Identification of Novel Biomarkers for Metabolic Syndrome Based on Machine Learning Algorithms and Integrated Bioinformatics Analysis

Guanzhi Liu
Xi'an Jiaotong University Second Affiliated Hospital https://orcid.org/0000-0003-1626-5006
Chen Chen
Department of Cardiovascular Medicine, First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China
Ning Kong
Bone and Joint Surgery Center, Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China
Yutian Lei
Bone and Joint Surgery Center, Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China
Sen Luo
Bone and Joint Surgery Center, Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China
Zhuo Huang
Bone and Joint Surgery Center, Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China
Kunzheng Wang
Bone and Joint Surgery Center, Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China
Pei Yang
Bone and Joint Surgery Center, Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China
Xin Huang (hearthx@126.com)
Department of Cardiovascular Medicine, First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, China

Research

Keywords: metabolic syndrome, WGCNA, diagnostic biomarkers, bioinformatics, machine learning

Posted Date: February 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-225591/v1

Abstract

Background: Metabolic syndrome is a common and complicated metabolic disorder and defined as a clustering of metabolic risk factors such as insulin resistance or diabetes, obesity, hypertension, and hyperlipidemia. However, its early diagnosis is limited because the lack of definitive clinical diagnostic biomarkers. In present study, we aim to select several candidate gene as a blood-based clinically applicable transcriptomics signature for metabolic syndrome.

Method: We collected so far the largest MetS-associated peripheral blood high-throughput transcriptomics data and put forward a novel feature selection strategy by combining weighted gene co-expression network analysis, protein-protein interaction network analysis, LASSO regression and random forest approaches. Then, based on selected hub gene signature, we performed logistic regression analysis and subsequently established a web nomogram calculator for metabolic syndrome risk to detect the diagnostic value of this hub gene signature. Finally, Receiver Operating Characteristic curve analysis, calibration curve analysis, Hosmer-Lemeshow good of fit test and decision curve analysis showed the classification and calibration performance as well as potential clinical benefit of this hub gene signature.

Results: Through weighted gene co-expression network analysis, protein-protein interaction network analysis, we identified 2 gene modules and 51 hub genes associated with metabolic syndrome. Then, we subsequently performed further feature selection via LASSO regression and random forest method. Finally, a 9-hub-gene signature with high diagnostic value and a web nomogram calculator for metabolic syndrome risk (https://xjtulgz.shinyapps.io/DynNomapp/) were developed. This 9-hub-gene signature showed excellent classification and calibration performance (AUC= 0.968 in training set, AUC= 0.883 in internal validation set, AUC= 0.861 in external validation set) as well as ideal potential clinical benefit.

Conclusions: The blood-based 9-hub-gene signature identified in present study and the web nomogram calculator for metabolic syndrome risk are possible to accurately achieve the noninvasive screening or diagnosis of MetS considering the excellent classification ability, calibration and potential clinical benefits.

1. Introduction

Metabolic syndrome (MetS) is defined as a complex abnormality which have several components such as insulin resistance or diabetes, obesity, hypertension, and hyperlipidemia[1–3]. The occurrence and development of MetS and its components always associated with poor cardiovascular outcomes, especially obesity and insulin resistance which were regarded as the core pathophysiological features of MetS[4,5]. The unclear molecular mechanism and complicated diagnosis method makes it difficult for the early intervention of MetS and metabolic syndrome-related diseases[6–8]. Some studies have reported the potential biomarkers of MetS, however, there is still a lack of definitive clinical MetS diagnostic biomarkers[9,10]. Currently, the researches about MetS biomarkers are limited to genomics level and the

association between MetS and single nucleotide polymorphisms (SNPs)[11,12]. Little attention has been focused on the MetS- specific biomarkers in transcriptomics aspect[13].

In the field of high-throughput technology of transcriptomics, microarrays technology and next-generation sequencing (NGS) have been widely used to measure RNA expression levels[14,15]. In addition, advanced bioinformatics approaches like weighted gene co-expression network analysis (WGCNA) can play significant roles in the identification of disease biomarkers with high sensitivity, specificity and efficiency, based on high-throughput transcriptomics data[16–18]. Compared to the traditional bioinformatics methods such as differentially expressed gene (DEG) analysis, network-focused algorithm WGCNA can established a weighted scale-free co-expression network and then more efficiently identified key gene modules and hub genes[19,20]. Machine learning (ML), as a main aspect of artificial intelligence, have got increasingly widely applied in many biomedicine fields such as biomarker identification, diagnosis signature development and drug target discovery[21,22]. Moreover, it have been proved that some machine learning methods like Least absolute shrinkage and selection operator (LASSO) regression and random forest (RF) can obtain much better performance in biomarker development of multifactorial and complicated disease[23–26].

Hence, in this study, to our best knowledge, we firstly performed integrated bioinformatics approaches such as WGCNA analysis based on the largest MetS-associated peripheral blood high-throughput transcriptomics data set at present. Then, we identified several hub genes via protein-protein interaction (PPI) network analysis and conducted further hub genes feature selection by combining LASSO regression and random forest algorithm. Finally, we established a logistic regression and a web nomogram calculator for MetS risk (https://xjtulgz.shinyapps.io/DynNomapp/) based on training set and measured the diagnostic value of these selected hub genes features by internal and external validation data. In addition, in order to further detect the hub gene expression difference in peripheral blood and plasma, we firstly carried out next-generation sequencing in plasma samples of MetS patients compared with control group patients. Current study aimed to identify several gene parameters with high diagnostic value and clinical implications for MetS by comprehensive bioinformatics and machine learning feature selection methods which can provide a novel strategy for further researches to develop biomarkers more effectively and reliably.

2. Materials And Methods

2.1 Data collection and preprocessing

It had been wildly accepted that metabolic syndrome can result in highly specific alteration of gene expression in peripheral blood. Therefore, in this work, a variety of public gene expression datasets based on peripheral blood samples containing metabolic syndrome-associated clinical diagnosis information were collected from the Gene Expression Omnibus database (GEO database, http://www.ncbi.nlm.nih.gov/geo/). Training set of this study consisted of randomly selected 70% of the

samples in GSE152073 (n = 90) and GSE98895 (n = 40) combined dataset (gene expression microarray data of peripheral blood), and the remaining 30% was used as internal validation data[27,28]. GSE124534 (n = 17, gene expression microarray data of peripheral blood) was used to achieve the external validation[29]. Subjects with diagnosis of other metabolic diseases or acute trauma, such as osteoporosis and femoral neck fracture which may causes gene expression changes, were excluded. The detailed information of these datasets was listed in **Additional file 1: Table S1**. After removed the outliers and probes which was duplicate or unable to annotate, we normalized these gene expression data and removed batch effects by the "limma" package in R. Missing data was imputed by R software package "impute".

2.2 Weighted gene co-expression network analysis

Weighted gene co-expression network analysis was performed based on GSE98895 datasets by R package "WGCNA"[19]. First, we calculated Pearson's correlation for all pairs of genes and established the similarity matrix. Second, an appropriate soft thresholding power of 2 was selected to meet the scale-free topology (scale free R2 > 0.9) criterion by the function "pickSoftThreshold". Third, topological overlap matrix (TOM) and the corresponding dissimilarity matrix (dissTOM) were constructed. Then, we used "blockwiseModules" function with the following major parameters: maxBlockSize of 5000, minModuleSize of 30 and mergeCutHeight = 0.25 and identified several gene modules through hierarchical clustering with dynamic tree-cutting algorithm. Finally, the correlation between gene modules and clinical phenotypes was calculated so as to identify the clinically significant modules.

2.3 Enrichment analysis of modules

In order to explore the function and signaling pathway associated with these modules, we performed Gene Ontology (GO) function enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis as well as Gene Set Enrichment Analysis (GSEA) via the "clusterProfiler" package, "enrichplot" package, "DOSE" package and "ggplot2" package in R software[30]. We set P value < 0.05 as the threshold.

