# Construction of diagnostic markers for hub lncRNAs in Parkinson's disease based on chip re-annotation

**Yi Quan**
Beijing Tiantan Hospital
**Shuo Wang**
Beijing Tiantan Hospital
**Jia Wang**
Beijing Tiantan Hospital
**Jizong Zhao** ( zhaojz1205@outlook.com )
https://orcid.org/0000-0002-5906-6149

**Research**

**Keywords:** LncRNA, WGCNA, SVM, Diagnostic biomarker, Parkinson's disease, KEGG pathway

**Posted Date:** April 28th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-435070/v1

## Abstract

# BACKGROUND

Parkinson's disease (PD) is a progressive neurodegenerative disease that is also the most common motor disorder and is accompanied by the loss of DA neurons in the brain. Long non-coding RNAs (lncRNAs) have recently been identified as new genetic entities that regulate cellular processes. One of the main functions of lncRNAs is the regulation of the expression of specific genes in multiple steps, including the regulation of transcriptional and post-transcriptional mechanisms and epigenetics.

# MATERIAL AND METHODS

Here we downloaded three sets of expression-spectrum data for PD from the GEO database. The data were re-annotated with R package, which were integrated into a set of expression profiles for the analysis of differentially expressed lncRNAs. Subsequently, lncRNA/mRNA co-expression modules were identified through a weighted co-expression analysis and lncRNAs were expressed based on binding differences. The diagnostic tags of PD were filtered with key modules and were finally used to build the PD diagnostic prediction model.

# RESULTS

Based on lncRNA re-annotation, a total of 1931 lncRNA expression values in the three sets of data were obtained and significant differences in expression ($P < 0.05$) for a total of 162 lncRNAs. A total of 21 modules were identified through WGCNA and five modules were selected. We screened 12 lncRNAs (AUC > 0.6) as PD diagnostic markers and as features to construct a SVM classification model. The model had good predictive ability in the training set and verification set (AUC of 0.9928 and 0.464, respectively), which illustrated their potential as diagnostic markers of PD.

# CONCLUSIONS

This study provided new molecular entities for the diagnosis of PD, which may promote the early detection of this disease and the development of personalized therapies.

## 1. Background

Parkinson's disease (PD) is a progressive disorder and one of the most common degenerative neurological diseases worldwide, after Alzheimer's disease (AD) [1]. Because of the diversity of the initial symptoms, PD diagnosis is highly difficult, resulting in confusion and delays in diagnosis and directly affecting the post-treatment stage. Therefore, there is an urgent need to identify useful biological markers of early-stage PD [2, 3]. The continued research on this subject included the exploration of body-fluid and

imaging markers; however, because of the drastic heterogeneity of this disease, no reliable biomarkers are available currently [4, 5].

The term non-coding RNA refers to functional RNA molecules that cannot be translated into proteins. Among them, common regulatory non-coding RNAs include small interfering RNAs (siRNAs), micro RNAs (miRNAs), Piwi-interacting RNAs (piRNAs), and long non-coding RNAs (lncRNAs). In particular, lncRNAs have become key regulators of different genetic regulatory layers. LncRNAs are typically expressed in more cell types and tissue-specific terms than are mRNAs or miRNAs; thus, they exhibit great advantages and are being prioritized as diagnostic and prognostic markers [6]. Increasing evidence suggests that lncRNAs have key biological functions in the brain, as lncRNAs have been associated with neurodegenerative diseases, such as AD and PD [7].

This research compared lncRNA data pertaining to a PD group from a database with that from the brain tissues (substantia nigra) of normal individuals using bioinformatics. The differential expression of lncRNAs was screened through the Gene Co-Expression Network (via WGCNA) and the SVM pattern was validated, thus laying a solid foundation for the identification of biological markers of PD.

## 2. Materials And Methods

# 2.1. LncRNA expression profiles in PD

Three sets of data were obtained from the HG-U133_Plus2 platform of the Gene Expression Omnibus (GEO) database, numbered GSE49036 [8], GSE20141 [9], and GSE7621 [10]. The date of download was 2019.1.5. The GSE49036 set contains 20 disease samples and eight control samples, GSE20141 contains 10 disease samples and eight control samples, and GSE7621 contains nine normal samples and 16 disease samples. We downloaded the original cell data of the three sets of data separately.

1) The robust multichip average (RMA) method of affy [11] was used to standardize the three sets of data; 2) The batch function was removed using the SVA package combat function in the R language; 3) Probes were mapped to genes/lncRNAs, with multiple probes corresponding to the median of a gene and one probe corresponding to the elimination of multiple genes. The data analysis process used here is shown in Supplementary Fig. S1.

# 2.2. LncRNA re-annotation

Based on the NetAffx annotation of the probes and the Refseq and Ensembl annotations of lncRNAs, we identified 2448 probes (1970 lncRNAs) that were represented on the Affymetrix HG-U133 Plus 2.0 arrays (S1.xls). Of these, 725 probes (510 genes) were annotated as lncRNAs by both the Refseq and the Ensembl databases; 512 probes (379 genes) were annotated only by the Refseq database, and 1211 probes (1081 genes) were annotated only by the Ensembl database. The probes that were annotated by both databases but had controversial definitions were excluded from our study.

## 2.3. Identification of differences between lncRNAs and mRNAs

To screen for genes and lncRNAs that were greatly changed in the PD samples, we used the limma package of the R software [12] to perform differential gene screening using a fold change < − 1.2 or > 1.2 and a *P* value < 0.05 as the threshold.

