WGCNA-Based Identification of Hub Genes and Key Pathways Involved in Obesity

Yin Yuan  
Tianjin University of Traditional Chinese Medicine

Xuan Sun  
Tianjin University of Traditional Chinese Medicine

Jinyuan Wang  
Tianjin University of Traditional Chinese Medicine

Zixuan Wu  
Tianjin University of Traditional Chinese Medicine

Zhikui Tian  
Tianjin University of Traditional Chinese Medicine

Xianyue Zeng  
Tianjin University of Traditional Chinese Medicine

Naijin Zhang  
Tianjin University of Traditional Chinese Medicine

Hongwu Wang (✉️ tianjinwh_w@163.com)  
Tianjin University of Traditional Chinese Medicine

Article

Keywords: obesity, WGCNA, machine learning, GEO

Posted Date: June 12th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2962268/v1

License: ☇️ ❶ This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License
Abstract

The prevalence of obesity is increasing, yet its pathogenesis and therapeutic targets are still unclear, and the efficacy of existing therapeutic modalities is limited. Therefore, it is crucial to elucidate the molecular mechanisms underlying the pathogenesis of obesity and to explore potential molecular targets for obesity drug therapy. The expression dataset (GSE73304) was downloaded from the Gene Expression Omnibus database (GEO), and the data were divided into C and P groups, which were normalized for differentially expressed genes analysis (DEGs), Gene Set Enrichment Analysis (GSEA) and weighted gene co-expression network analysis (WGCNA). The intersecting genes obtained by DEGs and WGCNA were analyzed by three machine learning methods (LASSO, RandomForest, SVM-REF), and the gene ROC curves, genetic differences between groups and correlations between genes were analyzed after obtaining the major genes. Based on WGCNA and machine learning, this study found that several genes were significantly different between healthy and obese populations and closely associated with multiple molecular mechanisms, and these genes may serve as potential targets for drug therapy and diagnostic biomarkers, which need further studies to elucidate.

Introduction

The prevalence of obesity is high in many age groups. With the improvement of socio-economic level and the change of people's life style, the invention rate of obesity has been increasing in recent years. Obesity is determined by the value of body mass index (BMI), and it is generally considered that a BMI > 25.0 kg/m$^2$ is diagnosed as obesity. In Asia, more than 1.5 billion people have been suffering from obesity, which has increased the burden of health care and social economy[1]. The main symptoms of obesity are weight gain, excessive accumulation of visceral fat, hyperlipidemia, etc[2]. In addition, obesity is a direct threat to human health and increases the risk of secondary diseases such as type 2 diabetes, non-alcoholic fatty liver, coronary atherosclerosis and tumors[3]. Obesity is a chronic systemic metabolic disease in which the excess adipose tissue produced by obesity produces inflammation through the secretion of cytokines, while the neuroendocrine and metabolic energy balance systems act as a basic survival mechanism to resist the loss of stored fat. Obesity is therefore a pathological metabolic state that disrupts energy homeostasis and causes damage to other organs and systems.

At present, the means of treating obesity is relatively single, mainly through adjusting lifestyle, adjusting diet structure and enhancing exercise to achieve the purpose of reducing weight and treating obesity[4]. In addition, orlistat and other drug treatment, surgery and other methods can also treat obesity, but because of the various causes of obesity, the target of the disease is still unclear and other factors, the above treatment methods have limited effect. Therefore, finding a stable and effective target for obesity treatment is an important basis for its treatment.

Weighted gene co-expression network analysis (WGCNA) is a method to analyze the gene expression patterns of multiple samples. It can cluster and form modules of genes by similar gene expression patterns and analyze the relationship between modules and specific features[5]. Machine learning, such
as Lasso logistic, SVM-RFE, and random forest, are also new methods to analyze the correlation between gene expression and genes. In this study, we analyzed the GSE73304 gene expression synthesis database in GEO database, divided the data into healthy population samples (Group C) and obese population samples (Group P), and constructed a gene co-expression network and performed modular analysis based on the WGCNA algorithm, assuming that the gene expression network follows a scale-free distribution, and analyzed key genes by machine learning, followed by analyzing information such as ROC curves and intergroup differences of key genes in order to analyze information such as major genes and related molecular pathways involved in obesity.