2.4 PPI network construction and hub gene identification

In this work, we constructed protein-protein interaction (PPI) network based on the STRING database (Search Tool for the Retrieval of Interacting Genes, version 11.0, combined score > 0.4). Then, connectivity degrees in the network were calculated and top 5% of the genes with the highest connectivity degree were identified as hub genes for further analysis. Visualization of hub genes in PPI network was achieved by Cytoscape software(version 3.7.0).

2.5 Clinical plasma sample collection

We obtained peripheral blood samples of 5 patients with metabolic syndrome and 5 patients in healthy volunteers group from First Affiliated Hospital of Xi'an Jiaotong University. We used the World Health Organization (WHO) metabolic syndrome definition. The diagnosis of MetS needs the presence of impaired fasting glucose (IFG), impaired glucose tolerance (IGT), type 2 diabetes mellitus (T2DM) or

Insulin resistance and two or more of the following: (1) waist-to-hip ratio > 0.90 in men; waist-to-hip ratio > 0.85 in women and/or BMI > 30 kg/m2; (2) serum triglyceride level ≥ 1.7 mmol/L (3) HDL cholesterol < 0.9 mmol/L in men, < 1.0 mmol/L in women, or treatment for dyslipidaemiaor; (4) blood pressure ≥ 140/90 mm Hg (5) microalbuminuria[31]. The plasma was separated and then used to perform high-throughput sequencing. This study was approved by the Ethics Committee of First Affiliated Hospital of Xi'an Jiaotong University granted (Ethical Approval number: XJTU1AF2019LSL-014). All subjects gave written informed consent in advance.

2.6 RNA extraction and high-throughput sequencing

Total RNA from plasma samples were extracted using TRIzol LS Reagent (Invitrogen) according to the manufacturer's instructions. Then, sequencing libraries were generated by NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs), RiboZero Magnetic Gold Kit (Epicentre, an Illumina Company) and KAPA Stranded RNA-Seq Library Prep Kit (Illumina). Agilent Bioanalyzer 2100 system (Agilent) was used to qualified the sequencing libraries. Finally, the high-throughput next-generation sequencing was carried out using TruSeq SR Cluster Kit (Illumina) based on the Illumina HiSeq 4000 sequencing platform (Illumina).

2.7 Plasma mRNA differential expression analysis

Trimmed reads were identified after raw sequencing data quality control and filtering process by Solexa pipeline program (version 1.8) and Cutadapt software. Subsequently, we obtained human reference genome indexing (hg38) via bowtie software (http://bowtie-bio.sourceforge.net/index.shtml). Then, sequencing alignment was achieved by Hisat2 program. R package "edgeR" was used to detect the differentially expressed genes (DEGs)[32]. We set the thresholds for DEGs as |log2FC|≥ 1 and P value < 0.05.

2.8 Hub gene feature selection strategy

Machine learning algorithm have been proved to be more powerful than traditional methods for complex classification like medical diagnosis and treatment.In this study, we combined two machine learning approaches LASSO regression and Random Forest by R packages "glmnet" and "randomForest" to achieve feature selection[33]. The feature selection was crosschecked and we selected several hub genes according to the classification accuracy. Hub genes both selected by LASSO regression and Random Forest feature selection strategy were further used to established diagnosis classifier.

2.9 Web nomogram calculator construction and validation of 9-hub-gene signature

R packages "rms" was used to established logistic regression model based on expression data in train set. Subsequently, corresponding web nomogram calculator for MetS risk was constructed to visualized the diagnosis effect of selected hub gene signature[34]. Then, we performed internal validation and external validation to detect the performance of this web nomogram calculator. Area Under Curve (AUC)

value of Receiver Operating Characteristic (ROC) curve was calculated by “pROC” package in R software which can depict the classification ability[35]. Hosmer-Lemeshow good of fit test and calibration curve analysis was conducted to indicate the calibration. In addition, decision curve analysis was carried out via “rmda” package to evaluate the clinical application value and net benefit of the nomogram.

3. Result

3.1 Weighted co-expression networks (WGCNA) construction and key modules identification

The workflow of this work was shown in Fig. 1. We combined the most comprehensive sets so far of MetS-associated high-throughput transcriptomics data from GEO database (**Additional file 1: Table S1)**. Gene expression profiles from GSE98895 were used to perform WGCNA analysis. After pre-processing and batch effect removal, we identified 25148 gene expression data of 20 MetS and 20 control group patients’ peripheral blood. Sample clustering analysis base on Pearson's correlation approach and average linkage approach showed no outliers (Fig. 2A). Then, in order to achieve scale-free topology (scale free $R2 > 0.9$), we selected soft-thresholding power $\beta = 2$ (Fig. 2B). Subsequently WGCNA network construction and average linkage hierarchical clustering detected 14 gene modules. Detailed hierarchical clustering information was shown in (Fig. 2C,D). Finally, through correlation analysis between these modules and MetS, we found red module (618 genes) and black module (546 genes) are highly associated with MetS (Fig. 2E). Hence, these two modules were identified as key module of MetS for further analysis. Scatter diagrams contain key module GS and MM information were shown in (Fig. 2F,G).

3.2 Gene ontology and pathway enrichment analysis

The results of Gene ontology (GO) functional enrichment analysis showed that these MetS-associated genes in red and black module mainly enriched in biology process (BP) such as receptor guanylyl cyclase signaling pathway, central nervous system neuron differentiation, response to calcium ion, platelet activation and so on. Besides, these genes were associated with molecular function (MF) such as tumor necrosis factor (TNF) receptor and lipid transporter activity. Cellular components (CC) like cellular junction and guanyl-nucleotide exchange factor complex may correlated with the development of MetS. As for KEGG signaling pathway enrichment analysis, our results indicated that these genes were significantly enriched in signaling pathways such as cell adhesion molecules, leukocyte transendothelial migration, calcium signaling pathway and so on (Fig. 3A,B). In addition, we performed GSEA analysis to further reveal the function and signaling pathway of these genes which showed the similar result as CC gene ontology and KEGG pathway enrichment analysis. BP GSEA analysis and MF GSEA analysis suggested that BP such as regulation of lymphocyte activation, drug metabolic process and MF such as lyase activity, hydrolase activity, molecular transducer activity, G-protein coupled receptor activity may involved with the development of MetS (Fig. 3C-F).

3.3 PPI network construction and hub gene identification

PPI network were established by STRING database based on genes in red and black modules. We calculated the connectivity degree and selected the top 5% genes (51 genes) with highest connectivity degree as hub genes associated with MetS. Hub genes with high connectivity degree such as MYC, UBE2E2, MIB2, ANAPC1, TCEB1, CTLA4, SPI1 may play important roles in the development of MetS and serve as potential biomarkers and therapy targets. The visualization of hub gene PPI network were shown in Fig. 4. These 51 hub genes (**Additional file 2: Table S2**) were used for further feature reduction analysis and model construction.

3.4 Hub gene expression level in plasma

A total of 12954 genes were discovered in our high-throughput sequencing data of plasma samples of 5 patients in metabolic syndrome group and 5 patients in healthy control group. Genes with P value < 0.05 and $|\log2FC| \geq 1$ were considered as differentially expressed genes (DEGs). Finally, we identified 45 up-regulated and 186 down-regulated DEGs in metabolic syndrome group compared with control group **(Additional file 3: Table S3)**. We found that the expression of these 51 hub genes in plasma have no significant difference between metabolic syndrome patients and healthy control patients (**Additional file 4: Table S4**). Our results indicated that we should focused on the potential function and diagnostic value of these 51 hub genes in peripheral blood cellular components instead of plasma component which defined the sampling type of further noninvasive MetS screening or diagnostic tools.

