OXFORD

DLRAPom: a hybrid pipeline of Optimized XGBoost-guided integrative multiomics analysis for identifying targetable disease-related lncRNA–miRNA–mRNA regulatory axes

Chen Shen[†], Huiyu Li[†], Miao Li[†], Yu Niu[†], Jing Liu, Li Zhu, Hongsheng Gui, Wei Han, Huiying Wang, Wenpei Zhang, Xiaochen Wang,

Xiao Luo, Yu Sun, Jiangwei Yan and Fanglin Guan (iD)

Corresponding authors: Fanglin Guan, Key Laboratory of National Ministry of Health for Forensic Sciences, School of Medicine & Forensics, Xi'an Jiaotong University, Xi'an 710061, China. Tel.: +86-29-82655117; Fax: +86-29-82655472; E-mail: fanglingguan@163.com; Jiangwei Yan, Department of Genetics, School of Medicine & Forensics, Shanxi Medical University, Taiyuan 030009, China. Tel.: +86-351-3985097; Fax: +86-351-3985097; E-mail: yanjw@sxmu.edu.cn; Yu Sun, Department of Endocrinology and Metabolism, Qilu Hospital of Shandong University, Ji'nan 250012, China. Tel.: +86-531-82169114; Fax: +86-531-82169114; E-mail: sunyujn@aliyun.com; Xiao Luo, Department of Physiology and Pathophysiology, School of Basic Medical Sciences, Xi'an Jiaotong University, Xi'an 710061, China. Tel.: +86-29-82657490; Fax: +86-29-82657490; E-mail: xluo@xjtu.edu.cn
[†]These authors contributed equally to this work.

Abstract

The lack of a reliable and easy-to-operate screening pipeline for disease-related noncoding RNA regulatory axis is a problem that needs to be solved urgently. To address this, we designed a hybrid pipeline, disease-related lncRNA–miRNA–mRNA regulatory axis prediction from multiomics (DLRAPom), to identify risk biomarkers and disease-related lncRNA–miRNA–mRNA regulatory axes by adding a novel machine learning model on the basis of conventional analysis and combining experimental validation. The pipeline consists of four parts, including selecting hub biomarkers by conventional bioinformatics analysis, discovering the most essential protein-coding biomarkers by a novel machine learning model, extracting the key lncRNA–miRNA–mRNA axis and validating experimentally. Our study is the first one to propose a new pipeline predicting the interactions between lncRNA and miRNA and mRNA by combining WGCNA and XGBoost. Compared with the methods reported previously, we developed an Optimized XGBoost model to reduce the degree of overfitting in multiomics data, thereby improving the generalization ability of the overall model for the integrated analysis of multiomics data. With applications to gestational diabetes mellitus (GDM), we predicted nine risk protein-coding biomarkers and some potential lncRNA–miRNA–mRNA regulatory axes, which all correlated with GDM. In those regulatory axes, the MALAT1/hsa-miR-144-3p/IRS1 axis was predicted to be the key axis and was identified as being associated with GDM for the first time. In short,

Chen Shen is currently working as a postgraduate student at the Shanghai Key Laboratory of Forensic Medicine, Academy of Forensic Science; Key Laboratory of National Ministry of Health for Forensic Sciences, School of Medicine & Forensics, Health Science Center, Xi'an Jiaotong University, Xi'an, China.

Huiyu Li is currently working as a postgraduate student at the Key Laboratory of National Ministry of Health for Forensic Sciences, School of Medicine & Forensics, Health Science Center, Xi'an Jiaotong University, Xi'an, China.

Miao Li is currently working as an associate professor at the Department of Ultrasound, the Second Affiliated Hospital, Xi'an Jiaotong University, Xi'an, China.

Yu Niu is currently working as an associate professor at the Department of Endocrinology and Metabolism, Ninth Hospital of Xi'an City, Xi'an, China.

Jing Liu is currently working as a postgraduate student at the Department of Physiology and Pathophysiology, School of Basic Medical Sciences, Health Science Center, Xi'an Jiaotong University, Xi'an, China.

Li Zhu is currently working as a lecturer at the Key Laboratory of National Ministry of Health for Forensic Sciences, School of Medicine & Forensics, Health Science Center, Xi'an Jiaotong University, Xi'an, China.

Hongsheng Gui is currently working as an associate professor at the Center for Behavior Health and Psychiatry Research, Henry Ford Health System, Detroit, MI, USA.

Wei Han is currently working as an associate professor at the Key Laboratory of National Ministry of Health for Forensic Sciences, School of Medicine & Forensics, Health Science Center, Xi'an Jiaotong University, Xi'an, China.

Huiying Wang is currently working as a postgraduate student at the Key Laboratory of National Ministry of Health for Forensic Sciences, School of Medicine & Forensics, Health Science Center, Xi'an Jiaotong University, Xi'an, China.

Wenpei Zhang is currently working as a postgraduate student at the Key Laboratory of National Ministry of Health for Forensic Sciences, School of Medicine & Forensics, Health Science Center, Xi'an Jiaotong University, Xi'an, China.

Xiaochen Wang is currently working as a undergraduate student at the Key Laboratory of National Ministry of Health for Forensic Sciences, School of Medicine & Forensics, Health Science Center, Xi'an Jiaotong University, Xi'an, China.

Xiao Luo is currently working as an associate professor at the Department of Physiology and Pathophysiology, School of Basic Medical Sciences, Health Science Center, Xi'an Jiaotong University, Xi'an, China.

Yu Sun is currently working as a professor at the Department of Endocrinology and Metabolism, Qilu Hospital of Shandong University, Ji'nan, China.

Jiangwei Yan is currently working as a professor at the Department of Genetics, School of Medicine & Forensics, Shanxi Medical University, Taiyuan, China.

Fanglin Guan is currently working as a professor at the Shanghai Key Laboratory of Forensic Medicine, Academy of Forensic Science; Key Laboratory of National Ministry of Health for Forensic Sciences, School of Medicine & Forensics, Health Science Center, Xi'an Jiaotong University, Xi'an, China.

as a flexible pipeline, DLRAPom can contribute to molecular pathogenesis research of diseases, effectively predicting potential disease-related noncoding RNA regulatory networks and providing promising candidates for functional research on disease pathogenesis.

Keywords: analysis framework, data integration, multiomics analysis, noncoding RNA regulatory

Introduction

Gene expression processes are influenced by the precise regulation and complex interactions of multiple dimensions, including methylation [1], genetic mutations [2], transcription factors [3] and noncoding RNAs [4]. These multiple levels of regulatory networks highlight multiomics integration as an important method for characterizing complex biological mechanisms underlying phenotypes. Currently, the rapid development of high-throughput sequencing technologies [5] and the emergence of new technologies in multiomics [6, 7] have boosted the generation of large amounts of multiomics data. Compared to the limitations of individual omics data in elucidating the biological mechanisms of diseases, multiomics data have more powerful complementary effects and are challenging because they not only include multiomics datasets with diverse data types of different unique characteristics and distributions [8] but also involve proficiency in applying the most advanced and appropriate machine learning methods to uncover the complex relationships between different dimensions of molecules [9]. Despite numerous attempts by researchers to address these issues [10–12], the lack of effective integrated analysis methods remains a fatal bottleneck to the interpretation of biological data and the translation of basic research. Therefore, constructing an effective integrated analysis method for the interpretation of multiomics data and its transformation is an urgent matter.