Materials and methods

1. Data source and search

We utilized the Gene Expression Omnibus (GEO) database and searched GEO for genomics experiments on obesity. Data selection for this study included the following selection criteria: (1) samples included obese patients and healthy populations, (2) biology was limited to Homo sapiens, (3) raw or processed data were publicly available and accessible, and (4) total sample size was greater than 20. Gene expression profiles of GSE73304 were selected for analysis in this study. The raw data of GSE73304 was downloaded from the GEO website. Correction was performed using normalizeBetweenArrays, and the data were normalized. DEGs analysis was performed on the samples. Differentially expressed genes with P.Value < 0.05 were obtained as screening conditions using the limma package in R. The final results were represented by volcano and heat maps.

2. GSEA analysis

GSEA is an analysis method for whole genome expression profile microarray data. It can identify functionally enriched genes by comparison with predefined gene sets. A gene set is a set of genes with the same localization, pathway, function or other characteristics. GSEA is performed using the clusterProfiler package (version 3.5). The fold change in gene expression between the two sets is calculated and a gene table is generated based on the change in |log2FC|. GSEA-based KEGG pathway enrichment analysis was performed by the clusterProfiler package, with adjusted p-values < 0.05 as the threshold value. The top 5 and bottom 5 GSEA results were taken for mapping after sorting, respectively.

3. WGCNA network construction and module identification

The "WGCNA" package in R software is used for network construction and visualization. (1) clustering of samples; (2) analysis of network topology using the pickSoftThreshold function in WGCNA; the connectivity between genes in the gene network satisfies the scale-free network distribution when the scale-free R2 = 0.9; (3) detection of modules using hierarchical clustering and dynamic tree cutting functions; (4) calculation of gene significance (GS) and module affiliation (MM) to correlate modules with clinical features. MM) to correlate modules with clinical features. The corresponding module gene information was extracted for further analysis. Finally, we visualized the network of feature genes.
4. Functional enrichment analysis.
The significant genes obtained from WGCNA were intersected with the significant genes obtained from the difference analysis. Gene ontology analysis (GO) was used to identify the biological trait properties of the intersected genes. The functional properties of the intersecting genes were determined by Kyoto Gene and Genome Encyclopedia (KEGG) pathway enrichment analysis. Significance was set at P < 0.05.

5. Machine learning algorithm diagnosis

The intersecting genes obtained in 4 were analyzed by three machine learning algorithms[9–11]. Lasso logistic regression is a machine learning method that identifies variables by finding the λ value that minimizes the classification error, and the glmnet package of R was used as a binomial LASSO model for the dataset. SVM-RFE is a support vector machine based machine learning method that finds the best variables by subtracting the support vector machine generated feature vectors to find the best variables. Random Forest is a machine learning integration method to analyze major genes. Random forests use bootstrap methods to generate random samples from a dataset and replace them. These samples are divided into training samples (two-thirds of the sample set) and test samples (one-third of the sample set) for the predictive performance of the model. The gene sets derived from the three machine learning algorithms are intersected to obtain the key gene set.

6. Key gene analysis

The key genes derived in 5 were analyzed. Diagnostic column line plots were obtained in R by the packages survival, regplot, survminer, and ggDCA to analyze the expression of each key gene and the total expression. And the ROC curve and AUC area of key genes were analyzed by pROC, timeROC and other R packages. The functional differences and enrichment between genes in healthy and obese populations were analyzed by ssgsea, and correlations between key genes were analyzed and plotted by R packages such as GSVA, GSEABase, ggplot2, ggpubr, etc.