## 2.4. Identification of lncRNA/mRNA co-expression modules

To identify the PD-related lncRNA and mRNA co-expression modules, we first combined the lncRNA and mRNA expression profiles to further remove outlier samples and mRNA/lncRNA modules with a variance < 0.5, to obtain large-variation mRNA/lncRNA modules (DVGLs). The weighted co-expression network analysis finds a module that has a co-expression relationship with a lncRNA and uses Fisher's exact test to screen the modules that are significantly enriched in that lncRNA. Specifically, we used the WGCNA [13] package in R to construct a scale-free co-expression network for the DVGLs. First, Pearson's correlation matrices and an average linkage method were both applied to all pairwise DVGLs. Subsequently, a weighted adjacency matrix was constructed using a power function, A$mn$ = |C$mn$|$\beta$ (C$mn$ = Pearson's correlation between DVGL m and DVGL n; A$mn$ = adjacency between DVGL m and DVGL n). $\beta$ is a soft-thresholding parameter that emphasizes strong correlations and penalizes weak correlations between DVGLs. After choosing the power of β, the adjacency was transformed into a topological overlap matrix (TOM), which measures the network connectivity of a DVGL defined as the sum of its adjacency with all other DVGLs for the network DVGL ratio, and the corresponding dissimilarity (1-TOM) was calculated. To classify DVGLs with similar expression profiles into DVGL modules, average linkage hierarchical clustering was conducted according to the TOM-based dissimilarity measure with a minimum size (DVGL group) of 30 for the DVGL dendrogram. To analyze the module further, we calculated the dissimilarity of module eigen DVGLs, chose a cut-off value for the module dendrogram, and merged several modules. Finally, the number of lncRNAs and mRNAs in each module was counted. Fisher's exact test was used to identify modules with significant enrichment of lncRNAs.

## 2.5. Functional enrichment analyses

Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were performed using clusterProfiler of the R package [14] for genes associated with lncRNAs in modules with significant enrichment, to identify over-represented KEGG pathways. In both analyses, statistical significance was set at *P* < 0.05.

## 2.6. Identification of co-expression modules related to Parkinson's disease

DisGeNET (http://www.disgenet.org) is a discovery platform that contains one of the largest publicly available collections of genes and variants associated with human diseases. Genes related to PD were screened in the DisGeNET database, and genes related to PD, Parkinsonian tremor, Parkinsonian

Disorders, and Parkinsonism were searched based on keywords; a total of 1598 genes were collected (S6.xlsx). After further filtering, 1168 unique IDs were retained, including 1040 genes that appeared in our data set. We counted the number of disease genes in each module, as shown in Table 1, where the *P* value represents the degree of significant aggregation of disease genes in this module. Subsequently, we determined the mRNA and Parkinson gene correlation of each module by Fisher's test.

## 2.7. Identification of lncRNA diagnostic markers

First, we selected the intersection of lncRNA and differential lncRNA genes in the disease-related and lncRNA-enriched modules and calculated the area under the receiver operating characteristic (ROC) curve (AUC) of each lncRNA. We then screened lncRNAs with an AUC > 0.6 as candidate diagnostic markers for PD and used a literature-mining approach to analyze the candidate diagnostic markers regarding their relevance to PD.

## 2.8. Construction and validation of a PD lncRNA diagnostic model

We used the PD-related candidate lncRNAs to build a diagnostic prediction model based on support vector machine (SVM) [15] classification to predict PD and normal healthy samples. The SVM classification is a supervised learning model in machine learning algorithms that can be used to analyze recognition patterns in data. An SVM constructs a hyperplane that can be used for classification and regression in high or infinite dimensional space. Given a set of training samples, each label belongs to two categories, and an SVM training algorithm establishes a model by assigning a new instance to one class or another, thus rendering it a non-probabilistic binary linear classification. We randomly divided all samples into a training data set and a verification data set. The model was built using the training data set based on the 10-fold cross-validation verification model classification ability. Subsequently, we used the established model to predict the samples in the validation data set. The model's predictive ability was evaluated using the AUC, and the model's predictive sensitivity and specificity for PD were analyzed. The 71 samples included in the set were randomly divided into two sets, the training data set and the verification data set, while ensuring that the ratio of PD to normal control samples was similar between the two data sets. The training data set contained 35 samples (23 PD samples and 12 normal control samples) and the validation data set contained 36 samples (23 PD samples and 13 normal control samples).

## 2.9. KEGG pathway enrichment analysis of lncRNAs

To assess further the function of the lncRNAs identified in the previous analyses, we used the single-sample gene set enrichment analysis (ssGSEA) [16] method of the GSVA package of the R software to perform a KEGG pathway enrichment analysis for each sample through gene expression profiling. We obtained an enrichment score for each pathway and further calculated the relevance of the lncRNA expression. The 20 most relevant KEGG pathways were selected.

## 2.10. Statistical analysis

All analyses that do not specify parameters used the default parameters of the software. Significance was set at $P < 0.05$. All statistical analyses were performed in R 3.4.3.

# 3. Results

## 3.1. Data processing

Each set of data obtained from the GEO database was pre-normalized and post-normalized using RMA. The results are shown in Supplementary Fig. S2. The results of lncRNA re-annotation are listed in S1.xlsx. Next, we extracted the three sets of data and finally obtained a combined data set.

## 3.2. Differential lncRNA and differential gene analysis

After removing the batch effect on the three sets of data, we extracted lncRNA probes according to the annotation information and transformed the probe ID into Gene-Symbol. Finally, the differences of 162 lncRNAs was identified, 91 of them were upregulated and 71 were downregulated in PD samples. Using the pheatmap R package to draw a heat map (Fig. 1), we can see from the diagram that these lncRNAs differ in PD from normal control samples.