Previous studies have indicated that RNA–protein interactions regulate gene expression through the control of various posttranscriptional processes, which in turn influence disease development directly or indirectly [13]. The dysregulation of noncoding RNAs, in particular microRNAs (miRNAs) and long noncoding RNAs (lncRNAs), is closely associated with various diseases [14]. miRNAs are found to directly or indirectly affect the development of diseases [15], and functional genomics studies have shown that lncRNAs are an important regulator in a variety of biological processes and disease development, partly through interaction with miRNAs or mRNAs. [16, 17]. Given the mechanisms by which lncRNAs regulate genes and the relationships between miRNA-targeted genes and diseases, it would be desirable to obtain more information on the lncRNA–miRNA–mRNA regulatory axes associated with diseases for more references and evidence for the elucidation of the disease molecular mechanisms [18]. Although a number of approaches have been developed for the prediction of disease-related ncRNAs, such as RWR [19], RWRHLD [20], LncRDNetFlow [21] and LncPriCNet [22],

there have been no reports about the methods or tools combined with experimental validation to identify the lncRNA–miRNA–mRNA network as a whole functional module. The lack of a reliable and easy-to-operate screening pipeline for disease-related noncoding RNA regulatory axis is a problem that needs to be solved urgently.

In this study, inspired by the currently well-performing extreme gradient boosting (XGBoost) model, we developed a first hybrid pipeline of integrative multiomics analysis for identifying targetable disease-related lncRNA–miRNA–mRNA regulatory axes based on a novel Optimized XGBoost model. The new pipeline of disease-related lncRNA–miRNA–mRNA regulatory axis prediction from multiomics (DLRAPom) added a novel Optimized XGBoost model on the basis of conventional weighted gene coexpression network analysis (WGCNA) analysis and combined with experimental verification to ensure the accuracy of extracting regulatory features from the ncRNA gene–disease association network. Compared with the methods that have been reported previously, we developed an Optimized XGBoost model to reduce the degree of overfitting in multiomics data, thereby improving the generalization ability of the overall model for the integrated analysis of multiomics data. Taking gestational diabetes mellitus (GDM) as an example, we utilized DLRAPom to evaluate the lncRNA–miRNA–mRNA regulatory network of GDM to reveal the value and reliability of the hybrid pipeline.

Materials and methods
Overall pipeline design

There are four steps in the DLRAPom pipeline: selecting hub biomarkers by conventional bioinformatics analysis, discovering the most essential protein-coding biomarkers by a novel machine learning model, extracting the key lncRNA–miRNA–mRNA axes and validating experimentally (Figure 1). To ensure the accuracy of the results, we first experimentally verified the results of key nodes (lncRNAs, miRNAs and mRNAs) by quantitative reverse transcription PCR (RT–qPCR), and then obtained supportive evidence for the pairwise target relationships within the predicted regulatory axes through literature search or dual-luciferase reporter assay. Through a reasonable combination of multiple platforms and methods to analyze disease-related multiomics data, we obtained reliable disease-related lncRNA–miRNA–mRNA regulatory axes for the further investigation of disease mechanisms. Next, we introduced in detail the specific methods of each part of the process, taking GDM as an example. All related analysis scripts were uploaded to GitHub (https://github.com/shenxiaochenn/DLRAPom).
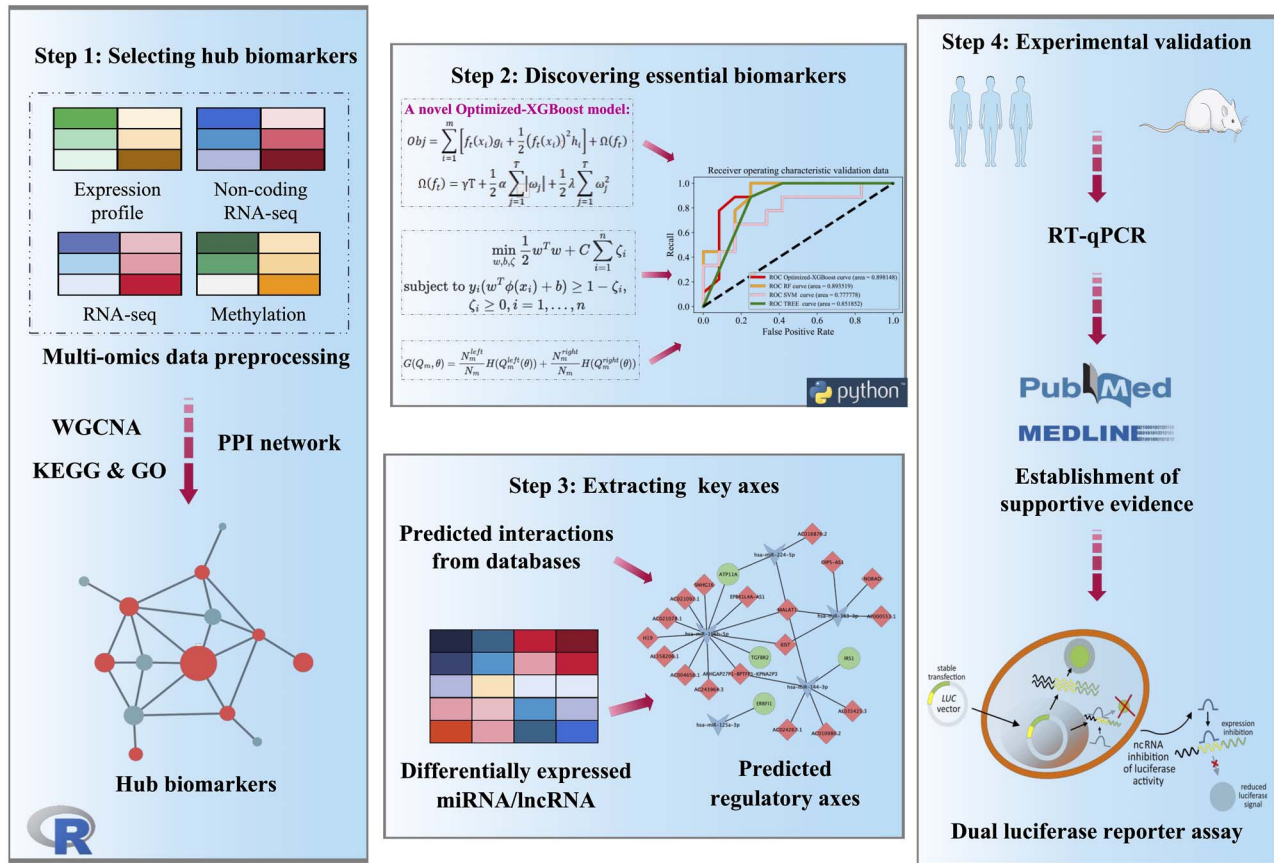
Figure 1. The DLRAPom pipeline for identifying targetable disease-related lncRNA–miRNA–mRNA regulatory axes by machine learning guided integrative multiomics analysis. The pipeline consists of four steps, namely selecting hub biomarkers by conventional bioinformatics analysis, discovering the most essential protein-coding biomarkers by a novel machine learning model, extracting the key lncRNA–miRNA–mRNA axis, and validating experimentally. In Step1, hub biomarkers are selected by conventional bioinformatics analyses. These results are used as the inputs of Step2 to discover essential protein-coding biomarkers and obtain the importance of each protein-coding biomarker. In Step3, the competing endogenous network is constructed based on the obtained information of lncRNA–miRNA and miRNA–target gene by databases. Among all the constructed regulatory axes, the regulatory axes containing the predicted risk protein-coding biomarkers in the novel Optimized XGBoost model are selected as the main outcomes of our pipeline and would be used for subsequent experimental verification. If there are multiple regulatory axes, the criticality of the regulatory axes is ranked in descending order according to the importance of the predicted protein-coding biomarker included in each axis. After the significant expression change of each RNA molecule in the predicted regulatory axes is confirmed, further supportive evidence for the pairwise targeting relationships within the predicted regulatory axes was required. If these targeting relationships have not been reported before, we need to determine whether these targeting relationships exist through the dual-luciferase reporter assay. For a predicted regulatory axis, only when the biological targeting relationships of lncRNA–miRNA and miRNA–mRNA have been both experimentally verified, can the predicted regulatory axis be considered to be targetable and reliable.

