**基于铁死亡相关基因探究糖尿病视网膜病变诊断标志物及其转录调控机制**

# 分析流程图

**2 、数据和方法**

* 1. **数据简介**

首先从Gene Expression Omnibus (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi>)搜索并从中下载相应的数据。

**2.2、差异表达分析**

利用edgeR[1]包进行差异表达分析，然后根据相应的参数分别筛选出（DR vs Ctrl; DM vs Ctrl）组间共有的差异表达基因。

**2.3、基因共表达分析**

WGCNA（weighted gene co-expression network analysis，权重基因共表达网络分析）是一种分析多个样本基因表达模式的分析方法，可将表达模式相似的基因进行聚类，并分析模块与特定性状或表型之间的关联关系。这里首先选取表达值差异较大的基因，然后利用R包WGCNA[2]进行共表达分析，并挑选表型相关性最高的模块。然后根据选定模块里基因共表达网络中基因间的相关性筛选出hub基因。

**2.4、基因功能富集分析**

基于Gene Ontology[3]数据库，以及KEGG PATHWAY DATABASE[4]生物化学通路的数据库，对候选基因进行了功能富集分析。利用统计学算法（Fisher’s exact test）来找出一组基因和哪些具体的功能条目联系最大，分析结果中每个条目都对应一个统计值P-value来表示显著性。

**2.5、ssGSEA分析**

这里根据各基因表达量及相应基因集，利用GSVA[5]包中的ssgsea法进行富集分析。最终得到各样本对应各基因集的ssGSEA得分

**2.6、Lasso COX回归模型构建**

利用R包glmnet [6]构建lasso回归模型（COX）。并根据构建得到的模型，预测各样本对应的Risk值。

## 2.7、ROC分析

利用pROC包进行ROC分析。

**2.8、miRNA预测**

根据miRNA与mRNA结合数据库（miRMap[7]、miRanda[8]、miRDB[9]、TargetScan[10]及miTarBase[11]）找出候选基因对应的miRNA。

**2.9、转录因子预测**

根据数据库（ChEA、ENCODE、hTFtarget、TRANSFAC和TRRUST）中基因与转录因子（TF）的关系找出候选基因对应的转录因子。

**2.10、lncRNA预测**

根据ENCORI（https://starbase.sysu.edu.cn/）结合数据库找出候选基因对应的lncRNA。

**3、分析结果**

**3.1、探究疾病组特异表达的基因及相关功能**

**3.1.1、差异表达分析**

根据GSE160306中的测序数据及样本分组信息,利用edgeR包进行差异表达分析，然后根据合适的参数筛选出 （DR vs Ctrl; DM vs Ctrl）组间共有的差异表达基因。

差异基因统计总览表

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **数据信息** | | **差异阈值** | | **差异基因数** | | **样本数** | | | |
| **数据来源** | **平台** | **FC** | **q\_value** | **Up** | **Down** | **DR** | **Ctrl** | **DM** | **Ctrl** |
| GSE160306 | 测序 | 1.2 | 0.05 | 986 | 640 | 39 | 20 |  |  |
| GSE160306 | 测序 | 1202 | 782 |  |  | 20 | 20 |

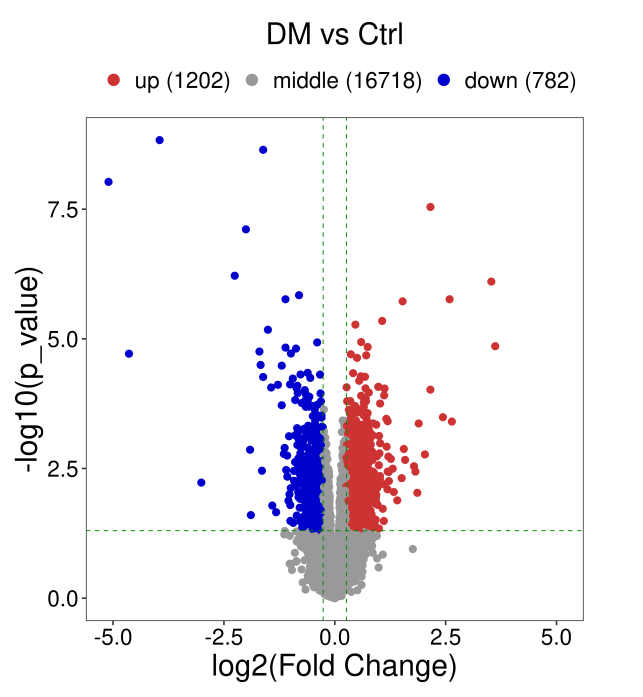
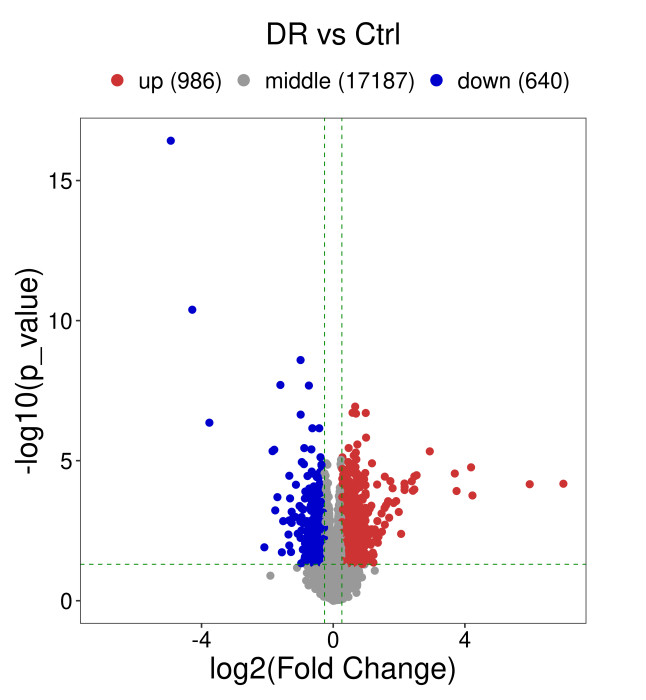
结果文件夹：diff\_exp