## 3.3. Co-expression analysis of lncRNAs and genes

A sample clustering analysis identified two abnormal samples, as shown in Fig. 2A. After removing the outlier samples (69 samples), 9755 genes/lncRNAs were finally obtained. Pearson's correlation coefficient was then used to calculate the distance between each gene and the lncRNA, and the WGCNA package of the R software was used to construct the weighted co-expression network and to select the soft threshold, as depicted in (Fig. 2B and C). Finally, 21 co-expression modules were identified (Fig. 2D). The number of differential lncRNAs in each module is shown in Table 2, where *P* values represent the degree of significant aggregation of the differential lncRNAs in this module. We found that five modules (black, blue, midnight-blue, tan, and turquoise) were significantly related to the differential lncRNAs.

The five modules were enriched in multiple KEGG pathways, with little intersection between these pathways, as shown in Fig. 3A. This suggests that different modules may perform different functions. The turquoise, blue, and black modules all contained the classic PD pathways. In addition, the first two Neuroactive ligands with the highest gene ratio among the 23 pathways that were enriched in black – receptor interaction, Dopaminergic synapse pathway. There are also reports in the literature [17, 18] were the most prominent pathways enriched in quantiles with PD miRNA patterns (e.g., Fig. 3E). The main pathways in another enrichment module (tan module) (e.g., Fig. 3C) were Alcoholism and PD, Drug addictions, etc., whereas Dopamine neurotransmission impairment underlies a wide range of disorders with motor control deficiencies [19].

## 3.4. Mining of disease-related modules

We obtained a total of 1168 genes from the DisGeNET database, among which 1040 genes underwent expression analysis. Moreover, the correlation between the expression of these 1040 genes and co-expression modules was assessed. We found that nine modules were significantly related to these genes (Supplementary Fig. 3 black and blue). In these three modules other with turquoise, the lncRNAs and differential lncRNAs were also significantly correlated. This module contained lncRNAs/mRNAs (GeneModuleClass.xlsx).

## 3.5. Screening of key lncRNAs

Using SVM to analyze the differential lncRNAs in the disease-related modules, 12 candidate lncRNAs were finally identified, as shown in Table 3. These lncRNAs exhibited high classification performance, with an average AUC > 0.6; thus, they were deemed potential lncRNA diagnostic markers of PD.

## 3.6. Construction and testing of a PD lncRNA diagnostic model

The 71 samples were randomly divided into two groups, the training data set (n = 35, 23 PD samples and 12 normal samples) and the validation dataset (n = 36, 23 PD samples and 13 normal samples). We used the 12 lncRNAs identified above as features in the training data set, to obtain their corresponding expression profiles and build an SVM classification model. The model was tested using a 10-fold cross-validation method. The classification accuracy rate was 94.28% (Fig. 4A), and 33 out of 35 samples were classified correctly. The sensitivity of the model regarding the identification of PD samples was 100%, with a specificity of 83.33% and an AUC of 0.9928 (Fig. 4C). Furthermore, the established model was used to predict the samples in the verification data set and test the predictive ability of the model. Thirty out of 36 samples were classified correctly, with a classification accuracy of 83.3%. The sensitivity of the model to PD was 86.95%, the specificity was 76.92% (Fig. 4B), and the AUC was 0.9464 (Fig. 4D). These results indicate that the diagnostic prediction model constructed in this study can effectively distinguish patients with PD from normal control populations, and that the 12 lncRNAs identified here can be used as reliable biomarkers for PD diagnosis.

## 3.7. KEGG pathway analysis of the 12 lncRNAs

To assess the function of each of the lncRNAs identified here, we analyzed the 12 most relevant pathways of lncRNA expression and found that seven lncRNAs were related to the Parkinson's disease pathway. Among them, AC093323.3 and COPG2IT1 showed a positive correlation, whereas AC120114.3, LOC153684, NCRNA00107, RP11-10O22.2, RP11-417J1.4, etc. were negatively correlated (Supplementary Fig. S4).

# 4. Discussion

The data pertaining to brain tissues (substantia nigra) of patients with PD were specifically selected here for data mining. We obtained the gene expression data of PD from the GEO database. Compared with the

TCGA database, the GEO data are scattered; therefore, we were only able to collect the data manually. Via chip re-annotation, 1970 lncRNA probes were obtained to study the regulatory mechanism of the mRNA/lncRNA co-expression network in PD and the possible regulatory mechanisms of disease pathways. Such a large sample in the study of PD is unique and will help improve the reliability of the research results [20].

The co-expression network analysis performed here identified five WGCNA modules, among which the midnight-blue module was most significantly enriched in neurodegenerative diseases. Furthermore, an enrichment analysis showed that the Parkinson's disease (PD) pathway (hsa05012) was one of the representative pathways related to this module. In addition, the remaining modules exhibited low intersection and were enriched in different pathways with different functions, such as the turquoise module, which was enriched for the *Helicobacter pylori* infection (hsa05120) and epithelial signal pathways; previous studies have reported that *Helicobacter pylori* infection is associated with PD [21] or that the ubiquitin-mediated proteolytic pathway (hsa04120) in astrocytic glutamine metabolism is associated with PD [22]. Thus, we inferred that this is a pathogenic entity. Because of the complexity of its underlying mechanisms, PD is a complex disease that cannot be attributed to the dysfunction of a single pathway. Therefore, further data mining was performed on the disease-related modules and 12 key lncRNAs were selected using the ROCR package in the R language. A PubMed literature search system was used for literature Dig and to explore the relationship between these lncRNAs and PD. According to previous reports, AC093323.3 exhibits differential expression in the midbrain of cocaine abusers [23], which shows that these 12 genes can be used as potential lncRNA diagnostic markers of PD. Subsequently, we used the SVM model to perform a disease prediction analysis of the 12 candidate lncRNAs. The 10-fold cross-validation of the PD dataset showed that our model had 12 lncRNAs. The sensitivity of RNA verification was 86.95% and the specificity was 76.92%, which further showed that these 12 lncRNAs can be used as reliable biomarkers for PD diagnosis. Finally, the 12 selected lncRNAs were re-analyzed through the KEGG pathway database. We identified 10 positive correlations for AC093323.3, including cancer-related pathways such as PD, and 10 negative correlations, mainly related to the JAK/STAT signaling pathway; several negative correlations for AC120114.3, including Huntington's disease, AD, PD, and other related pathways; and a negative correlation between LOC153684 and the AD pathway.