3.5 Novel hub gene feature selection strategy

In this study, LASSO regression analysis and random forest approach were used to achieve feature selection. The expression data of these 51 hug genes were entered into LASSO regression models and 10-fold cross-validation was performed to detect the optimal classification accuracy (Fig. 5A,B). Hence, we obtained 15 hub gene features based on LASSO regression analysis including ADRA2A, CXCR5, FZD1, HLA.DPA1, HSPA5, KCTD7, KLHL9, P2RY14, P2RY2, PRKACG, PSMD1, PTTG1, REEP4,SPTAN1, TSPAN14. In addition, We constructed random forest model via these 51 hug genes expression profiles and measured the classification importance of hub gene feature by mean decrease in Gini coefficient (MeanDecreaseGini). Then, 15 hub genes features were chosen by random forest approach including SPTAN1, KCTD7, IL2RG, ITPR3, PSMD1, ITGB7, FZD1, DCTN4, KLHL9, PTTG1, TSPAN14, RNF19B, XCR1, P2RY2, CXCR5 (Fig. 5C**)**. Finally, we combined the results of these two gene feature selection method by taking the intersection and selected 9 hub gene features (SPTAN1, KCTD7, PSMD1, FZD1, KLHL9, PTTG1, TSPAN14, P2RY2, CXCR5) for further analysis.

3.6 Web nomogram calculator construction and validation of 9-hub-gene signature

The expression profiles of these 9 selected gene features were entered into a logistic regression and then, in order to validate the diagnostic value of this 9-hub-gene signature, we established a web nomogram calculator for MetS risk based on training set (https://xjtulgz.shinyapps.io/DynNomapp/). Then, ROC curve analysis (Fig. 6A) showed this MetS diagnostic nomogram had excellent classification ability (AUC = 0.968 in training set, AUC = 0.883 in internal validation set, AUC = 0.861 in external validation set). The

ROC curves of every single hub genes are shown in **Additional file 5: Figure S1**. In addition, we performed calibration curve analysis and Hosmer-Lemeshow good of fit test (P = 0.915) which showed good calibration of this nomogram (Fig. 6B). Furthermore, decision curve ploted the standardized net benefit of our MetS diagnostic nomogram in different decision thresholds (Fig. 6C). These results indicated that the application of this MetS diagnostic nomogram would lead to ideal clinical outcomes.

4. Discussion

In the past few years, a great deal of attention has been devoted to metabolic syndrome, however, it is still difficult to achieve early diagnosis and intervention because of the lack of effective biomarkers and treatment target[36,37]. In this study, to our best knowledge, we firstly identified a key gene module and 51 MetS-associated hub genes by combining WGCNA bioinformatics approach and PPI network analysis. Genes in this key module were mainly enriched in signaling pathways like cell adhesion, leukocyte transendothelial migration signaling, NF-κB (Nuclear factor kappa B) and functions like lymphocyte activation. Moreover, These 51 hub genes may play important roles in the development of MetS. Cheung et al suggested that MYC (MYC Proto-Oncogene) can serve as an important mediator of impaired insulin secretion and β-cell apoptosis[38]. Some researches have indicated that the SNPs in UBE2E2 (Ubiquitin Conjugating Enzyme E2) were associated with the development of type 2 diabetes mellitus[39,40]. Besides, MIB2 (Mindbomb E3 Ubiquitin Protein Ligase 2), ANAPC1 (Anaphase Promoting Complex Subunit 1) and TCEB1 (ELOC, Elongin C) are also involved in the process of ubiquitination which can affect the development of insulin resistance and metabolic syndrome[41–43]. CTLA4 (Cytotoxic T-Lymphocyte Associated Protein 4) was reported to be involved in T-cell immune responses and thus regulated the pathogenesis of insulin resistance and insulin-dependent diabetes mellitus[44,45]. Moreover, the up-regulation of SPI1(Spi-1 Proto-Oncogene, also known as PU.1) in adipocyte can cause insulin resistance by stimulating ROS production and inflammatory cytokine gene expression[46,47]. It is apparent that these hub genes could serve as biomarkers for MetS and many of its contributing components.

In addition, through machine learning feature selection methods, we finally obtained a 9-hub-gene signature with high diagnostic value and clinical implications for MetS. Dhana et al found that PSMD1 (Proteasome 26S Subunit, Non-ATPase) gene was associated with both BMI (body mass index) and WC (waist circumference) and could serve as biomarkers for obesity-related diseases[48]. Some studies showed that FZD1 (Frizzled Class Receptor 1) is related to the occurrence of insulin resistance[49,50]. Besides, Frendo et al indicated that KLHL9 (Kelch Like Family Member 9) can induce insulin resistance by regulating IRS1 (Insulin Receptor Substrate-1) degradation[51]. PTTG1 (Pituitary Tumor-Transforming Gene 1) is a crucial factor in the development and physiological responses of pancreatic beta-cell and its dysregulation could result in diabetes[52,53]. TSPAN14 (Tetraspanin 14) can interact with ADAM10 (ADAM Metallopeptidase Domain 10) and then regulate leukocyte development and inflammatory immunity function[54]. Previous studies have demonstrated that P2RY2 (Purinergic Receptor P2Y2) contributes to the development of chronic high-fat diet-induced metabolic dysfunction and insulin resistance[55,56]. Besides, Merz et al suggested that P2RY2 is involved in the process of immune cell infiltration in metabolic

syndrome[57]. Follicular helper T (Tfh) cells of diabetes patients expressed elevated levels of CXCR5 (C-X-C Motif Chemokine Receptor 5) and Wang et al found the dysregulation of circulating CD4 + CXCR5 + T cells in diabetes patients[58,59]. Finally, we verified the classification ability, calibration and potential clinical benefit of this blood-based 9-hub-gene signature in internal and external validation set. Previous studies did not investigate the diagnostic value of these 9 hub genes for MetS and so far a early screening or diagnostic tool for MetS has not been developed[60]. However, In this study, we combined the blood-based 9-hub-gene signature by logistic regression and visualized as a nomogram which have excellent classification and calibration performance. The AUC of ROC curves can reach 0.883 in internal validation set and 0.861 in external validation set. Hosmer-Lemeshow good of fit test P value = 0.915, which showed good calibration. Further decision curve analysis showed that this nomogram have much more ideal net benefit than any single gene signature almost in all range of decision thresholds. Overall, our results indicated that this 9-hub-gene signature is useful for further MetS-associated blood-based screening or diagnosis in clinical application.

In this study, we collected the largest MetS-associated peripheral blood high-throughput transcriptomics data set so far, However, it should be noted that further large independent patient cohorts validation researches is still needed to establish a diagnostic model for clinical application.

5. Conclusion

In conclusion, the 9-hub-gene signature identified in present study is possible to accurately achieve the screening or diagnosis of MetS considering the excellent classification ability, calibration and potential clinical benefits. Besides, we firstly put forward a novel diagnostic biomarkers selection method by combining WGCNA approaches, PPI network analysis, LASSO regression and RF feature selection algorithm. In addition, we firstly performed high-throughput sequencing to detect the plasma cell-free mRNA expression level in MetS patients compared with healthy control patients, which can provide a reliable basis for sampling type in the construction of MetS diagnosis of screening tool.

Declarations

Conflict of Interest

The authors declare that there is no conflict of interest.

Author Contributions

Conception and design: Xin Huang, Pei Yang, Kunzheng Wang, Guanzhi Liu; collection and assemble of data: Ning Kong, Fangze Xing, Yutian Lei, Zhuo Huang, Sen Luo; analysis and interpretation of the data: Guanzhi Liu, Chen Chen; draft of the article: Guanzhi Liu; All authors read, critically revised and approved the final manuscript.

Funding

The research was funded by Key Project for Science Research and Development of Shaanxi Province (2019SF-164) and Project for Science Research of First Affiliated Hospital of Xi'an Jiaotong University (XJTU1AF-CRF-2019-014).

Acknowledgments

We would like to thanks all participants for their commitment and cooperation.

Availability of data and materials

The data that support the findings of the this study are available from the corresponding author on reasonable request. The datasets for this study can be found in the Gene Expression Omnibus (GEO) database [https://www.ncbi.nlm.nih.gov/geo/].

Ethics approval and consent to participate

This study was approved by the Ethics Committee of First Affiliated Hospital of Xi'an Jiaotong University granted (Ethical Approval number: XJTU1AF2019LSL-014). All patients gave written informed consent in advance.

Consent for publication

Not applicable.