Results

1. Data downloading, collation and analysis of differential genes

We downloaded GSE73304 and its clinical performance data from the GEO database with the dataset platforms GPL11154 and GPL17077, 10 specimens each from the obese patient group and the normal body composition population, and analyzed them directly using the gene expression matrix provided on the website. The raw data were processed and 18,148 genes in the samples were analyzed and normalized to the raw data (Fig. 1A). Differential genes between the two groups (healthy population samples: C; obese population samples: P) were analyzed using the limma package, P < 0.05, 1937 differentially expressed genes (DIFF) were obtained, and DIFF volcano and heat maps were obtained (Fig. 1B, C).

2. GSEA analysis
A total of 32 prominent KEGG pathways were selected by GSEA analysis (Fig. 2). The activated pathways mainly included "Aminoacyl-tRNA biosynthesis", "Butanoate metabolism", "Folate biosynthesis", "Terpenoid backbone biosynthesis", "Mitophagy - animal", etc., most of which are closely related to biosynthesis. The inhibited signaling pathways are more diverse, such as "p53 signaling pathway", "Renin-angiotensin system", "Staphylococcus aureus infection" and "Neuroactive ligand-receptor interaction", "Complement and coagulation cascades", etc.

3.WGCNA analysis

Next, we performed the WGCNA analysis (Fig. 3). First, the samples were clustered to determine if there were any significant outliers (Fig. 3A). The height cut-off value was set to 60, and all samples were included in the analysis.

The genes screened from the GEO database were applied to the WGCNA network module. First, the network topology was analyzed using the pickSoft Threshold function in WGCNA; the connectivity between genes in the gene network satisfied the scale-free network distribution when the scale-free R2 = 0.9 (Fig. 3B).

The one-step network construction function of the WGCNA R package was used to construct the gene networks and perform module identification. To perform cluster splitting, the minimum module size was set to 50, deepSplit was set to 2, and finally 40 gene co-expression modules were constructed (Fig. 3C). Subsequently, the gene clusters were visualized and correlations between modules were analyzed (Fig. 3D, E, F). The results showed that the ORANGE module had the highest positive correlation with Obesity (correlation coefficient = 0.85, P = 3E-06).

4.Acquisition and functional enrichment of disease genes

The 45 major genes intersected by WGCNA and DIFF were subjected to GO and KEGG enrichment analysis (Fig. 4A). The results of GO analysis showed that BP was mainly enriched in negative regulation of cell development, neuroepithelial cell differentiation, camera-type eye development. CC was mainly enriched in respiratory chain complex IV, smooth endoplasmic reticulum, sarcoplasmic reticulum membrane, cytochrome complex. MF was mainly enriched in protein tyrosine phosphatase activity, growth factor activity, intronic transcription regulatory region sequence-specific DNA binding (Fig. 4B). KEGG analysis showed that genes were mainly enriched in Signaling pathways regulating pluripotency of stem cells, Dilated cardiomyopathy, Ovarian steroidogenesis, Longevity regulating pathway - multiple species, Insulin secretion and other pathways (Fig. 4C). Taken together, these data suggest a strong correlation between major genes and cell differentiation, mitochondria, and multiple endocrine factors and hormones.

5.Diagnosis by three machine learning algorithms

The main genes derived from LASSO, RandomForest, and SVM-REF analysis were analyzed and the intersecting genes derived from the three machine learning algorithms were analyzed (Fig. 5). The results showed that LASSO analysis yielded XLOC_004699, RIMBP2, COX6B2, OR5T1, RXFP2, etc. as the main
genes producing a role in the disease (Fig. 5A); Random Forest analysis showed that the main genes were XLOC_008079, XLOC_008408, VAX1, XLOC_007170, etc. (Fig. 5B); SVM-REF analysis RIMBP2, XLOC_008408, XLOC_004699, XLOC_000702, RXFP2, etc. were the main genes (Fig. 5C). The results obtained from the three analyses were intersected and Venny plots were drawn (Fig. 5D), and the results showed that 11 genes, including XLOC_004699, RIMBP2, COX6B2, OR5T1, RXFP2, XLOC_003676, XLOC_013038, VAX1, Q07610, XLOC_011515, and PTPN3 were the main disease genes.