We analyzed the lncRNA/mRNA network and related pathways in PD using bioinformatics techniques. These results can help understand the occurrence and development of PD. However, our research also had some limitations. First, we used probes to re-annotate the pipeline and identify functional lncRNAs related to PD; although this approach has been widely used in many bioinformatics studies, we admit that this pipeline filters out many lncRNAs that do not match the probe sequence. Second, in addition to gene expression, epigenetic- and protein-level information also plays a very important role in the drug-response mechanism; therefore, this information should be included in the pre-expression model. Third, in the field of bioinformatics, the validity of the results is often assessed based on statistical significance and literature verification, which were used here to validate the accuracy and reliability of the

lncRNA/mRNA network, lncRNA-related functional modules, or the diagnostic potential of the lncRNA biomarkers.

In this study, we used the GEO database to analyze systematically potential lncRNA molecular markers in PD based on lncRNA re-annotation. We screened out 12 lncRNA molecules and verified them through the SVM model, to obtain satisfactory results. It is concluded that the expression of these 12 lncRNAs may be related to the occurrence and development of PD. This study provided new molecular entities for the diagnosis of PD, which may promote the early detection of this disease and the development of personalized therapies.

# 5. Conclusions

This research provides new molecular features for Parkinson's diagnosis through database screening and machine learning verification model, which is helpful for early Parkinson's diagnosis and personalized treatment.

# Declarations

### Ethics approval and consent to participate.

Not applicable

### Consent for publication

Not applicable

### Availability of data and materials

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

### Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

### Author Contributions

Yi Quan had full access to all of the study data and takes responsibility for the integrity and accuracy of the data analysis; Study concept and design: Yi Quan, Jia Wang; Critical revision of the manuscript for important intellectual content: All authors; Statistical analysis: Yi Quan, Jia Wang; Administrative, technical, and study supervision: Shuo Wang, Jizong Zhao

### Funding

Not applicable

### Acknowledgments

We thank the Central laboratory of China National Clinical Research Center for Neurological Diseases for technical support.

# References

1. Miller DB, O'Callaghan JP. Biomarkers of Parkinson's disease: present and future. *Metabolism.* 2015;64(3 Suppl 1):S40-46.
2. Silveira-Moriyama L, Sirisena D, Gamage P, Gamage R, de Silva R, Lees AJ. Adapting the Sniffin' Sticks to diagnose Parkinson's disease in Sri Lanka. *Mov Disord.* 2009;24(8):1229-1233.
3. The Lancet N. Biomarker promise for Parkinson's disease. *Lancet Neurol.* 2010;9(12):1139.
4. Khan AR, Hiebert NM, Vo A, et al. Biomarkers of Parkinson's disease: Striatal sub-regional structural morphometry and diffusion MRI. *Neuroimage Clin.* 2019;21:101597.
5. Yilmaz R, Hopfner F, van Eimeren T, Berg D. Biomarkers of Parkinson's disease: 20 years later. *J Neural Transm (Vienna).* 2019;126(7):803-813.
6. Cheetham SW, Gruhl F, Mattick JS, Dinger ME. Long noncoding RNAs and the genetics of cancer. *Br J Cancer.* 2013;108(12):2419-2425.
7. Wu P, Zuo X, Deng H, Liu X, Liu L, Ji A. Roles of long noncoding RNAs in brain development, functional diversification and neurodegenerative diseases. *Brain Res Bull.* 2013;97:69-80.
8. Dijkstra AA, Ingrassia A, de Menezes RX, et al. Evidence for Immune Response, Axonal Dysfunction and Reduced Endocytosis in the Substantia Nigra in Early Stage Parkinson's Disease. *PLoS One.* 2015;10(6):e0128651.
9. Zheng B, Liao Z, Locascio JJ, et al. PGC-1alpha, a potential therapeutic target for early intervention in Parkinson's disease. *Sci Transl Med.* 2010;2(52):52ra73.
10. Lesnick TG, Papapetropoulos S, Mash DC, et al. A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. *PLoS Genet.* 2007;3(6):e98.
11. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics.* 2004;20(3):307-315.
12. 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.
13. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics.* 2008;9:559.
14. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS.* 2012;16(5):284-287.