References

1. Jahani V, Kavousi A, Mehri S, Karimi G. Rho kinase, a potential target in the treatment of metabolic syndrome. *Biomed Pharmacother*. 2018;106(May):1024-1030. doi:10.1016/j.biopha.2018.07.060
2. Ruderman NB, Carling D, Prentki M, Cacicedo JM. AMPK, insulin resistance, and the metabolic syndrome. *J Clin Invest*. 2013;123(7):2764-2772. doi:10.1172/JCI67227.2764
3. Martínez MC, Andriantsitohaina R. Extracellular vesicles in metabolic syndrome. *Circ Res*. 2017;120(10):1674-1686. doi:10.1161/CIRCRESAHA.117.309419
4. Rask-Madsen C, Kahn CR. Tissue-specific insulin signaling, metabolic syndrome, and cardiovascular disease. *Arterioscler Thromb Vasc Biol*. 2012;32(9):2052-2059. doi:10.1161/ATVBAHA.111.241919
5. Ren J, Anversa P. The insulin-like growth factor i system: Physiological and pathophysiological implication in cardiovascular diseases associated with metabolic syndrome. *Biochem Pharmacol*. 2015;93(4):409-417. doi:10.1016/j.bcp.2014.12.006
6. Cornier MA, Dabelea D, Hernandez TL, et al. The metabolic syndrome. *Endocr Rev*. 2008;29(7):777-822. doi:10.1210/er.2008-0024
7. Gong L li, Yang S, Zhang W, et al. Discovery of metabolite profiles of metabolic syndrome using untargeted and targeted LC–MS based lipidomics approach. *J Pharm Biomed Anal*. 2020;177. doi:10.1016/j.jpba.2019.112848

8. Nolan CJ, Prentki M. Insulin resistance and insulin hypersecretion in the metabolic syndrome and type 2 diabetes: Time for a conceptual framework shift. *Diabetes Vasc Dis Res*. 2019;16(2):118-127. doi:10.1177/1479164119827611
9. Chen PY, Cripps AW, West NP, Cox AJ, Zhang P. A correlation-based network for biomarker discovery in obesity with metabolic syndrome. *BMC Bioinformatics*. 2019;20(Suppl 6):1-10. doi:10.1186/s12859-019-3064-2
10. Robinson MD, Mishra I, Deodhar S, et al. Water T2 as an early, global and practical biomarker for metabolic syndrome: An observational cross-sectional study. *J Transl Med*. 2017;15(1):1-19. doi:10.1186/s12967-017-1359-5
11. Kong S, Cho YS. Identification of female-specific genetic variants for metabolic syndrome and its component traits to improve the prediction of metabolic syndrome in females. *BMC Med Genet*. 2019;20(1):1-13. doi:10.1186/s12881-019-0830-y
12. Moon S, Lee Y, Won S, Lee J. Multiple genotype-phenotype association study reveals intronic variant pair on SIDT2 associated with metabolic syndrome in a Korean population. *Hum Genomics*. 2018;12(1):1-10. doi:10.1186/s40246-018-0180-4
13. Paczkowska-Abdulsalam M, Niemira M, Bielska A, et al. Evaluation of transcriptomic regulations behind metabolic syndrome in obese and lean subjects. *Int J Mol Sci*. 2020;21(4). doi:10.3390/ijms21041455
14. Su Z, Fang H, Hong H, et al. An investigation of biomarkers derived from legacy microarray data for their utility in the RNA-seq era. *Genome Biol*. 2014;15(12):523. doi:10.1186/s13059-014-0523-y
15. Shiino S, Matsuzaki J, Shimomura A, et al. Serum miRNA–based Prediction of Axillary Lymph Node Metastasis in Breast Cancer. *Clin Cancer Res*. 2019;25(6):1817-1827. doi:10.1158/1078-0432.CCR-18-1414
16. Liu GZ, Chen C, Kong N, et al. Identification of potential miRNA biomarkers for traumatic osteonecrosis of femoral head. *J Cell Physiol*. 2020;235(11):8129-8140. doi:10.1002/jcp.29467
17. Guillotin D, Taylor A, Platé M, et al. Transcriptome analysis of IPF fibroblastic foci identifies key pathways involved in fibrogenesis. *Thorax*. 2020:1-10. doi:10.1101/2020.03.10.984955
18. Chen C, Liu GZ, Liao YY, et al. Identification of Candidate Biomarkers for Salt Sensitivity of Blood Pressure by Integrated Bioinformatics Analysis. *Front Genet*. 2020;11(September):1-10. doi:10.3389/fgene.2020.00988
19. Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9. doi:10.1186/1471-2105-9-559
20. Wang G, Yu J, Yang Y, et al. Whole-transcriptome sequencing uncovers core regulatory modules and gene signatures of human fetal growth restriction. *Clin Transl Med*. 2020;9(1). doi:10.1186/s40169-020-0259-0
21. Rahul C. Deo. Machine Learning in Medicine. *Circulation*. 2017;25(5):1032-1057. doi:10.1111/mec.13536.Application

22. Mamoshina P, Vieira A, Putin E, Zhavoronkov A. Applications of Deep Learning in Biomedicine. *Mol Pharm*. 2016;13(5):1445-1454. doi:10.1021/acs.molpharmaceut.5b00982
23. Wei Q, Fang W, Chen X, et al. Establishment and validation of a mathematical diagnosis model to distinguish benign pulmonary nodules from early non-small cell lung cancer in Chinese people. *Transl Lung Cancer Res*. 2020;9(5):1843-1852. doi:10.21037/tlcr-20-460
24. Cánovas R, Cobb J, Brozynska M, et al. Genomic risk scores for juvenile idiopathic arthritis and its subtypes. *Ann Rheum Dis*. 2020;(figure 1):1-8. doi:10.1136/annrheumdis-2020-217421
25. Howard F, Kochanny S, Koshy M, Spiotto MT, Pearson AT. Machine learning guided adjuvant treatment of head and neck cancer. *J Clin Oncol*. 2020;38(15_suppl):6567-6567. doi:10.1200/jco.2020.38.15_suppl.6567
26. Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. *Brief Bioinform*. 2019;20(2):492-503. doi:10.1093/bib/bbx124
27. Jales Neto LH, Wicik Z, Torres GHF, et al. Overexpression of SNTG2, TRAF3IP2, and ITGA6 transcripts is associated with osteoporotic vertebral fracture in elderly women from community. *Mol Genet Genomic Med*. 2020;8(9):1-12. doi:10.1002/mgg3.1391
28. D'Amore S, Härdfeldt J, Cariello M, et al. Identification of miR-9-5p as direct regulator of ABCA1 and HDL-driven reverse cholesterol transport in circulating CD14 + cells of patients with metabolic syndrome. *Cardiovasc Res*. 2018;114(8):1154-1164. doi:10.1093/cvr/cvy077
29. Matualatupauw JC, O'Grada C, Hughes MF, Roche HM, Afman LA, Bouwman J. Integrated Analys of High-Fat Challenge-Induced Changes in Blood Cell Whole-Genome Gene Expression. *Mol Nutr Food Res*. 2019;63(20):1-9. doi:10.1002/mnfr.201900101
30. Yu G, Wang LG, Han Y, He QY. ClusterProfiler: An R package for comparing biological themes among gene clusters. *Omi A J Integr Biol*. 2012;16(5):284-287. doi:10.1089/omi.2011.0118
31. Balkau B, Charles MA. Comment on the provisional report from the WHO consultation. *Diabet Med*. 1999;16(5):442-443. doi:10.1046/j.1464-5491.1999.00059.x
32. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2009;26(1):139-140. doi:10.1093/bioinformatics/btp616
33. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1-22. doi:10.18637/jss.v033.i01
34. Ying W, Riopel M, Bandyopadhyay G, et al. Adipose Tissue Macrophage-Derived Exosomal miRNAs Can Modulate in Vivo and in Vitro Insulin Sensitivity. *Cell*. 2017;171(2):372-384.e12. doi:10.1016/j.cell.2017.08.035
35. Robin X, Turck N, Hainard A, et al. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12(1):77. doi:10.1186/1471-2105-12-77
36. O'Neill S, O'Driscoll L. Metabolic syndrome: A closer look at the growing epidemic and its associated pathologies. *Obes Rev*. 2015;16(1):1-12. doi:10.1111/obr.12229