DR vs Ctrl 部分差异表达结果表

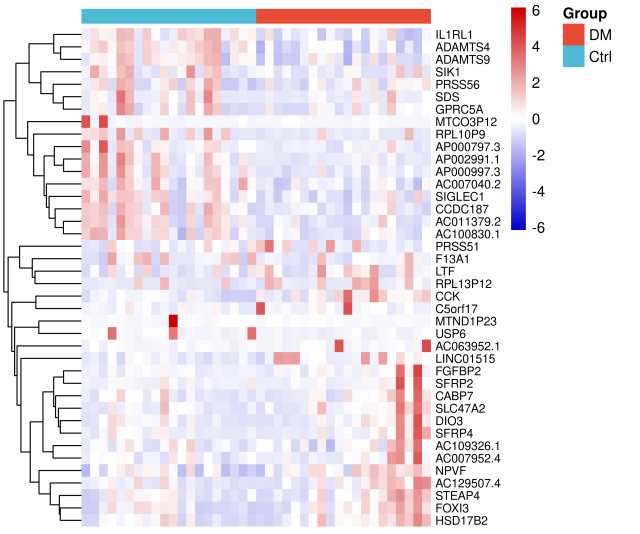
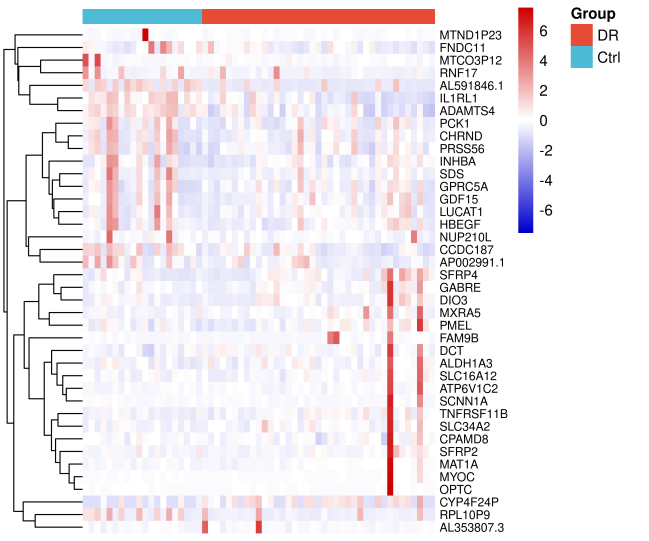
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Gene\_Name** | **log2FC** | **FC** | **p\_value** | **q\_value** | **DR** | **Ctrl** |
| OPTC | 6.99232 | 127.32059 | 0.00007 | 0.01333 | 2.47329 | -4.51904 |
| MYOC | 5.96930 | 62.65249 | 0.00007 | 0.01366 | 1.48193 | -4.48737 |
| FAM9B | 4.22999 | 18.76526 | 0.00017 | 0.02083 | 1.04024 | -3.18975 |
| SFRP2 | 4.18751 | 18.22079 | 0.00002 | 0.00775 | 2.92603 | -1.26149 |
| SCNN1A | 3.74736 | 13.42977 | 0.00012 | 0.01706 | 0.60516 | -3.14220 |

DM vs Ctrl 部分差异表达结果表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Gene\_Name** | **log2FC** | **FC** | **p\_value** | **q\_value** | **DM** | **Ctrl** |
| SFRP2 | 3.61758 | 12.27438 | 0.00001 | 0.01441 | 2.33288 | -1.28470 |
| AC063952.1 | 3.53016 | 11.55273 | 0.00000 | 0.00211 | 0.79851 | -2.73165 |
| SFRP4 | 2.64031 | 6.23467 | 0.00039 | 0.06647 | 1.50571 | -1.13460 |
| FGFBP2 | 2.58898 | 6.01672 | 0.00000 | 0.00322 | 1.62015 | -0.96883 |
| C5orf17 | 2.43828 | 5.41996 | 0.00032 | 0.06135 | 0.32102 | -2.11726 |