15. Sanz H, Valim C, Vegas E, Oller JM, Reverter F. SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinformatics.* 2018;19(1):432.
16. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics.* 2013;14:7.
17. Gui Y, Liu H, Zhang L, Lv W, Hu X. Altered microRNA profiles in cerebrospinal fluid exosome in Parkinson disease and Alzheimer disease. *Oncotarget.* 2015;6(35):37043-37053.
18. Kong Y, Liang X, Liu L, et al. High Throughput Sequencing Identifies MicroRNAs Mediating alpha-Synuclein Toxicity by Targeting Neuroactive-Ligand Receptor Interaction Pathway in Early Stage of Drosophila Parkinson's Disease Model. *PLoS One.* 2015;10(9):e0137432.
19. Daadi MM. Differentiation of Neural Stem Cells Derived from Induced Pluripotent Stem Cells into Dopaminergic Neurons. *Methods Mol Biol.* 2019;1919:89-96.
20. Mendoza JL, Stafford KL, Stauffer JM. Large-sample confidence intervals for validity and reliability coefficients. *Psychol Methods.* 2000;5(3):356-369.
21. Suwarnalata G, Tan AH, Isa H, et al. Augmentation of Autoantibodies by Helicobacter pylori in Parkinson's Disease Patients May Be Linked to Greater Severity. *PLoS One.* 2016;11(4):e0153725.
22. Sidoryk-Wegrzynowicz M, Lee E, Mingwei N, Aschner M. Disruption of astrocytic glutamine turnover by manganese is mediated by the protein kinase C pathway. *Glia.* 2011;59(11):1732-1743.
23. Bannon MJ, Savonen CL, Jia H, et al. Identification of long noncoding RNAs dysregulated in the midbrain of human cocaine abusers. *J Neurochem.* 2015;135(1):50-59.

# Tables

**Table 1**

Statistical differences in different modules of lncRNAs.

| Module | All | Lnc | DElnc | p.value |
|---|---|---|---|---|
| Black* | 200 | 7 | 7 | 2.60E-08 |
| Blue* | 1242 | 53 | 18 | 8.32E-08 |
| Brown | 1037 | 70 | 4 | 0.853 |
| Green | 316 | 11 | 1 | 0.619 |
| Greenyellow | 114 | 2 | 1 | 0.160 |
| Grey60 | 58 | 2 | 1 | 0.160 |
| Lightcyan | 59 | 1 | 0 | 1 |
| Lightgreen | 52 | 2 | 1 | 0.1608 |
| Lightyellow | 50 | 4 | 0 | 1 |
| Magenta | 147 | 3 | 0 | 1 |
| Midnightblue* | 59 | 12 | 5 | 0.001 |
| Pink | 153 | 3 | 1 | 0.231 |
| Purple | 143 | 4 | 1 | 0.295 |
| Red | 252 | 1 | 0 | 1 |
| Tan* | 108 | 2 | 2 | 0.007 |
| Turquoise* | 2449 | 104 | 37 | 5.59E-16 |
| Yellow | 470 | 6 | 0 | 1 |

*p.value calculation was measured by Fisher’s test.

**Table 2**

The number of differential lncRNAs in each module where *P* values represent the degree of significant aggregation of the differential lncRNAs in this module.

| Module | All | PCG | Disgene | p.value |
|---|---|---|---|---|
| Black* | 200 | 193 | 24 | 0.002 |
| Blue* | 1242 | 1189 | 84 | 0.0001 |
| Brown | 1037 | 967 | 41 | 0.803 |
| Cyan | 88 | 88 | 9 | 0.024 |
| Green | 316 | 305 | 14 | 0.595 |
| Greenyellow | 114 | 112 | 10 | 0.041 |
| Grey60 | 58 | 56 | 4 | 0.276 |
| Lightcyan | 59 | 58 | 5 | 0.141 |
| Lightgreen | 52 | 50 | 5 | 0.088 |
| Lightyellow | 50 | 46 | 1 | 0.894 |
| Magenta | 147 | 144 | 16 | 1 |
| Midnightblue* | 59 | 47 | 0 | 0.001 |
| Pink | 153 | 150 | 16 | 1.695e-05 |
| Purple | 143 | 138 | 14 | 8.84e-08 |
| Red | 252 | 251 | 28 | 2.707e-05 |
| Royalblue | 37 | 37 | 7 | 0.241 |
| Salmon | 94 | 94 | 4 | 0.6606 |
| Tan* | 108 | 106 | 7 | 0.0016 |
| Turquoise* | 2449 | 2345 | 166 | 0.006 |
| Yellow | 470 | 464 | 20 | 0.709 |

* indicates p.value < 0.05, there are significant statistical differences

**Table 3**

Candidate lncRNAs which using SVM to analyze.

| ID | best_gm | best_cost | accurary | sensitivity | specificity | AUC |
|---|---|---|---|---|---|---|
| LOC147727 | 1000 | 1 | 0.943 | 1 | 0.84 | 0.6117 |
| RP11-342C20.3 | 1000 | 2 | 0.647 | 1 | 0 | 0.6061 |
| RP4-751H13.6 | 1000 | 2 | 0.929 | 0.978 | 0.84 | 0.6583 |
| RP11-417J1.4 | 1 | 1 | 0.746 | 0.934 | 0.4 | 0.627 |
| AC120114.3 | 10000 | 2 | 0.647 | 1 | 0 | 0.673 |
| NCRNA00107 | 1000 | 1 | 0.788 | 0.9787 | 0.44 | 0.6443 |
| LOC153684 | 1000 | 1 | 0.859 | 1 | 0.6 | 0.6409 |
| AL360001.1 | 100 | 2 | 0.788 | 0.956 | 0.48 | 0.6296 |
| COPG2IT1 | 100 | 2 | 0.830 | 0.978 | 0.56 | 0.6165 |
| AC093323.3 | 100 | 1 | 0.802 | 0.978 | 0.48 | 0.6826 |
| RP11-10O22.2 | 1000 | 2 | 0.887 | 0.978 | 0.72 | 0.6504 |
| RP11-29H23.1 | 1 | 1 | 0.746 | 0.913 | 0.44 | 0.6191 |