37. Carrier A. Metabolic syndrome and oxidative stress: A complex relationship. *Antioxidants Redox Signal*. 2017;26(9):429-431. doi:10.1089/ars.2016.6929
38. Cheung L, Zervou S, Mattsson G, et al. C-Myc directly induces both impaired insulin secretion and loss of β-cell mass, independently of hyperglycaemia in vivo. *Islets*. 2010;2(1):37-45. doi:10.4161/isl.2.1.10196
39. Kazakova E V., Wu Y, Zhou Z, et al. Association between UBE2E2 variant rs7612463 and type 2 diabetes mellitus in a Chinese Han Population. *Acta Biochim Pol*. 2015;62(2):241-245. doi:10.18388/abp.2014_936
40. Yamauchi T, Hara K, Maeda S, et al. A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A-C2CD4B. *Nat Genet*. 2010;42(10):864-868. doi:10.1038/ng.660
41. Yang XD, Xiang DX, Yang YY. Role of E3 ubiquitin ligases in insulin resistance. *Diabetes, Obes Metab*. 2016;18(8):747-754. doi:10.1111/dom.12677
42. Song R, Peng W, Zhang Y, et al. Central role of E3 ubiquitin ligase MG53 in insulin resistance and metabolic disorders. *Nature*. 2013;494(7437):375-379. doi:10.1038/nature11834
43. Yang S, Wang B, Humphries F, Hogan AE, O'Shea D, Moynagh PN. The E3 Ubiquitin ligase pellino3 protects against obesity-induced inflammation and insulin resistance. *Immunity*. 2014;41(6):973-987. doi:10.1016/j.immuni.2014.11.013
44. Moraes-Vieira PM, Castoldi A, Aryal P, Wellenstein K, Peroni OD, Kahn BB. Antigen presentation and T-cell activation are critical for RBP4-induced insulin resistance. *Diabetes*. 2016;65(5):1317-1327. doi:10.2337/db15-1696
45. Cabrera SM, Engle S, Kaldunski M, et al. Innate immune activity as a predictor of persistent insulin secretion and association with responsiveness to CTLA4-Ig treatment in recent-onset type 1 diabetes. *Diabetologia*. 2018;61(11):2356-2370. doi:10.1007/s00125-018-4708-x
46. Liu Q, Yu J, Wang L, et al. Inhibition of PU.1 ameliorates metabolic dysfunction and non-alcoholic steatohepatitis. *J Hepatol*. 2020;73(2):361-370. doi:10.1016/j.jhep.2020.02.025
47. Lin L, Pang W, Chen K, et al. Adipocyte expression of PU.1 transcription factor causes insulin resistance through upregulation of inflammatory cytokine gene expression and ROS production. *Am J Physiol - Endocrinol Metab*. 2012;302(12). doi:10.1152/ajpendo.00462.2011
48. Dhana K, Braun KVE, , Jana Nano, Trudy Voortman, Ellen W. Demerath W, Guan, Myriam Fornage, Joyce B.J. van Meurs, Andre G. Uitterlinden, Albert Hofman OH, Franco AD. An Epigenome-Wide Association Study of Obesity-Related Traits. *Am J Epidemiol*. 2017;186(2):227-236. doi:10.1093/aje/kwy025/4995327
49. Yang X, Jansson PA, Nagaev I, et al. Evidence of impaired adipogenesis in insulin resistance. *Biochem Biophys Res Commun*. 2004;317(4):1045-1051. doi:10.1016/j.bbrc.2004.03.152
50. Karczewska-Kupczewska M, Stefanowicz M, Matulewicz N, Nikołajuk A, Straczkowski M. Wnt signaling genes in adipose tissue and skeletal muscle of humans with different degrees of insulin sensitivity. *J Clin Endocrinol Metab*. 2016;101(8):3079-3087. doi:10.1210/jc.2016-1594

51. Frendo-Cumbo S, Jaldin-Fincati JR, Coyaud E, et al. Deficiency of the autophagy gene ATG16L1 induces insulin resistance through KLHL9/KLHL13/CUL3-mediated IRS1 degradation. *J Biol Chem*. 2019;294(44):16172-16185. doi:10.1074/jbc.RA119.009110
52. Yu R, Cruz-Soto M, Calzi SL, Hui H, Melmed S. Murine pituitary tumor-transforming gene functions as a securin protein in insulin-secreting cells. *J Endocrinol*. 2006;191(1):45-53. doi:10.1677/joe.1.06885
53. Manyes L, Arribas M, Gomez C, Calzada N, Fernandez-Medarde A, Santos E. Transcriptional profiling reveals functional links between RasGrf1 and Pttg1 in pancreatic beta cells. *BMC Genomics*. 2014;15(1):1-20. doi:10.1186/1471-2164-15-1019
54. Matthews AL, Koo CZ, Szyroka J, Harrison N, Kanhere A, Tomlinson MG. Regulation of leukocytes by TspanC8 tetraspanins and the “molecular scissor” ADAM10. *Front Immunol*. 2018;9(JUL):1-9. doi:10.3389/fimmu.2018.01451
55. Adamson SE, Montgomery G, Seaman SA, Peirce-Cottler SM, Leitinger N. Myeloid P2Y2 receptor promotes acute inflammation but is dispensable for chronic high-fat diet-induced metabolic dysfunction. *Purinergic Signal*. 2018;14(1):19-26. doi:10.1007/s11302-017-9589-9
56. Zhang Y, Ecelbarger CM, Lesniewski LA, Müller CE, Kishore BK. P2Y2 Receptor Promotes High-Fat Diet-Induced Obesity. *Front Endocrinol (Lausanne)*. 2020;11(June):1-19. doi:10.3389/fendo.2020.00341
57. Merz J, Albrecht P, von Garlen S, et al. Purinergic receptor Y2 (P2Y2)- dependent VCAM-1 expression promotes immune cell infiltration in metabolic syndrome. *Basic Res Cardiol*. 2018;113(6). doi:10.1007/s00395-018-0702-1
58. Kenefeck R, Narendran P, Walker LSK, et al. Follicular helper T cell signature in type 1 diabetes Find the latest version : Follicular helper T cell signature in type 1 diabetes. *J Clin Invest*. 2015;125(1):292-303. doi:10.1172/JCI76238.although
59. Wang Q, Zhai X, Chen X, Lu J, Zhang Y, Huang Q. Dysregulation of circulating CD4+CXCR5+ T cells in type 2 diabetes mellitus. *Apmis*. 2015;123(2):146-151. doi:10.1111/apm.12330
60. O’Neill S, Bohl M, Gregersen S, Hermansen K, O’Driscoll L. Blood-Based Biomarkers for Metabolic Syndrome. *Trends Endocrinol Metab*. 2016;27(6):363-374. doi:10.1016/j.tem.2016.03.012