6. ROC and differential gene analysis

The expression and total expression of each key gene were analyzed by diagnostic column line plots, and their ROC curves and AUC areas were obtained. Nomogram results showed that the total nomogram score was helpful to provide a quantitative method for predicting the effect of key genes on obesity (Fig. 6A, B). The ROC results showed that the AUC values of 11 key genes were all greater than 0.6, and the AUG value of PTPN3 was the highest (AUG = 8.90); COX6B2 had the smallest AUG value (AUG = 0.700) (Fig. 6C). Analysis of the expression differences of major genes in healthy and obese populations showed that eight genes, including XLOC_003676, XLOC_011515, VAX1, Q07610, RIMBP2, OR5T1, XLOC_013038, and PTPN3, had decreased expression in the obese population; XLOC_004699, COX6B2, and RXFP2 and three other genes had increased expression in the obese population (Fig. 6D).

7. Marker gene set analysis

Functional differences existed between genes in healthy and obese populations by ssgsea analysis (Fig. 7A). ns represents p-values between 0.2 and 1, # represents p-values between 0.05 and 0.2, * represents P < 0.05. Enrichment results showed that UV_RESPONSE_UP, MYC_TARGETS_V2 differed significantly between the two groups. PEROXISOME, BILE_ACID_METABOLISM, ANGIOGENESIS, HEME_METABOLISM, OXIDATIVE_PHOSPHORYLATION, FATTY_ACID_METABOLISM, XENOBIOTIC_METABOLISM, MYC_TARGETS_V1, PI3K_AKT_MTOR_SIGNALING, MTORC1_SIGNALING, HEDGEHOG_SIGNALING, and G2M_CHECKPOINT showed large differences between the two groups. The gene and function enrichment correlations were analyzed (Fig. 7B). The results showed the most significant correlations between PTPN3 and HEME_METABOLISM, RXFP2 and ANGIOGENESIS, and XLOC_004699 and ANGIOGENESIS. Correlations between key genes were analyzed (Fig. 7C). The results showed that the correlations between VAX1 and OR5T1, VAX1 and XLOC_003676, Q07610 and OR5T1, and PTPN3 and COX6B2 were high.

Discussion

Obesity is a significant cause of many diseases and is an important risk factor for human health. Reducing the incidence of obesity can help improve the incidence of metabolic diseases and other related diseases, such as tumors and cardiovascular diseases. At present, the molecular mechanisms underlying the pathophysiology of obesity are still unclear, and there are no complete treatment guidelines for clinical treatment. Therefore, it is crucial to explore the genes associated with obesity. The main genes known to be associated with obesity pathogenesis are C/EBPα, PPAR, SREBP, etc., regulating a wide
range of downstream genes such as fatty acid synthesis/metabolism[12–14]. In addition, studies have shown that mitochondrial autophagy, inflammation, apoptosis and other related pathways also regulate obesity. Therefore, the biomarkers of obesity are complex and diverse, and need to be fully explored and explored. In this study, the preliminary analysis showed that there were several differential genes (DIFF) between healthy and obese population samples, and they were enriched to several signaling pathways, such as "Aminoacyl-tRNA biosynthesis", "Butanoate metabolism" were significantly activated; "p53 signaling pathway", "Renin-angiotensin system", "Staphylococcus spp. "Staphylococcus aureus infection" signaling pathways were significantly inhibited. The existing findings show that the p53 signaling pathway regulates lipid homeostasis, including lipid transport and storage, fatty acid and cholesterol biosynthesis, and fatty acid desaturation, etc[15]. The Renin-angiotensin system plays an important role in the development of obesity, and the local Renin-angiotensin system is activated in the obese state[16].