# Figures

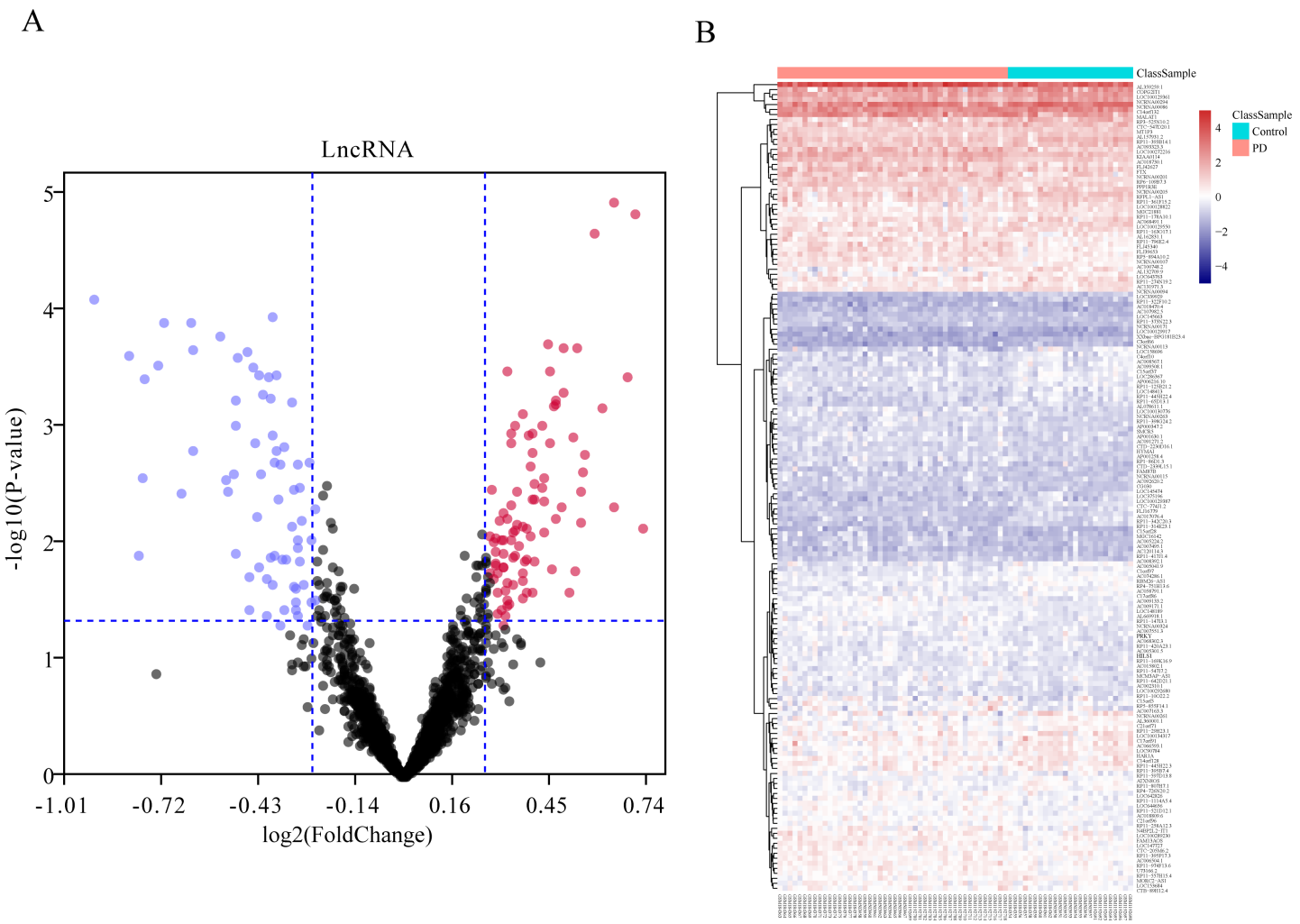
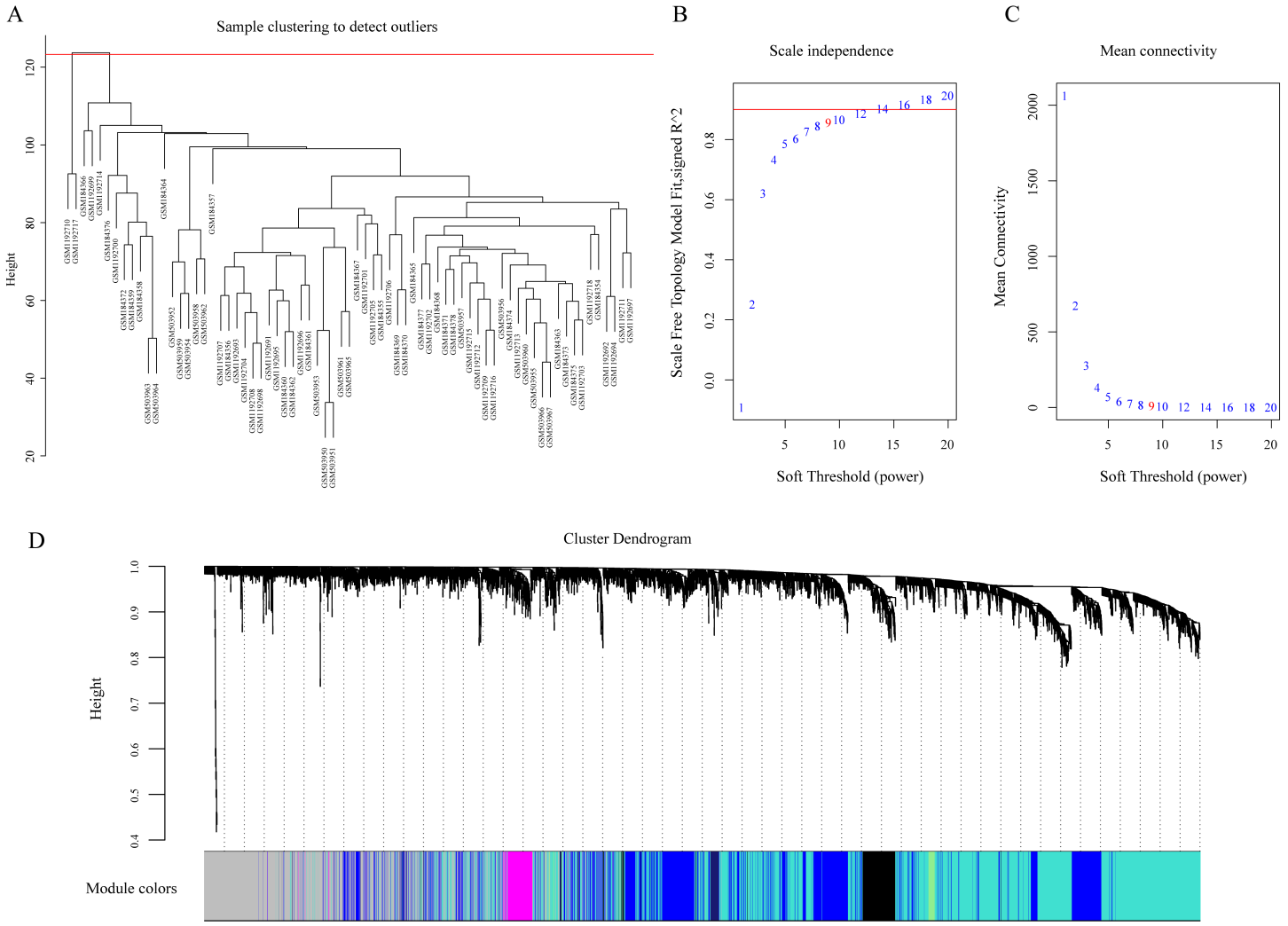