Selection of hub biomarkers by bioinformatics analysis

Discovery of differentially expressed biomarkers in multiomics datasets

Taking GDM as an example, we downloaded RNA-sequencing data (GSE154377 [23], GSE150621 [24]), mRNA expression profiling microarray data (GSE87295), DNA methylation microarray data (GSE88929 [25]) and noncoding RNA profiling sequencing data (GSE112168 [26]) from the Gene Expression Omnibus database, which were preprocessed to retain suitable data for differentially expressed genes (DEGs), methylated genes and miRNAs. According to the results of the t-SNE algorithm and correlation matrix analysis, samples with significant differences between the GDM group and the control group were retained. Thus, from the 134 samples in GSE154377, 49 disease-related samples were selected, including 32 GDM samples and 17 control samples.

GSE112168 included six GDM samples and six control samples. A total of eight samples from GSE150621 were screened, consisting of five GDM samples and three control samples. Differentially expressed miRNAs and DEGs were screened using the DESeq2 package in R statistical software. The R-script for DESeq2 analysis has been uploaded to the GitHub (https://github.com/shenxiaochenn/DLRAPom). Using the limma package [27] in R statistical software, seven samples in GSE87295 were retained, including five GDM samples and two control samples. Among the two datasets (A and B) of GSE88929, we selected dataset B with more samples, including 23 GDM samples and 45 control samples. The ChAMP package [28] in R statistical software was used to screen differentially methylated positions. WGCNA is a method to find coexpressed gene modules by exploring the association of gene networks with phenotypes and the core genes in the network [29]. The WGCNA package

in R statistical software was used to perform gene expression matrix and coexpression analyses based on GSE154377 data (meeting the minimum sample size requirements of WGCNA) to extract coexpressed genes in disease-related modules. The differentially expressed biomarkers and the coexpressed genes in WGCNA were divided into two groups, an upregulated group and a downregulated group, via the ggVennDiagram package in R statistical software.

Enrichment analysis

Gene ontology (GO) enrichment analysis explores and characterizes the functions of genes from three aspects: cellular component, biological process and molecular function. Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database resource for large-scale molecular databases to better understand the high-level functions and utilities of biological systems [30]. The clusterProfiler package in R statistical software was used to perform and determine the enrichment pathways in biological process, and $P < 0.05$ was considered to be statistically significant. Based on the results of GO enrichment analysis, disease-related metabolic pathways were selected and plotted using the ClueGO plugin (http://apps.cytoscape.org/apps/cluego) in Cytoscape software.

Construction of the protein–protein interaction network

Based on the disease-related metabolic pathways, the protein–protein (PPI) network was constructed using the STRING database (https://string-db.org/), and then the key nodes were screened using the CentiScaPe2.2 plugin (http://apps.cytoscape.org/apps/centiscape) in Cytoscape software with the default criteria (degree = 5.555, centrality = 57.999).

Discovering the most essential protein-coding biomarker by a novel Optimized XGBoost algorithm

After screening out the hub genes, three datasets including expression data (GSE87295, GSE154377, GSE150621) were combined and then divided into 80% train set and 20% validation set by the random split algorithm. Furthermore, a novel machine learning integration classification algorithm Optimized XGBoost was designed by us, and the importance of each gene in the Optimized XGBoost model was evaluated by a comprehensive approach of the weight (the number of times one biomarker was used to split the data across all trees), gain (the average gain of the biomarker when it was used in trees) and cover (the average coverage of the biomarker when it was used in trees). In the model, the objective function was identified as Equation (1). In this objective function, g_i and h_i were the first and second derivatives, respectively.

$$Obj = \sum_{i=1}^{m} \left[f_t(x_i)\, g_i + \frac{1}{2}(f_t(x_i))^2 h_i \right] + \Omega(f_t) \qquad (1)$$

$$\Omega(f_t) = \gamma T + \frac{1}{2}\alpha \sum_{j=1}^{T} |\omega_j| + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \qquad (2)$$

The complexity of the model was estimated by the $\Omega(f_t)$ (Equation (2)). T was identified as the amount of leaf nodes. Meanwhile, L1 regularization and L2 regularization were utilized to control the complexity of the model. The detailed explanations for both formulae could be found in the Supplementary Document.

Another three machine learning models were also used to validate the novel Optimized XGBoost model, including support vector machine (SVM) [31], random forest (RF) [32] and decision tree from scikit-learn (github.com/scikit-learn/scikit-learn). The area under the computing receiver-operating characteristic (ROC) curve (AUC) and accuracy were calculated to estimate the different machine learning models. The most vital biomarker in the novel Optimized XGBoost model was determined by the importance of each protein-coding biomarker. All source codes used in this step were uploaded to GitHub (https://github.com/shenxiaochenn/DLRAPom).

Extracting the key lncRNA–miRNA–mRNA axes

Differentially expressed miRNAs/lncRNAs were screened using the DESeq2 package in R statistical software. If there were no differentially expressed lncRNA data, the StarBase database version 2.0 (https://starbase.sysu.edu.cn/starbase2/index.php) was used to search for lncRNAs that may regulate differentially expressed miRNAs. Then, the differentially expressed miRNAs/lncRNAs were utilized to construct the competing endogenous network based on the predicted interacting relationships by StarBase database Version 2.0 (the standard: clipExpNum > 10).

Using the miRWalk database (http://mirwalk.uni-hd.de/), target genes of differentially expressed miRNAs were predicted with the following criteria: P value = 0.01, 'positions' = 3UTR and TargetScan database or miRDB database = 1. The hub genes in the PPI network intersected with target genes to obtain disease-related genes. Finally, the lncRNA-miRNA and miRNA-target gene networks were combined to construct a competing endogenous network in Cytoscape. Among all the constructed regulatory axes, the regulatory axes containing the predicted risk protein-coding biomarkers in the novel Optimized XGBoost model are selected as the main outcomes of our pipeline and can be forwarded for subsequent experimental verification. If there are multiple regulatory axes, the criticality of the regulatory axes is ranked in descending order according to the importance of the predicted gene included in each axis.

Experimental validation

In population samples or animal models, blood or disease-related tissue samples are collected, and the three RNA molecules of the predicted key regulatory axes

are experimentally verified. Only when the expression change of each RNA molecule in the predicted regulatory axes has a statistically significant difference, are the predicted regulatory axes considered to have been initially verified. Considering that the placenta, as a vital link between pregnancy and offspring, plays an important role in the pathogenesis of GDM, placental tissues from five GDM patients and five normal glucose tolerance controls were recruited to perform RT–qPCR verification of the predicted key regulatory axes. All participants underwent an oral glucose tolerance test at 24–28 weeks of gestation. GDM was diagnosed according to American Diabetes Association 2011 guidelines [33]. Five pregnant women with normal glucose tolerance were recruited as a normal control group. Individuals with the following conditions were excluded from the study: multiple pregnancy, maternal–fetal blood type incompatibility, abnormal placenta or umbilical cord, inflammation, severe diabetes, hypertension, other pregnancy diseases, malignant tumors or other serious organic diseases.

Placentas of all participants were collected after delivery and frozen in liquid nitrogen immediately for RNA extraction. Subjects' characteristics are shown in Supplementary Table S1 available online at http://bib.oxfordjournals.org/. Data are presented as mean ± SD. Differences were tested by Student's *t*-tests. The significance threshold was set at $P < 0.05$. Total RNA from placental tissue was isolated using TRIzol reagent (Invitrogen, San Diego, CA, USA), and 1 μg of total RNA from each sample was reverse-transcribed to cDNA using a commercial RT–PCR kit (Thermo Scientific, Waltham, MA, USA) according to the manufacturer's instructions. RT–qPCR was performed as previously described [34]. Gene expression changes were calculated with the $2^{-\Delta\Delta Ct}$ method relative to the *YWHAZ* housekeeping gene and standardized to the control group. All primers were synthesized by AUGCT (Beijing, China). Sequences are shown in Supplementary Table S2 available online at http://bib.oxfordjournals.org/. This study was conducted in accordance with the ethical guidelines of the Declaration of Helsinki (version 2002) and was approved by the Ethics Committee of the First Affiliated Hospital of Xi'an Jiaotong University (XJTU1AF2020LSK-275). All participants provided written informed consent forms.

