12.943 (1.217 - 137.775), 16.116 (1.061 - 244.915). lncRNA DLX6-AS1 was not associated with DR in other subgroups. LncRNA PRINS was associated with DR in all subgroups of the population. LncRNA FAM190A-3 was associated with DR only in males whose age > 60, non-alcohol drinking group, no fatty liver group with OR and 95% CI of 0.247 (0.076 - 0.798), 0.076 (0.010 - 0.521), 0.227 (0.071 - 0.716), and 0.238 (0.075 - 0.747) and was not associated with DR in other subgroups (Table 3). No significant interaction effects were observed between sex, age, smoking status, alcohol

Table 3. Multivariate logistic regression analysis in different subgroups*.

	DLX-6AS1	$P_{interaction}$	PRINs	$P_{interaction}$	FAM190A-3	$P_{interaction}$
Sex						
Male	6.689 (2.136 - 20.953)	0.096	0.152 (0.049 - 0.475)	0.802	0.247 (0.076 - 0.798)	0.393
Female	1.370 (0.285 - 6.611)		0.048 (0.007 - 0.380)		0.370 (0.073 - 1.843)	
Age						
≤ 60	2.592 (0.944 - 7.122)	0.166	0.095 (0.029 - 0.333)	0.501	0.563 (0.170 - 1.862)	0.215
> 60	10.062 (1.474 - 68.720)		0.130 (0.019 - 0.881)		0.076 (0.010 - 0.521)	
Smoke						
Yes	8.049 (1.260 - 51.458)	0.201	0.049 (0.006 - 0.394)	0.913	0.180 (0.028 - 1.150)	0.496
No	2.717 (0.870 - 8.483)		0.146 (0.042 - 0.514)		0.382 (0.111 - 1.311)	
Drink						
Yes	12.943 (1.217 - 137.775)	0.272	0.093 (0.009 - 0.913)	0.335	0.489 (0.059 - 4.091)	0.501
No	2.399 (0.908 - 6.343)		0.150 (0.052 - 0.436)		0.227 (0.071 - 0.716)	
Fatty liver						
Yes	16.116 (1.061 - 244.915)	0.173	0.026 (0.001 - 0.898)	0.978	0.768 (0.086 - 6.754)	0.687
No	2.853 (1.057 - 7.702)		0.130 (0.043 - 0.396)		0.238 (0.075 - 0.747)	

* Adjusting for confounding variables including age, sex, BMI, WHR, duration of diabetes, total bilirubin, and blood glucose.

consumption, or fatty liver status and the association between exosomal lncRNAs and DR. LncRNA DLX6-AS1 and lncRNA PRINS had predictive value for the diagnosis of DR.

Conclusion

A deep forest-based framework integrating multi-source medical data was developed to predict lncRNA-diabetes associations. By combining GWAS information, gene co-expression network analysis, and network representation learning, the model enabled systematic identification of candidate lncRNAs related to type 2 diabetes and its complications. Comparative genomic and interaction network analyses further supported the potential genetic linkage between obesity-related loci and diabetes. In addition, several exosomal lncRNAs were identified as significantly associated with diabetic retinopathy. The proposed framework provided a robust computational strategy for exploring lncRNA-disease relationships in complex metabolic disorders. The model might facilitate the prioritization of candidate lncRNAs for subsequent experimental validation. Future studies incorporating expanded datasets and biological validation will help refine the predictive performance and clarify the functional roles of the identified lncRNAs.

Acknowledgements

This research was supported by the Nantong Natural Science Foundation and Social and Livelihood Science and Technology Program Project (Grant No. MSZ2024122), Doctoral Research Startup Fund Project of Nantong Institute of Technology (Grant No. 2025XKB29), Jiangsu Province Higher Education Informatisation Research Major Project (Grant No. 2025JSETKT036), Nantong Institute of Technology Artificial Intelligence General Education Teaching Reform Research Special Project (Grant No. 2025JJG005), Nantong Institute of Technology Master's Construction Project of Electronic Information (Grant No. 879002), Nantong Institute of Technology Higher Education Teaching Reform Research Project (Grant No. 2025JJG042).

References

1. Lipovich L, Johnson R, Lin CY. 2010. MacroRNA underdogs in a microRNA world: Evolutionary, regulatory, and biomedical significance of mammalian long non-protein-coding RNA. BBA-Gene Regul Mech. 1799(9):597-615.