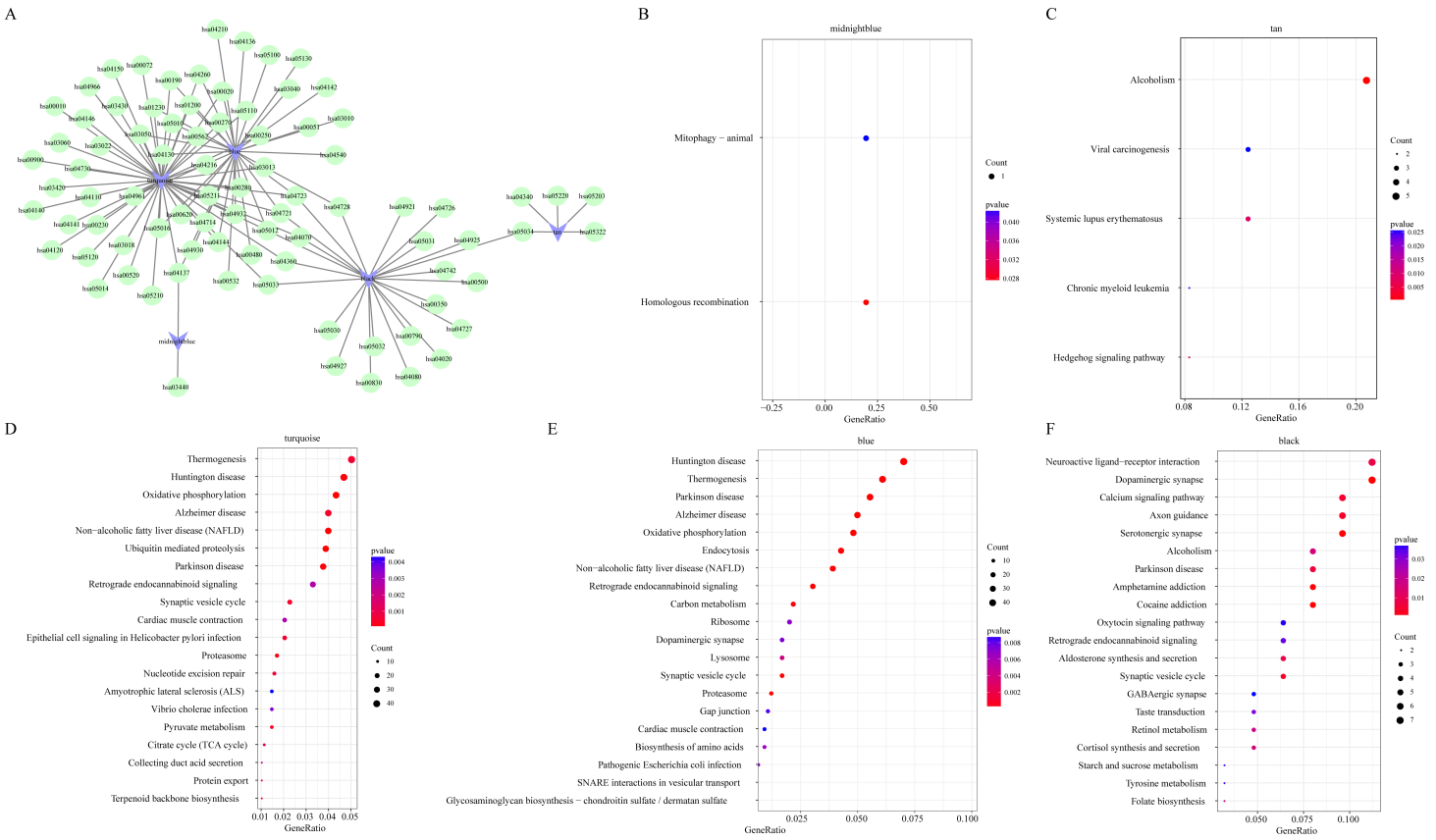

Figure 1

A) In Volcano Plot, the blue part represents downregulated expression and the red part represents upregulated expression. B) Heat map, on the top part, red represents PD samples, blue part which below red represents normal control samples.