Figures

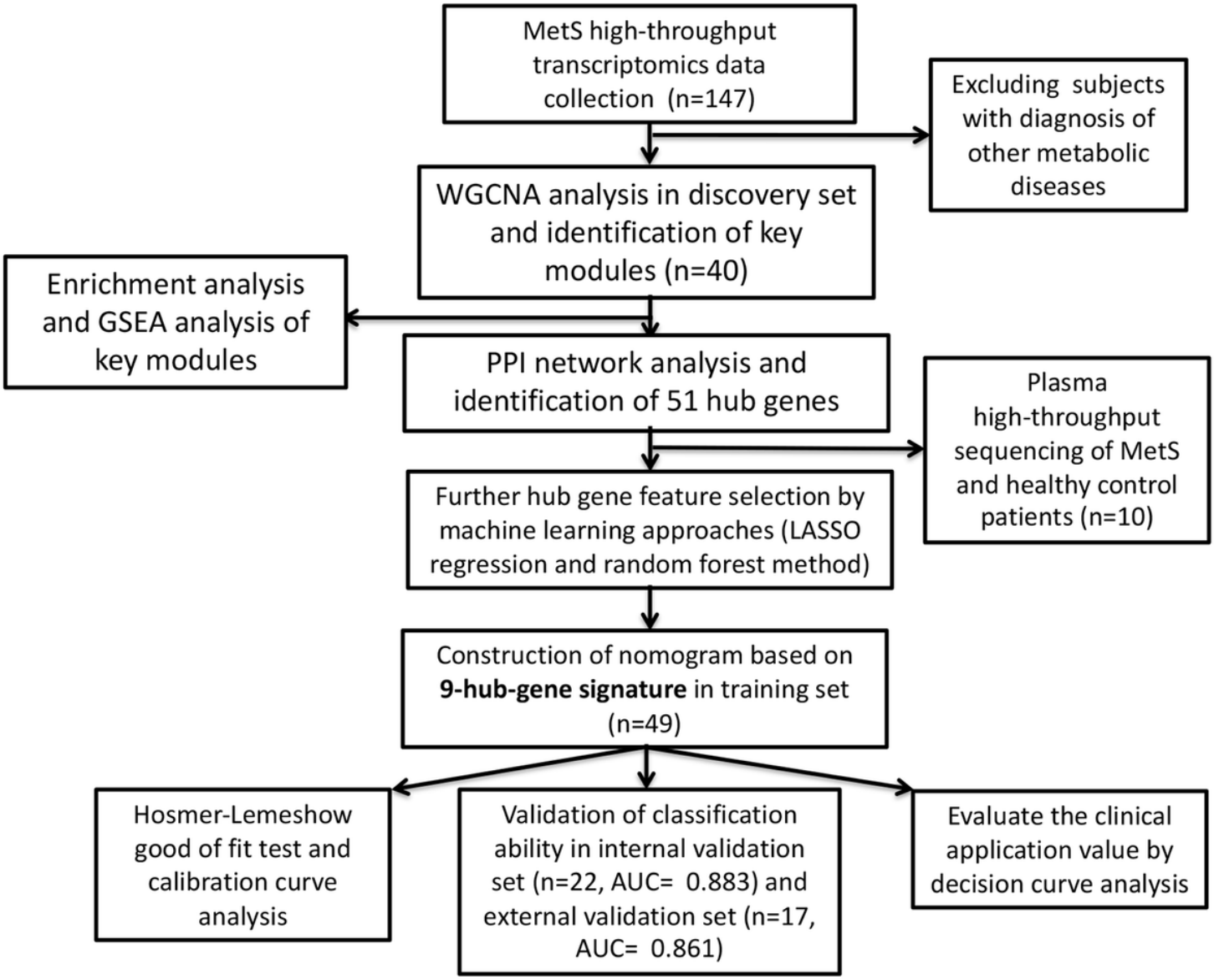

Figure 1

Flow chart of data processing and analysis.

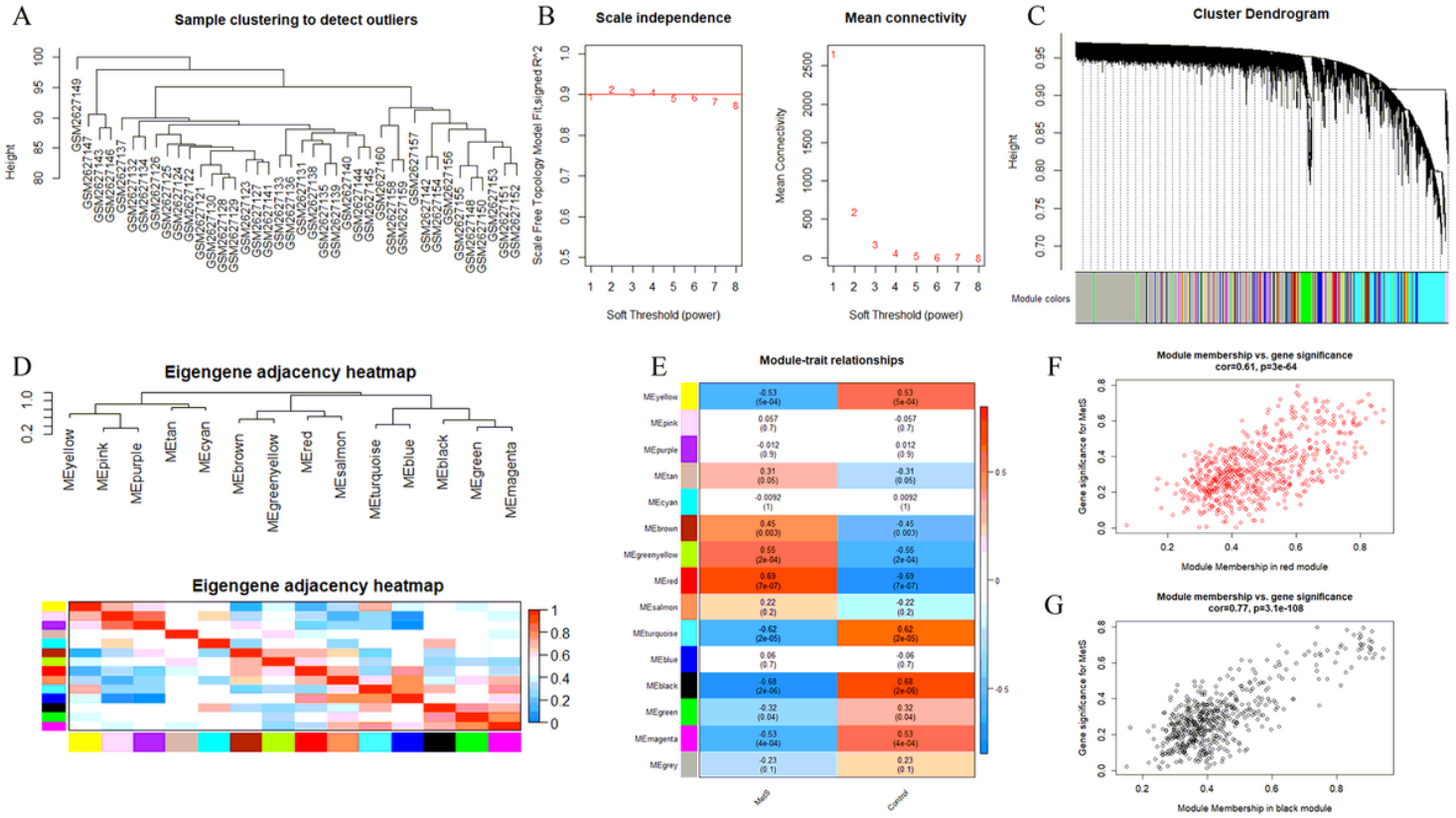

Figure 2

Weight gene correlation network analysis (WGCNA). (A) Sample clustering dendrogram and outliers detection. (B) Selection of the soft threshold. Scale-free topology fitting index R2 analysis (left) and mean connectivity for various soft threshold powers (right). The red line in the left panel means R2 = 0.9. (C) Clustering diagram of gene modules represented by different colors. (D) Clustering tree of gene modules and the correlation heatmap of the module eigengenes. (E) Heatmap of the relationship between modules and MetS: red for positive correlation and blue for negative correlation. (F,G) Scatter diagrams of genes in red module and black module. X-axis represents gene significance and y-axis represents module membership.