The Renin-angiotensin system is activated in the obese state, thereby regulating obesity and inflammation.

To further identify key genes and pathways in the pathogenesis of obesity, we performed WGCNA, identified gene modules that differed significantly between the two groups, and identified 45 major genes that play an important role in the development of obesity by screening the intersection of WGCNA and DIFF, which were analyzed by GO and KEGG. The results of GO analysis showed that the major genes were associated with KEGG was enriched for Signaling pathways regulating pluripotency of stem cells, Ovarian steroidogenesis, Insulin secretion and other pathways. Insulin resistance is closely related to obesity[17, 18]. Insulin sensitivity decreases with increasing body fat content, while insulin resistance increases, further leading to the development of obesity. It has also been shown that altered levels of adipokines (e.g. leptin, lipocalin, etc.) in the obese state affect steroidogenesis and synergize with insulin resistance and elevated inflammatory status to affect oocytes and ovaries[19]. In addition, stem cells are closely associated with adipogenesis[20]. There are multiple mechanisms of excessive fat accumulation in obese patients, of which overproduction of adipocytes is an important mechanism of influence. Adipose tissue contains a large number of pluripotent stem cells, called adipose-derived stem cells (ASCs), capable of developing into mature adipocytes. The enhanced adipogenesis in obese people may be due to the enhanced ability of ASCs in obese people to undergo adipogenesis and differentiation. In summary, this study found that intersecting genes may contribute to the development of obesity through various pathways such as cell differentiation and regulation of multiple endocrine factors and hormone levels.

Second, we analyzed 45 intersecting genes of WGCNA and DIFF by LASSO, RandomForest, and SVM-REF, and the results obtained from the three analyses were taken to show that XLOC_004699, RIMBP2, COX6B2, OR5T1, RXFP2, XLOC_003676, XLOC_013038, VAX1, Q07610, XLOC_011515, PTPN3 and 11 other genes were the major genes for obesity. Subsequently, the ROC curves of the 11 major genes were analyzed and their expression differences in healthy population samples and obese population samples were compared, and finally the enrichment analysis of the major genes was performed. The results showed that eight genes, including XLOC_003676, XLOC_011515, VAX1, Q07610, RIMBP2, OR5T1, XLOC_013038 and PTPN3, had decreased expression in the obese population; three genes, including
XLOC_004699, COX6B2 and RXFP2, had increased expression in the obese population. The results of existing studies showed that COX6B2 was closely associated with pancreatic cancer development and metastasis, and was confirmed to be differentially expressed between groups in a study of colonic gene expression in obese rats in which flaxseed polysaccharide altered lipid metabolism and energy metabolism, indicating that it may have a role in affecting lipid metabolism[21]. The main functional enrichments involved in differentially expressed genes include PEROXISOME, BILE_ACID_METABOLISM, OXIDATIVE_PHOSPHORYLATION, and FATTY_ACID_METABOLISM. Peroxisomes are closely associated with obesity and play an important role in lipid metabolism, including fatty acid oxidation and plasma protein synthesis, and can interact with mitochondria, thereby regulating mitochondrial dynamics and lipid thermogenesis[22]. Its associated peroxisome proliferator-activated receptor is a key regulator of adipocyte differentiation and regulates insulin and adipokine production and secretion. Bile acids are important physiological agents for nutrient absorption and biliary secretion of lipids[23]. Bile acids regulate hepatic lipid metabolism through a variety of molecular signals to maintain metabolic homeostasis. A variety of factors, such as a high-fat diet, can alter bile acid metabolism, leading to metabolic disorders and obesity. Fatty acids play an important role in obesity and metabolic diseases[24], mainly through the regulation of malonyl coenzyme A synthesis by acetyl coenzyme A carboxylase 1 and 2 (ACC1 and ACC2) for fatty acid synthesis. The above findings show that obesity is regulated by multiple pathways and targets, and the main targets obtained by machine learning and their enriched molecular functions help us to conduct subsequent experimental studies.