## Figure 2

A: Cluster analysis of different samples; B and C diagrams are the network topology analysis of various soft threshold powers; D The picture shows the gene tree and module color, we used WGCNA software to execute and finally identified 21 co-expression modules.




## Figure 3

A) The network relationship between the five-module enrichment results. B-F) The result of top 20 genes in each module.

A

| Trainingdataset | RealPD | RealNormal | |
|---|---|---|---|
| PredictPD | 23 | 2 | |
| PredictNormal | 0 | 10 | Totals |
| Totals | 23 | 12 | 35 |
| Correct | 23 | 10 | 33 |
| Sensitivity(%) | 100 | | |
| Specificity(%) | | 83.33 | |

B

| Validationdataset | RealPD | RealNormal | |
|---|---|---|---|
| PredictPD | 20 | 3 | |
| PredictNormal | 3 | 10 | Totals |
| Totals | 23 | 13 | 36 |
| Correct | 20 | 10 | 30 |
| Sensitivity(%) | 86.95 | | |
| Specificity(%) | | 76.92 | |

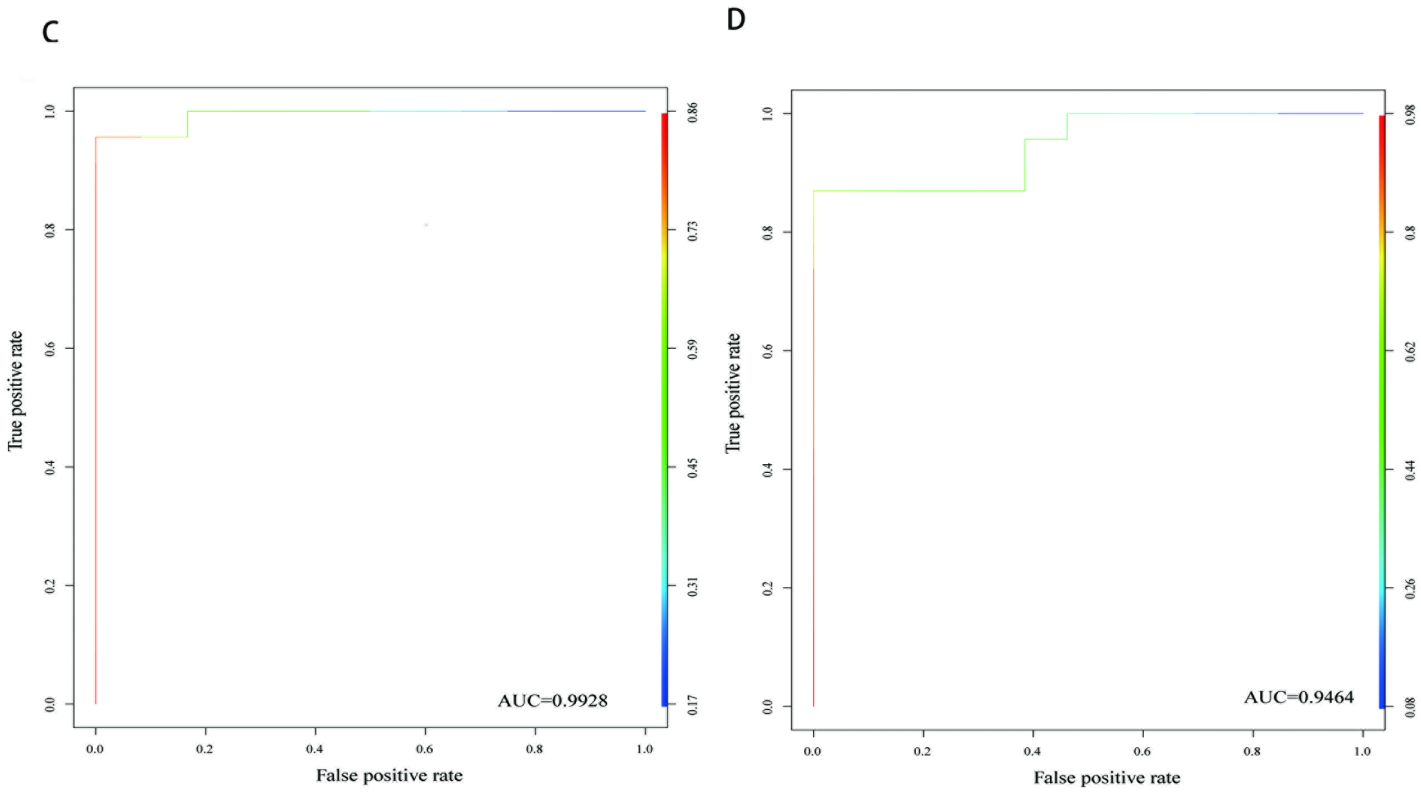



**Figure 4**

Classification in training model and the ROC curve of the model. A) Training data set classification. B) Validate the classification data set. C) ROC curve of training data set. D) Validate the ROC curve of data set.

# Supplementary Files

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

- SupplementaryFig.S4.pdf

- SupplementaryMaterial.xlsx
- SupplementaryFig.S1.tif
- SupplementaryFig.S2.tif
- SupplementaryFig.S3.tif