After the significant expression change of each RNA molecule in the predicted regulatory axes was confirmed, further supportive evidence for the pairwise targeting relationships within the predicted regulatory axes were required. If these targeting relationships have not been reported before, we need to determine whether these targeting relationships exist through the dual-luciferase reporter assay. For a predicted regulatory axis, only when the biological targeting relationships of lncRNA-miRNA and miRNA-mRNA have been both experimentally verified, can the predicted regulatory axis be considered to be targetable and reliable.

Results
Significant differences in miRNA and mRNA expression and gene coexpression modules

GDM-related multiomics datasets (GSE154377, GSE150621, GSE87295, GSE88929 and GSE112168) were used to analyze the core gene sets and pathways. Given that the GSE154377 dataset was an expression profile obtained by high-throughput sequencing, dimension reduction and cluster analyses of the GSE154377 dataset were performed by the t-SNE algorithm (Supplementary Figure S1A available online at http://bib.oxfordjournals.org/). The heatmap and volcano plots of DEGs are shown in Supplementary Figure S1B and C available online at http://bib.oxfordjournals.org/. The GSE150621 dataset was also used for expression profiling by high-throughput sequencing. Thus, cluster analysis of the GSE150621 dataset was conducted with a cluster heatmap plot (Supplementary Figure S1D available online at http://bib.oxfordjournals.org/), and its heatmap and volcano plots of DEGs are shown in Supplementary Figure S1E and F available online at http://bib.oxfordjournals.org/. The cluster analysis of the GSE87295 dataset, an expression profiling by array, was carried out with a cluster heatmap plot (Supplementary Figure S1G available online at http://bib.oxfordjournals.org/), and the corresponding heatmap and volcano plots of DEGs are shown in Supplementary Figure S1H and I available online at http://bib.oxfordjournals.org/. The GSE88929 dataset was subjected to methylation profiling by a genome tiling array, and the volcano plot of differentially methylated genes is shown in Supplementary Figure S1J available online at http://bib.oxfordjournals.org/. The top eight differentially methylated genes are labeled in Supplementary Figure S1J available online at http://bib.oxfordjournals.org/. The GSE112168 dataset, which contains noncoding miRNA profiling by high-throughput sequencing, had 25 differentially expressed miRNAs, which are listed in Supplementary Table S3 available online at http://bib.oxfordjournals.org/. There were 856 DEGs in the GSE154377 dataset, 1174 DEGs in the GSE150621 dataset, 726 DEGs in the GSE87295 dataset and 1869 differentially methylated genes in the GSE88929 dataset. Considering that the minimum sample size in stable WGCNA must be greater than 15 in a single group, the GSE154377 dataset that contained 49 samples (32 GDM patients and 17 controls) was utilized to establish the WGCNA network. Through the analysis of the relationships between pairwise gene coexpression modules and eigengenes, the gene expression was comparatively independent between modules (Supplementary Figure S2 available online at http://bib.oxfordjournals.org/). The formation communication of eigengenes was estimated, and the most significant associations of key modules were identified between the GDM group and the control group. The green–yellow module (Supplementary Figure S2D and F available online at http://bib.oxfordjournals.org/) was mostly positively correlated
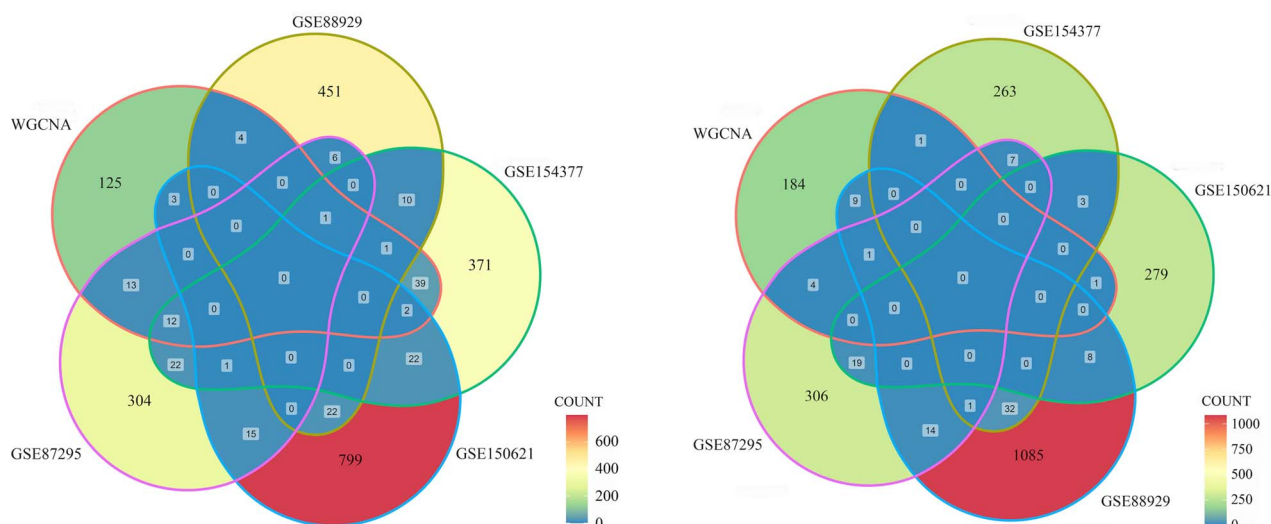
Figure 2. The Venn-diagram of upregulated and downregulated genes in different datasets. (**A**) The intersection of upregulated genes in the five data sets (GSE154377, GSE150621, GSE87295, GSE88292 and WGCNA). (**B**) The intersection of downregulated genes in the five data sets. Redder colors indicate more genes while bluer colors indicate fewer genes. WGCNA: weighted gene coexpression network analysis.

with GDM. Furthermore, the topological overlap matrix plot is shown in Supplementary Figure S2G available online at http://bib.oxfordjournals.org/. There were 173 genes screened from the upregulated group (Figure 2A), and 100 genes were screened from the downregulated group (Figure 2B). A total of 273 key genes were utilized in the subsequent enrichment analysis.

Analyses of GO and KEGG enrichment and construction of the PPI network

The hub pathways in each group are colored in the center of the circle, and other metabolic disease-related pathways are sorted around the hub pathways (Supplementary Figure S3 available online at http://bib.oxfordjournals.org/). The genes selected from the hub pathways were utilized to construct the PPI network. The key points in the PPI network were screened by the degree centrality and betweenness centrality algorithm and are labeled in green in Figure 3. The nine key genes were selected in the PPI network (Figure 3) constructed in Cytoscape software based on the default criteria (degree = 5.555, centrality = 57.999).

Establishment of an Optimized XGBoost model and discovery of the most essential protein-coding biomarker

The nine selected key genes in the PPI network were utilized to construct the multiple machine learning models. The three models of SVM, RF and decision tree were used for horizontal comparison with the Optimized XGBoost model, of which the aim was to demonstrate that the Optimized XGBoost model had the best performance. The original boosting model showed slight overfitting, which is shown by the red and yellow lines (Figure 4A). The Optimized XGBoost machine learning model is shown by the green and blue lines (Figure 4A) through applying 3-fold cross-validation by the function of xgboost.cv to

discover the best parameters and hyperparameters, the details of which could be found in our codes in the GitHub (https://github.com/shenxiaochenn/DLRAPom). We preformed 3-fold, 5-fold and 10-fold cross-validation, and the results were presented in Supplementary Figure S4 available online at http://bib.oxfordjournals.org/. As indicated in Supplementary Figure S4, available online at http://bib.oxfordjournals.org/, these results were not significantly different in our datasets. However, in a small sample size data set, these results will show obvious differences, which will have potential impacts on the prediction results generated by the model. Thus, given the sample size of data sets used by users and the universality of our pipeline, in order to reduce the error of high-fold cross-validation in small sample data sets, we applied 3-fold cross-validation in the pipeline, which is also the default parameter of the function of xgboost.cv.