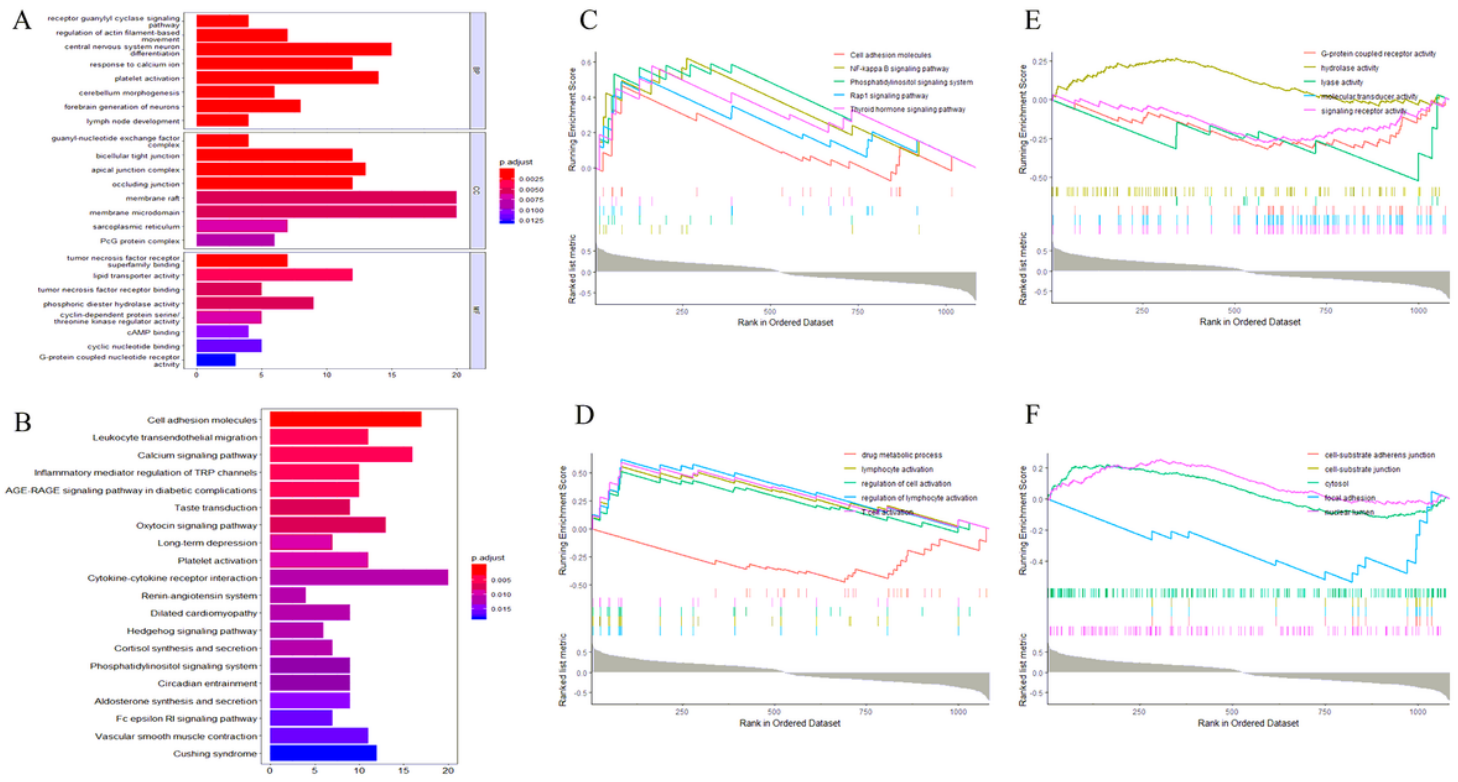

Figure 3

(A) Enrichment analysis of Gene Ontology (GO) function. (B) Enrichment analysis of Kyoto Encyclopedia of Genes and Genomes (KEGG) signaling pathway. The color represents the P value and X-axis represents gene number. (C) Gene Set Enrichment Analysis (GSEA) of KEGG signaling pathway. (D) Gene Set Enrichment Analysis of biology process (BP). (E) Gene Set Enrichment Analysis of molecular function (MF). (F) Gene Set Enrichment Analysis of cellular component (CC)

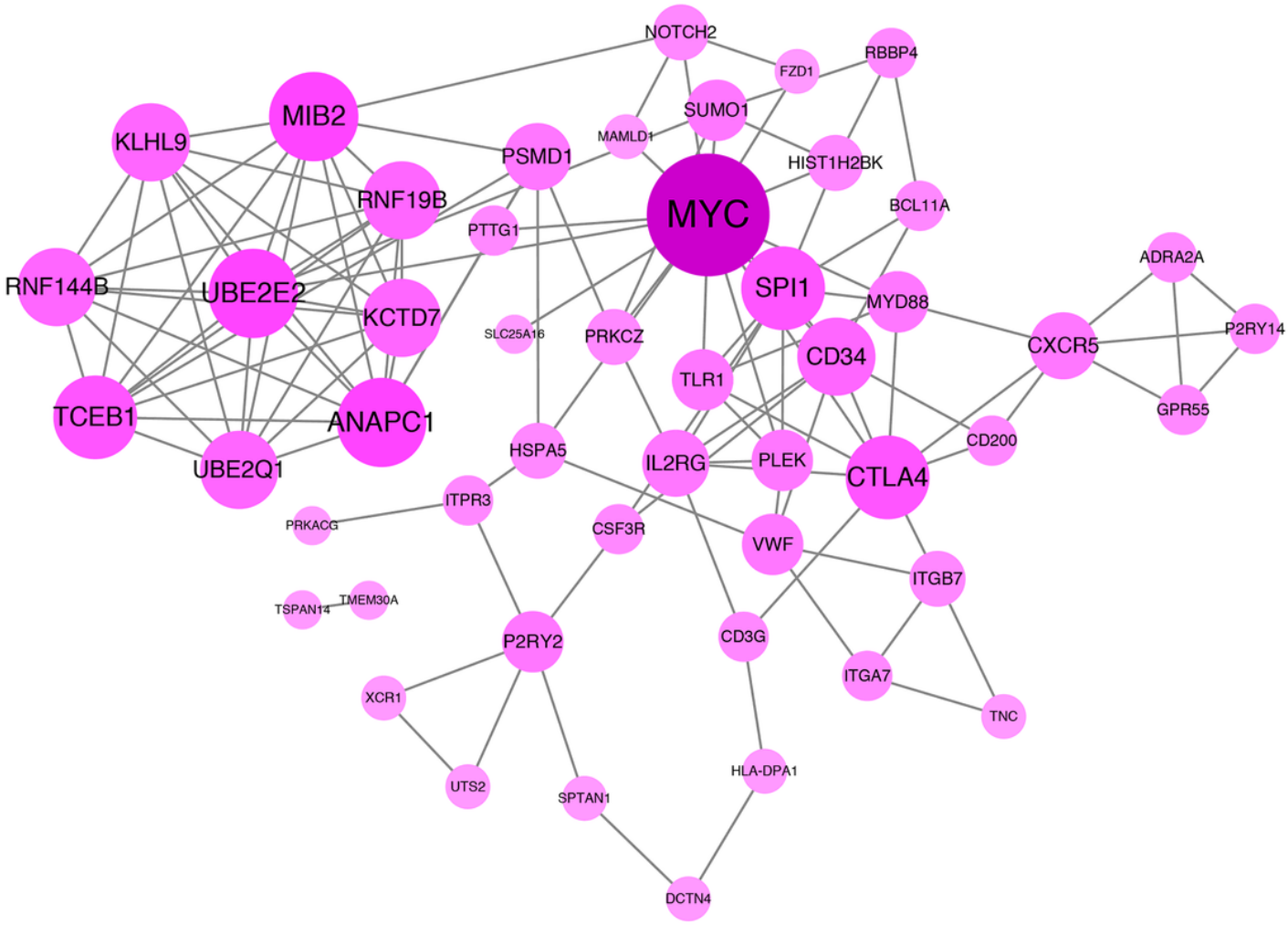

Figure 4

Protein-protein interaction (PPI) network. The gradual color and spot size represents the connectivity degree.