In summary, the analysis of key genes and their enriched signaling pathways and molecular functions that lead to obesity by using WGCNA and various machine learning methods in this study is of guiding significance in exploring the pathogenesis of obesity and seeking new obesity treatment methods and drugs. However, this study still needs to be improved: firstly, a more rigorous analysis by expanding the sample size is needed to obtain more generalized results. Secondly, this study was not validated in animal experiments, and the results obtained from bioinformatics should be validated in experiments in the follow-up study. Finally, we still need to validate the bioinformatics results in the clinical setting.

This is the first study to use the WGCNA approach to construct co-expression networks combined with machine learning to explore obesity-related susceptibility modules and important genes. This study reveals that several modules play an important role in the etiology of obesity, and several genes such as RIMBP2, COX6B2, and RXFP2 are important influences in the development of obesity. This study helps to improve the genetic study of the etiology of obesity.

Data Availability statement

All data generated or analysed during this study are included in this published article [and its supplementary information files].

Declarations
Data Availability statement

All data generated or analysed during this study are included in this published article [and its supplementary information files].

Author contributions

Yin Yuan and Xuan Sun participated in all experimental work, analyzed data, and contributed to both artwork design and manuscript writing; Jinyuan Wang and Zixuan Wu responsible for code debugging and drawing; Zhikui Tian, Xianyue Zeng and Naijin Zhang responsible for collecting data and analyzing; Hongwu Wang proposed the idea for the paper.

Funding

National Key Research and Development Program of China (No. 2017YFC1703305). Tianjin Science and Technology Plan Project (No. 21ZLGCGX00030).

Competing interests

The authors declare no competing interests.

References


---

**Figures**

![Figure 1](image1.png)

**Figure 1**
Data variation analysis. (A) Raw data and normalized data. (B) DIFF volcano plot. Gray dots represent non-differentially expressed genes, blue dots represent down-regulated genes, and red dots represent up-regulated genes. (C) Heat map of DIFF. Red represents upregulated genes and blue represents downregulated genes.

Figure 2

GSEA pathway enrichment analysis.
Figure 3

Construction of coexpression modules. (A) Clustering dendrogram of samples based on their Euclidean distance. (B) Analysis of the scale-free fit index for various soft-thresholding powers ($\beta$) and analysis of the mean connectivity for various soft-thresholding powers. (C) The cluster dendrogram of genes in GSE73304. Each color represented a module. (D) Module–trait associations. Each row corresponds to a module, and each column corresponds to a trait. Each cell contains the corresponding correlation and P
value. The table is color-coded by correlation according to the color legend. (E) Visualization of the WGCNA network using a heatmap plot. The heatmap depicts the topological overlap matrix (TOM) among all modules included in the analysis. The light color represents a low overlap, and the progressively darker red color represents an increasing overlap. (F) Eigengene dendrogram and eigengene adjacency plot.

Figure 4

Functional enrichment analysis of WGCNA and DIFF crossover genes. (A) Venny diagram of WGCNA and DIFF intersection genes. (B) GO analysis of the genes. (C) KEGG analysis of genes.
Figure 5

Machine learning algorithm diagnosis. (A) Tenfold cross-validation for tuning parameter selection in the LASSO mode and LASSO coefficient profiles. (B) RandomForest analysis of major genes. (C) Tenfold cross-validation for SVM-REF analysis. (D) Machine learning Venny diagram.
Figure 6

Differential gene analysis. (A-B) Nomogram diagnostic column line graph. (C) Key gene ROC curves. (D) Comparison of key gene expression in samples from healthy and obese populations.
Figure 7

Marker gene set analysis. (A) Functional differences between genes in healthy and obese populations. (B) Gene and function enrichment correlations. (C) Correlation between key genes.

Supplementary Files

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

- GSE73304.txt