The degree of overfitting was released by sacrificing the accuracy in the train data. As shown in Figure 4B, the importance of each key gene in the Optimized XGBoost model is presented. The *IRS1* gene was more essential than other genes. The AUCs of the ROC curve were 0.940 in the train data and 0.898 in the validation data (Figure 4C and D). For the SVM, RF and decision tree models (their specific parameters presented in Supplementary) in the validation data, the AUCs were 0.778, 0.894 and 0.852, respectively (Figure 4D). Compared with the Optimized XGBoost model with an accuracy of 80.95%, the accuracies of the SVM, RF and decision tree models in the validation data were 71.43%, 71.43% and 76.19%, respectively (presented in the source_code.ipynb script). All related source codes were uploaded to GitHub (https://github.com/shenxiaochenn/DLRAPom). Combined with the accuracy and AUC of the four models, the Optimized XGBoost model developed by us performed the best in related analyses.
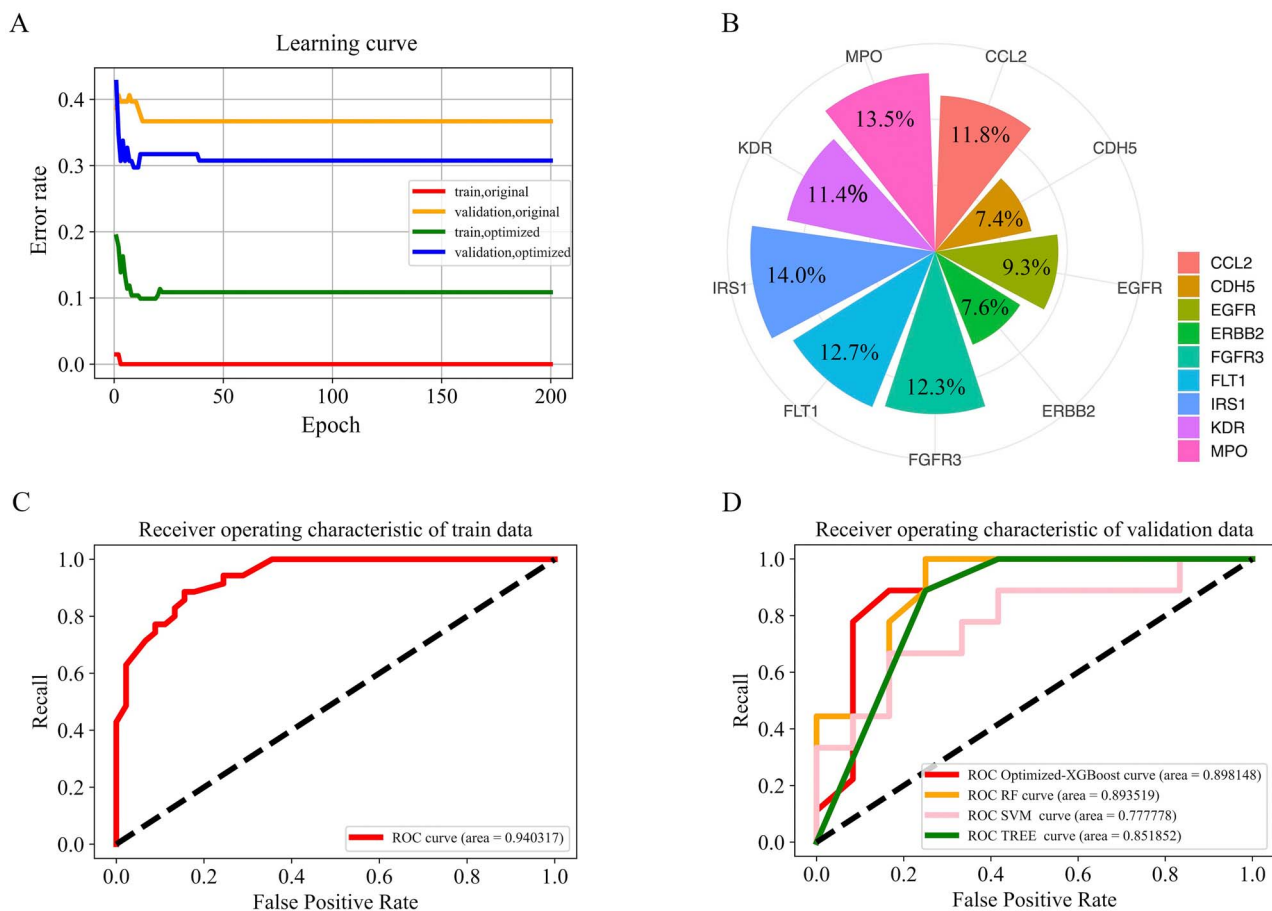
Figure 3. Construction of protein–protein interaction network. Circles represent genes and lines represent interactions. Green points represent predicted key genes in the whole network.

Prediction of the key lncRNA–miRNA–mRNA axes

The key genes in the PPI network, the differentially expressed miRNAs and lncRNAs (the predicted lncRNAs using starBase database Version 2.0) were utilized to establish the competing endogenous network (Figure 5). As shown in Figure 5, only one predicted protein-coding biomarker (*IRS1*) was included in the potential lncRNA–miRNA–mRNA axes. In the analyses based on multiple machine learning algorithms, the *IRS1* gene was the most vital gene according to its importance in the optimal model of Optimized XGBoost. Additionally, the enrichment analysis also regulated some GDM-related pathways. By analyzing the competing endogenous network (Figure 5), the *IRS1* gene was only linked with hsa-miR-144-3p, and hsa-miR-144-3p was shown to interact with the most active lncRNA, MALAT1. Therefore, the MALAT1/hsa-miR-144-3p/*IRS1* regulatory axis could be considered to be a key axis in GDM.

Validation of the key MALAT1/hsa-miR-144-3p/*IRS1* axis

To investigate the reliability of the predicted key MALAT1/hsa-miR-144-3p/*IRS1* regulatory axis and the Optimized XGBoost model developed by us, we examined the expression pattern of the key regulatory axis in the placenta by RT–qPCR. As shown in Figure 6, the

expression levels of the *IRS1* gene and hsa-miR-144-3p were both lower in placenta from GDM patients than in placenta from controls, while MALAT1 was highly expressed in GDM patients. All the *P* values (*t*-tests) were less than 0.05 (Figure 6). Furthermore, when searching the literature to obtain supportive evidence for the target interactions of MALAT1/hsa-miR-144-3p and hsa-miR-144-3p/*IRS1* in the key regulatory axis, we found in recently published articles that their biological targeting relationships were both confirmed by the dual-luciferase reporter assay [35–38]. However, the relationship of this key regulatory axis with GDM has not yet been reported. Therefore, given that biological targeting relationships of MALAT1/hsa-miR-144-3p and hsa-miR-144-3p/*IRS1* have no disease or cell specificity, the predicted key MALAT1/hsa-miR-144-3p/*IRS1* regulatory axis can be considered to be targetable and reliable, which also confirmed the reliability of our DLRAPom pipeline and the accuracy of the Optimized XGBoost machine learning model developed by us.