2. Jin T. 2021. LncRNA DRAIR is a novel prognostic and diagnostic biomarker for gastric cancer. Mamm Genome. 32(6):503-507.
3. Reddy MA, Amaram V, Das S, Tanwar VS, Ganguly R, Wang M, *et al*. 2021. lncRNA DRAIR is downregulated in diabetic monocytes and modulates the inflammatory phenotype *via* epigenetic mechanisms. JCI Insight. 6(11):e143289.
4. Li G, Kryczek I, Nam J, Li X, Li S, Li J, *et al*. 2021. LIMIT is an immunogenic lncRNA in cancer immunity and immunotherapy. Nat Cell Biol. 23(5):526-537.
5. Lu X, Tan Q, Ma J, Zhang J, Yu P. 2022. Emerging role of LncRNA regulation for NLRP3 inflammasome in diabetes complications. Front Cell Dev Biol. 9:792401.
6. Yan J, Wang R, Tan J. 2023. Recent advances in predicting lncRNA-disease associations based on computational methods. Drug Discov Today. 28(2):103432.
7. Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, *et al*. 2012. LncRNADisease: A database for long-non-coding RNA-associated diseases. Nucleic Acids Res. 41(D1):D983-D986.
8. Ning S, Zhang J, Wang P, Zhi H, Wang J, Liu Y, *et al*. 2016. Lnc2Cancer: A manually curated database of experimentally supported lncRNAs associated with various human cancers. Nucleic Acids Res. 44(D1):D980-D985.
9. Asim MN, Ibrahim MA, Asif T, Dengel A. 2025. RNA sequence analysis landscape: A comprehensive review of task types, databases, datasets, word embedding methods, and language models. Heliyon. 11(2):e41488.
10. Yang X, Gao L, Guo X, Shi X, Wu H, Song F, *et al*. 2014. A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases. Plos One. 9(1):e87797.
11. Fu G, Wang J, Domeniconi C, Yu G. 2018. Matrix factorization-based data fusion for the prediction of lncRNA-disease associations. Bioinformatics. 34(9):1529-1537.
12. Lan W, Lai D, Chen Q, Wu X, Chen B, Liu J, *et al*. 2020. LDICDL: LncRNA-disease association identification based on collaborative deep learning. IEEE/ACM Trans Comput Biol Bioinform. 19(3):1715-1723.
13. Lan W, Wu X, Chen Q, Peng W, Wang J, Chen YP. 2022. GANLDA: Graph attention network for lncRNA-disease associations prediction. Neurocomputing. 469:384-393.
14. Yang Q, Li X. 2021. BiGAN: LncRNA-disease association prediction based on bidirectional generative adversarial network. BMC Bioinformatics. 22:1-17.
15. Zeng M, Lu C, Fei Z, Wu FX, Li Y, Wang J, *et al*. 2020. DMFLDA: A deep learning framework for predicting lncRNA-disease associations. IEEE/ACM Trans Comput Biol Bioinform. 18(6):2353-2363.
16. Yao D, Zhan X, Zhan X, Kwoh CK, Li P, Wang J. 2020. A random forest based computational model for predicting novel lncRNA-disease associations. BMC Bioinformatics. 21:1-18.
17. Baião AR, Cai Z, Poulos RC, Robinson PJ, Reddel RR, Zhong Q, *et al*. 2025. A technical review of multi-omics data integration methods: From classical statistical to deep generative approaches. Brief Bioinform. 26(4):bbaf355.
18. Alqudah AM, Moussavi Z. 2025. A review of deep learning for biomedical signals: Current applications, advancements, future prospects, interpretation, and challenges. Comput Mater Contin. 83(3):1-89.
19. Li Y, Zhang Q, Liu Z, Wang C, Han S, Ma Q, *et al*. 2021. Deep forest ensemble learning for classification of alignments of non-coding RNA sequences based on multi-view structure representations. Brief Bioinform. 22(4):bbaa354.
20. Ji S, Wu J, An F, Lou M, Zhang T, Guo J, *et al*. 2025. Umami-gcForest: Construction of a predictive model for umami peptides based on deep forest. Food Chem. 464:141826.