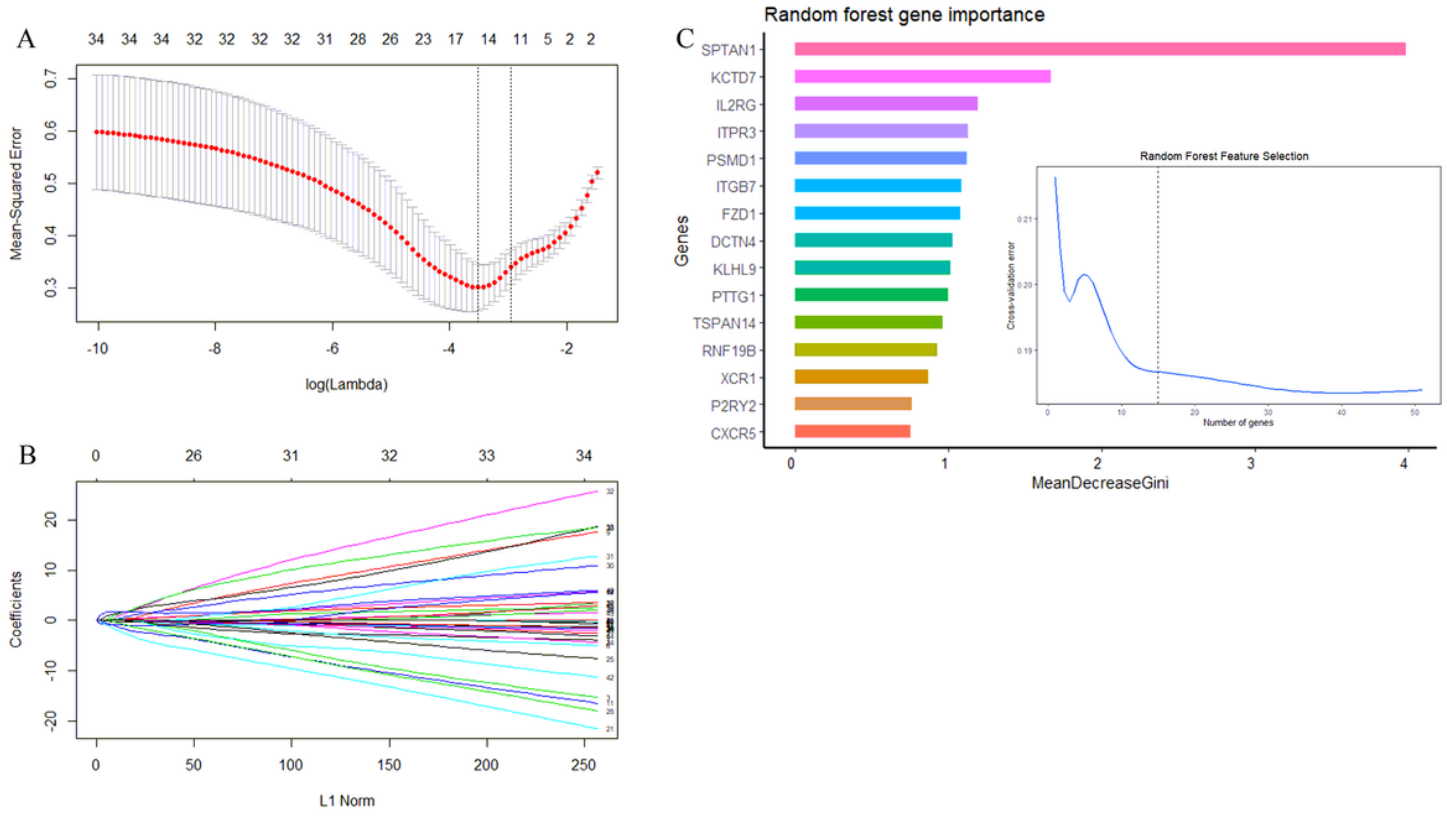

Figure 5

(A) The mean-squared error of LASSO regression. Y-axis represents mean-squared error. X-axis represents the ideal gene feature amount on various of lambda value. Left dotted line means the minimum of mean-squared error and the right dotted line means one standard deviation above minimum of mean-squared error. (B) Coefficients distribution trend of LASSO regression. (C) The importance of hub gene features based on random forest algorithm and the ideal gene feature amount.

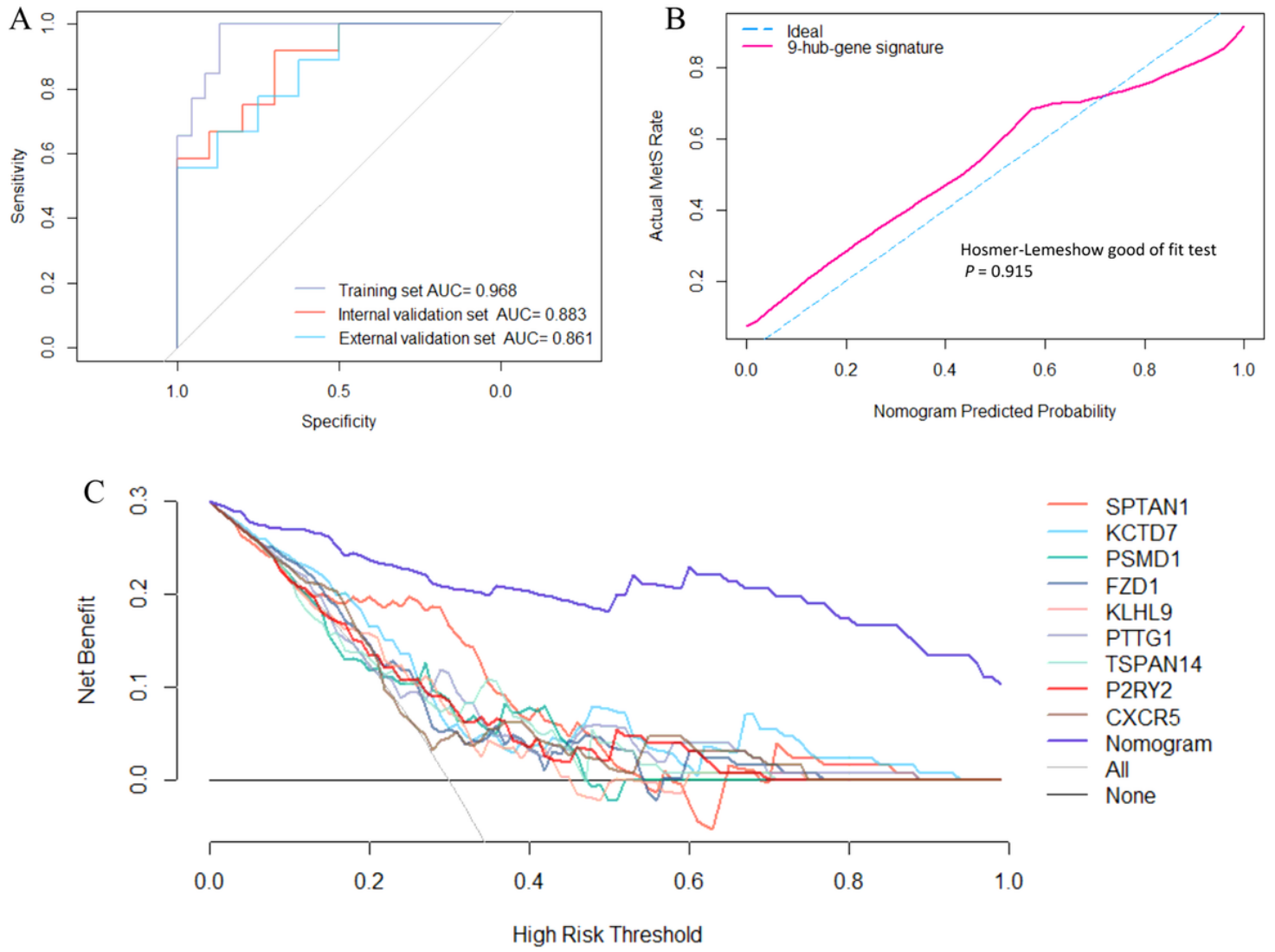

Figure 6

(A) Receiver Operating Characteristic curves of the web nomogram calculator based on the 9-hub-gene signature. (B) Calibration curve analysis and Hosmer-Lemeshow good of fit test of the web nomogram calculator based on the 9-hub-gene signature. (C) Decision curve analysis of every single gene feature and the web nomogram calculator based on the 9-hub-gene signature.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1.xlsx
- TableS2.xlsx
- TableS3.xlsx
- TableS4.xlsx
- FigureS1.tif