Discussion

Most of the previously reported methods focused on one or two types of noncoding RNA, and there were few predictions combined with experimental validation to identify the lncRNA–miRNA–mRNA network as a whole

Figure 4. Evaluation by multiple machine learning algorithms. (**A**) The X-axis reflects the times of iteration and Y-axis reflects the error rate. The red and yellow lines reflect the results of the original model. The green and blue lines reflect the results of optimized model. (**B**) The importance of genes in the model. The bigger the area, the more vital the genes. (**C**) The receiver-operating characteristic (ROC) curve of the train set. (**D**) ROC curves of the validation set by different machine learning algorithms.

functional module. The lack of a reliable and easy-to-operate screening pipeline for disease-related noncoding RNA regulatory axis is a problem that needs to be solved urgently. To our knowledge, our study is the first one to propose a new pipeline predicting the interactions between lncRNA and miRNA and mRNA by combining WGCNA and XGBoost with experimental verification to improve the accuracy of prediction results. Compared with the methods reported previously, we developed an Optimized XGBoost model to reduce the degree of overfitting in multiomics data, thereby improving the generalization ability of the overall model for the integrated analysis of multiomics data. We took GDM as an example to introduce in detail the application of the DLRAPom pipeline to identify targetable disease-related lncRNA–miRNA–mRNA regulatory axes. The characteristics of competing endogenous networks of GDM were depicted well. Importantly, a potential key GDM-related MALAT1/hsa-miR-144-3p/IRS1 regulatory axis was captured, and it was further validated by relevant experiments. To our knowledge, this key GDM-related regulatory axis was identified for the first time. The results of the stability test indicate that our machine learning-guided multiomics integrative analysis pipeline

is robust and reliable. Additionally, DLRAPom can effectively predict potential disease-related lncRNA–miRNA–mRNA regulatory networks and provide more promising candidates for functional research on disease pathogenesis.

One of the prominent advantages of the DLRAPom pipeline is that it fully combines conventional bioinformatics analysis methods and machine learning models, hence providing more biological mechanism-targeted candidates for downstream experimental verification. In terms of these methods, models and experiments alone, they are very popular in various disease research fields [39]. In fact, in addition to disease-related lncRNA–miRNA–mRNA regulatory axes, other essential factors, such as disease-related metabolic pathways, the machine learning model for accurate assessment of disease-related risk biomarkers deserves our attention as well. Flexibility is also a major advantage of the DLRAPom pipeline. Four machine learning algorithms were applied: three traditional machine learning algorithms (SVM, RF and decision tree) and one novel machine learning algorithm (Optimized XGBoost). The advantage of the model is that it can deal with datasets that include missing values. If a certain risk biomarker
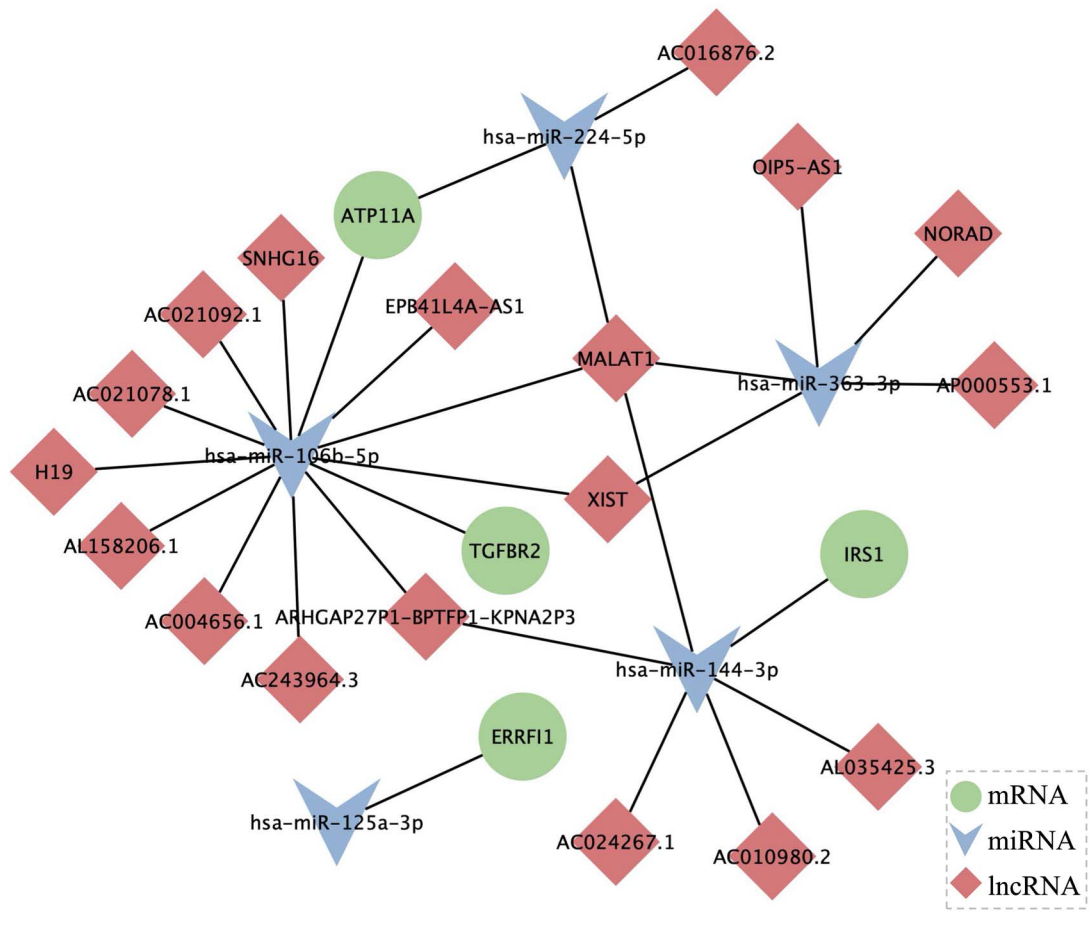
Figure 5. Construction of competing endogenous regulatory networks. Green circle nodes: target genes; red diamond: lncRNAs; blue arrow: miRNAs.
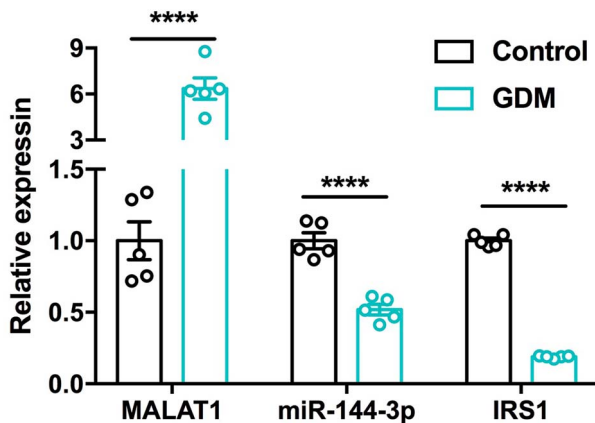

Figure 6. Experimental validation of expression change of MALAT1/hsa-miR-144-3p/IRS1. RT-qPCR analysis of the expression level of MALAT1/hsa-miR-144-3p/IRS1 in placentas from GDM patients and controls. The data show the relative expression changes of MALAT1/hsa-miR-144-3p/IRS1 in placentas from both groups. Values are means ± SD and the expression is corrected for the housekeeping gene YWHAZ. Differences were tested by Student's t-test. Significance was set at $P < 0.05$.

corresponding biomarker values for the samples of the datasets. This technique reduced the time to find split points for sparse discrete biomarkers [40]. The original XGBoost model was usually prone to overfitting, which was also confirmed when we compared the learning curves between the original XGBoost and Optimized XGBoost in the multiomics data (Figure 4A). This is most likely due to the spatiotemporal specificity of transcriptome data and genetic heterogeneity among different biological samples. Therefore, we optimized the original XGBoost model in this study to reduce the degree of overfitting in multiomics data, thereby improving the generalization ability of the overall model to meet the requirements of DLRAPom. The Optimized XGBoost model was tested again and again by applying 3-fold cross-validation. Scalability was the best performing factor compared to other machine learning models, such as SVM, multiple logistic regression, LASSO regression and RF [40]. Meanwhile, the weighted quantum sketch program could solve instance weights in approximate tree learning, which was essential to estimate the weight of each risk biomarker in our model. A comprehensive approach was used to estimate the importance of each risk biomarker, including weight (the number of times one biomarker was used to split the data across all trees), gain (the average gain of the biomarker when it was used

was not measured in the datasets, our model could evaluate it as well. When looking for the best split point on biomarker α, it would not traverse the biomarkers with missing values in the datasets but only traverse the

in trees) and cover (the average coverage of the biomarker when it was used in trees). In the future, by obtaining more multiomics data, the Optimized XGBoost machine learning model could be better perfected. Additionally, we encourage other researchers to design more optimized machine learning algorithms based on ours to fit their own data and make the process more flexible.

In the example of GDM, nine risk protein-coding biomarkers were identified as the key points. The other eight genes were computed with less weight than the IRS1 gene in the best performing Optimized XGBoost model, but their correlations with the pathogenesis of GDM could not be dissociated. As the most vital gene among the nine biomarkers, the IRS1 gene encodes one of the most important substrates for insulin receptor and insulin-like growth factor-1 receptor tyrosine kinases [41] and was downregulated in both the GSE154377 and GSE87295 datasets. The decreased expression level of the IRS1 gene was validated by RT–qPCR in the GDM group, which was in line with previous studies [42, 43]. The impairment of IR and IRS1 signaling was reported to be closely connected with the mechanisms underlying chronic insulin resistance [44]. Due to chronic insulin resistance, GDM may convert to type 2 diabetes. As a key regulatory factor, the IRS1 gene was reported to be regulated by lncRNAs and miRNAs in the complicated endogenous network to affect the development of insulin resistance. A previous study by Chen et al. demonstrated that MALAT1 ablation-mediated insulin-induced activation of IRS-1 to regulate insulin responses, which indicated that MALAT1 played an important role in regulating insulin sensitivity [45]. Moreover, a recent study reported that Dnmt3a-dependent promoter methylation and MALAT1 cooperatively downregulated the expression of the IRSI gene by activating oxidative stress, which could lead to hindered insulin signaling and impaired insulin-dependent glucose uptake in skeletal muscle and ultimately promote the development of insulin resistance [46]. This evidence suggested that MALAT1 may negatively regulate IRS1 to a certain extent. Furthermore, in the competing endogenous network constructed by DLRAPom, hsa-miR-144-3p, one of the differentially expressed miRNAs in the GSE112168 dataset, was indicated to regulate the IRS1 gene and be regulated by MALAT1. The pairwise biological targeting relationships between them (MALAT1/hsa-miR-144-3p and hsa-miR-144-3p/IRS1) have been confirmed by dual-luciferase reporter assays [35–38]. These results powerfully illustrated that the IRS1 gene was regulated in a complicated competing endogenous network as a key factor, which was well reflected in our Optimized XGBoost machine learning model.

To date, there have been no additional research data on the expression of hsa-miR-144-3p in the placenta of GDM patients except the GSE112168 dataset. Studies of the relationship between hsa-miR-144-3p expression and diabetes or GDM are mainly focused on the expression profile in peripheral blood. Considering that there are many confounding factors in peripheral blood, these results may not have a reliable reference and directivity. Notably, hsa-miR-144-3p was reported to have decreased expression in the placenta of patients with preeclampsia [47], belonging to the gestational metabolic disorders together with GDM. Moreover, the expression level of hsa-miR-144-3p was found to be lower in epicardial adipose tissue in response to hyperglycemia [48]. Although research has suggested that downregulation of hsa-miR-144-3p in response to hyperglycemia may be the cause of proliferation promotion [49], the exact mechanism underlying hyperglycemia-induced downregulation of hsa-miR-144-3p is unclear. A potentially reasonable explanation is the endogenous competition mechanism between lncRNAs and miRNAs. In fact, it was reported that lncRNA MALAT1 spoused hsa-miR-144-3p and promoted cell proliferation and migration [35, 36, 38]. As shown in our experimental results, there was higher expression of MALAT1 and lower expression of hsa-miR-144-3p in the placenta of GDM patients. In terms of a single target interaction, as the target gene of hsa-miR-144-3p, the expression of the IRS1 gene should show an increasing trend in the GDM group. Unfortunately, the regulatory network formed by the endogenous competition mechanism of noncoding RNAs is very complex, and it is often the result of multiple targeted regulatory effects. In the results, the possible explanation for the decreased expression of the IRS1 gene in the GDM group was that the negative regulatory effect of MALAT1 on IRS1 may be stronger than the regulatory effect of hsa-miR-144-3p on the IRS1 gene. Considering the complexity of the regulatory network constituted by the endogenous competition mechanism of noncoding RNAs and the importance of the network to the mechanisms of disease occurrence and development, it would be desirable to predict and identify them to elucidate the molecular mechanisms of diseases. To date, there has been no directly supportive evidence for the interactive relationships between MALAT1 and hsa-miR-144-3p and IRS1 in GDM. Through DLRAPom, their interactive relationships were identified for the first time, suggesting that the targetable MALAT1/hsa-miR-144-3p/IRS1 regulatory axis may be closely related to the pathogenesis of GDM.

Inevitably, the DLRAPom pipeline also had some limitations. First, since machine learning algorithms were used in the DLRAPom pipeline, the larger the sample size, the more advantageous it is to obtain more reliable and stable prediction results, which is also determined by the nature of the machine learning algorithm. In addition, exploring the targetable lncRNA–miRNA–mRNA axes is the main purpose of the DLRAPom pipeline, but the current design of the DLRAPom pipeline only included genomics, transcriptomics and epigenomics datasets for integrative analyses. Certain datasets do not fit the current DLRAPom pipeline, such as proteomics data, metabolomics data and microbiomics data. Considering that these omics data are also vital in

a variety of biological processes of disease development and mediating various mechanisms of the pathogenesis of diseases, we will update the DLRAPom pipeline with novel optimization algorithms to fit these omics data and make the predictive process more flexible and the results more reliable in the future. Additionally, taking into account the heterogeneity between individuals and the heterogeneity caused by the different characteristics of the different developmental stages of the disease, the use of multiomics data from the same sample would allow more promising predictive results to be obtained.

In conclusion, by adding a novel machine learning model on the basis of conventional analysis and combining experimental verification, the hybrid DLRAPom pipeline was developed to identify targetable disease-related lncRNA–miRNA–mRNA regulatory axes. As stated, DLRAPom is a flexible pipeline, which is an essential contribution to molecular pathogenesis research of diseases, as it can effectively predict potential disease-related ncRNA regulatory networks and provide more promising candidates.

Key Points

- Our study is the first one to propose a new pipeline of machine learning-guided integrative multiomics analysis, which added a novel Optimized XGBoost model on the basis of conventional WGCNA analysis and combined experimental validation to precisely predict targetable disease-related lncRNA–miRNA–mRNA regulatory axes.
- A novel machine learning algorithm, Optimized XGBoost, is developed to reduce the degree of overfitting in multiomics data, and to quantify the importance of each gene for discovering the most essential protein-coding biomarker.
- We identified the association of the MALAT1/hsa-miR-144-3p/IRS1 axis with gestational diabetes mellitus by the pipeline for the first time.
- Our work presents a new solution for the reliable prediction of disease-related lncRNA–miRNA–mRNA regulatory networks, providing useful information for mechanistic studies of noncoding regulatory networks implicated in complex diseases.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Funding

References

1. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *elife* 2013;**2**:e00523.
2. Perez-Perez JM, Candela H, Micol JL. Understanding synergy in genetic interactions. *Trends Genet* 2009;**25**(8):368–76.
3. Zlotorynski E. Gene expression: the yin and yang of enhancer-promoter interactions. *Nat Rev Mol Cell Biol* 2018;**19**(2):75.
4. Henderson CA, Vincent HA, Stone CM, *et al.* Characterization of MicA interactions suggests a potential novel means of gene regulation by small non-coding RNAs. *Nucleic Acids Res* 2013;**41**(5):3386–97.
5. Pai JA, Satpathy AT. High-throughput and single-cell T cell receptor sequencing technologies. *Nat Methods* 2021;**18**(8):881–92.
6. Nam AS, Chaligne R, Landau DA. Integrating genetic and non-genetic determinants of cancer evolution by single-cell multiomics. *Nat Rev Genet* 2021;**22**(1):3–18.
7. Sun YV, Hu YJ. Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. *Adv Genet* 2016;**93**:147–90.
8. Oh M, Park S, Kim S, *et al.* Machine learning-based analysis of multi-omics data on the cloud for investigating gene regulations. *Brief Bioinform* 2021;**22**(1):66–76.
9. Subramanian I, Verma S, Kumar S, *et al.* Multi-omics data integration, interpretation, and its application. *Bioinform Biol Insights* 2020;**14**:1177932219899051.
10. Paczkowska M, Barenboim J, Sintupisut N, *et al.* Functional interpretation working, J. Reimand, P. Consortium, integrative pathway enrichment analysis of multivariate omics data. *Nat Commun* 2020;**11**(1):735.
11. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res* 2018;**46**(20):10546–62.
12. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res* 2019;**47**(2):1044.
13. Liu S, Li B, Liang Q, *et al.* Classification and function of RNA-protein interactions. *Wiley Interdiscip Rev RNA* 2020;**11**(6):e1601.
14. Siomi H, Siomi MC. On the road to reading the RNA-interference code. *Nature* 2009;**457**(7228):396–404.
15. Gabisonia K, Prosdocimo G, Aquaro GD, *et al.* MicroRNA therapy stimulates uncontrolled cardiac repair after myocardial infarction in pigs. *Nature* 2019;**569**(7756):418–22.
16. Yamamura S, Imai-Sumida M, Tanaka Y, *et al.* Interaction and cross-talk between non-coding RNAs. *Cell Mol Life Sci* 2018;**75**(3):467–84.
17. Engreitz JM, Haines JE, Perez EM, *et al.* Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* 2016;**539**(7629):452–5.
18. Huang Y. The novel regulatory role of lncRNA-miRNA-mRNA axis in cardiovascular diseases. *J Cell Mol Med* 2018;**22**(12):5768–75.
19. Sun J, Shi H, Wang Z, *et al.* Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network. *Mol BioSyst* 2014;**10**(8):2074–81.
20. Zhou M, Wang X, Li J, *et al.* Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network. *Mol BioSyst* 2015;**11**(3):760–9.
21. Zhang J, Zhang Z, Chen Z, *et al.* Integrating multiple heterogeneous networks for novel LncRNA-disease association inference. *IEEE/ACM Trans Comput Biol Bioinform* 2019;**16**(2):396–406.

22. Yao Q, Wu L, Li J, *et al.* Global prioritizing disease candidate lncR-NAs via a multi-level composite network. *Sci Rep* 2017;**7**:39516.

23. Del Vecchio G, Li Q, Li W, *et al.* Cell-free DNA methylation and transcriptomic signature prediction of pregnancies with adverse outcomes. *Epigenetics* 2021;**16**(6):642–61.

24. Pinney SE, Joshi A, Yin V, *et al.* Exposure to gestational diabetes enriches immune-related pathways in the transcriptome and methylome of human amniocytes. *J Clin Endocrinol Metab* 2020;**105**(10):3250–64.

25. Haertle L, El Hajj N, Dittrich M, *et al.* Epigenetic signatures of gestational diabetes mellitus on cord blood methylation. *Clin Epigenetics* 2017;**9**:28.

26. Nair S, Jayabalan N, Guanzon D, *et al.* Human placental exosomes in gestational diabetes mellitus carry a specific set of miRNAs associated with skeletal muscle insulin sensitivity. *Clin Sci (Lond)* 2018;**132**(22):2451–67.

27. Ritchie ME, Phipson B, Wu D, *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**(7):e47.

28. Tian Y, Morris TJ, Webster AP, *et al.* ChAMP: updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics* 2017;**33**(24):3982–4.

29. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;**9**:559.

30. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**(1):27–30.

31. Cristianini N, Ricci E. Support vector machines. In: Kao M-Y (ed). *Encyclopedia of Algorithms.* Boston, MA, USA: Springer, 2008, 928–32.

32. Breiman L. Random forests. *Mach Learn* 2001;**45**(1):5–32.

33. Basevi V, Di Mario S, Morciano C, *et al.* Comment on: American Diabetes Association. Standards of medical care in diabetes—2011. *Diabetes Care* 2011;**34**(Suppl. 1):S11–61, Diabetes Care 2011;**34**(5):e53; author reply e54.

34. Wei X, Jia R, Yang Z, *et al.* NAD(+)/sirtuin metabolism is enhanced in response to cold-induced changes in lipid metabolism in mouse liver. *FEBS Lett* 2020;**594**(11):1711–25.

35. Ye W, Ma J, Wang F, *et al.* LncRNA MALAT1 regulates miR-144-3p to facilitate epithelial-mesenchymal transition of lens epithelial cells via the ROS/NRF2/Notch1/snail pathway. *Oxidative Med Cell Longev* 2020;**2020**:8184314.

36. Wang Y, Zhang Y, Yang T, *et al.* Long non-coding RNA MALAT1 for promoting metastasis and proliferation by acting as a ceRNA of miR-144-3p in osteosarcoma cells. *Oncotarget* 2017;**8**(35): 59417–34.

37. Bai J, Hu Y, Chen X, *et al.* miR-144-3p inhibits the invasion and metastasis of lung adenocarcinoma cells by targeting IRS1. *Zhongguo Fei Ai Za Zhi* 2021;**24**(5):323–30.

38. Gong X, Zhu Y, Chang H, *et al.* Long noncoding RNA MALAT1 promotes cardiomyocyte apoptosis after myocardial infarction via targeting miR-144-3p. *Biosci Rep* 2019;**39**(8):BSR20191103.

39. Olivier M, Asmis R, Hawkins GA, *et al.* The need for multi-omics biomarker signatures in precision medicine. *Int J Mol Sci* 2019;**20**(19):4781.

40. Chen T, Guestrin C. *XGBoost, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 2016, pp. 785-94.

41. Hayashi T, Kubota T, Mariko I, *et al.* Lack of brain insulin receptor Substrate-1 causes growth retardation, with decreased expression of growth hormone-releasing hormone in the hypothalamus. *Diabetes* 2021;**70**(8):1640–53.

42. Barbour LA, McCurdy CE, Hernandez TL, *et al.* Chronically increased S6K1 is associated with impaired IRS1 signaling in skeletal muscle of GDM women with impaired glucose tolerance postpartum. *J Clin Endocrinol Metab* 2011;**96**(5): 1431–41.

43. Colomiere M, Permezel M, Riley C, *et al.* Defective insulin signaling in placenta from pregnancies complicated by gestational diabetes mellitus. *Eur J Endocrinol* 2009;**160**(4):567–78.

44. Friedman JE, Kirwan JP, Jing M, *et al.* Increased skeletal muscle tumor necrosis factor-alpha and impaired insulin signaling persist in obese women with gestational diabetes mellitus 1 year postpartum. *Diabetes* 2008;**57**(3):606–13.

45. Chen J, Ke S, Zhong L, *et al.* Long noncoding RNA MALAT1 regulates generation of reactive oxygen species and the insulin responses in male mice. *Biochem Pharmacol* 2018;**152**: 94–103.

46. Wei J, Hao Q, Chen C, *et al.* Epigenetic repression of miR-17 contributed to di(2-ethylhexyl) phthalate-triggered insulin resistance by targeting Keap1-Nrf2/miR-200a axis in skeletal muscle. *Theranostics* 2020;**10**(20):9230–48.

47. Hu S, Li J, Tong M, *et al.* MicroRNA1443p may participate in the pathogenesis of preeclampsia by targeting Cox2. *Mol Med Rep* 2019;**19**(6):4655–62.

48. Oclon E, Latacz A, Zubel-Lojek J, *et al.* Hyperglycemia-induced changes in miRNA expression patterns in epicardial adipose tissue of piglets. *J Endocrinol* 2016;**229**(3):259–66.

49. Plebani M, Masiero M, Sciacovelli L, *et al.* A rapid, specific enzyme immunoassay for follitropin and lutropin determination. *Clin Chem* 1988;**34**(4):772.