差异表达火山图

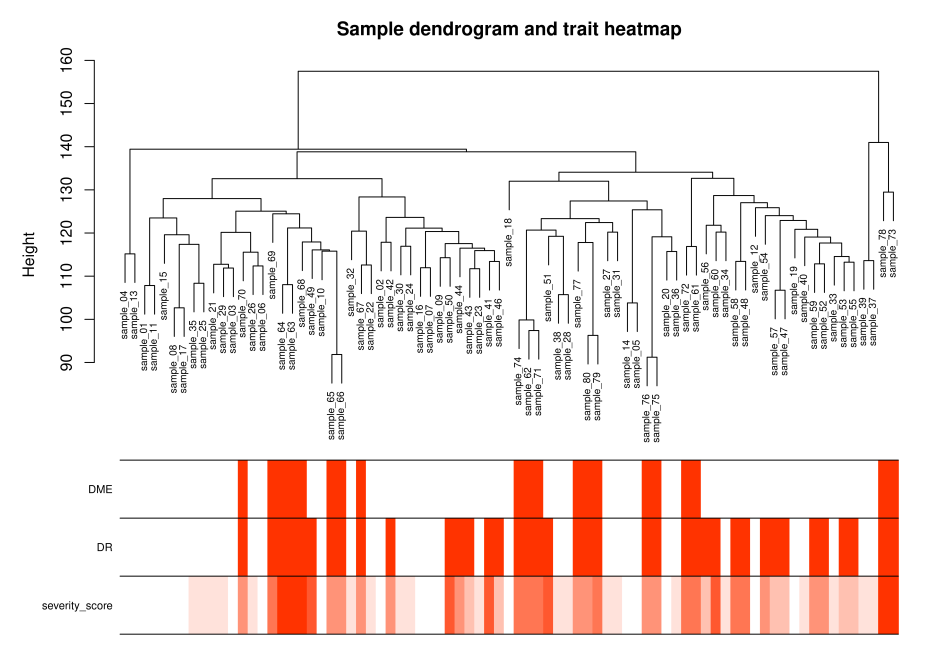


上下调各top20基因表达热图

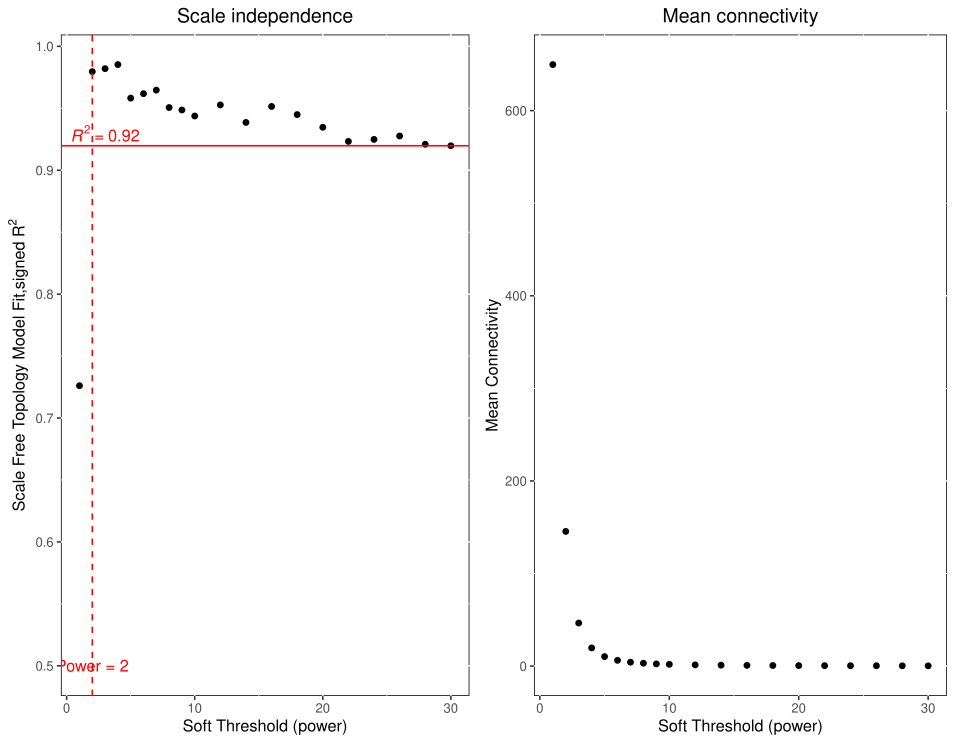
**3.1.2、WGCNA共表达分析**

结果文件夹：WGCNA

数据来源：GEO：GSE160306

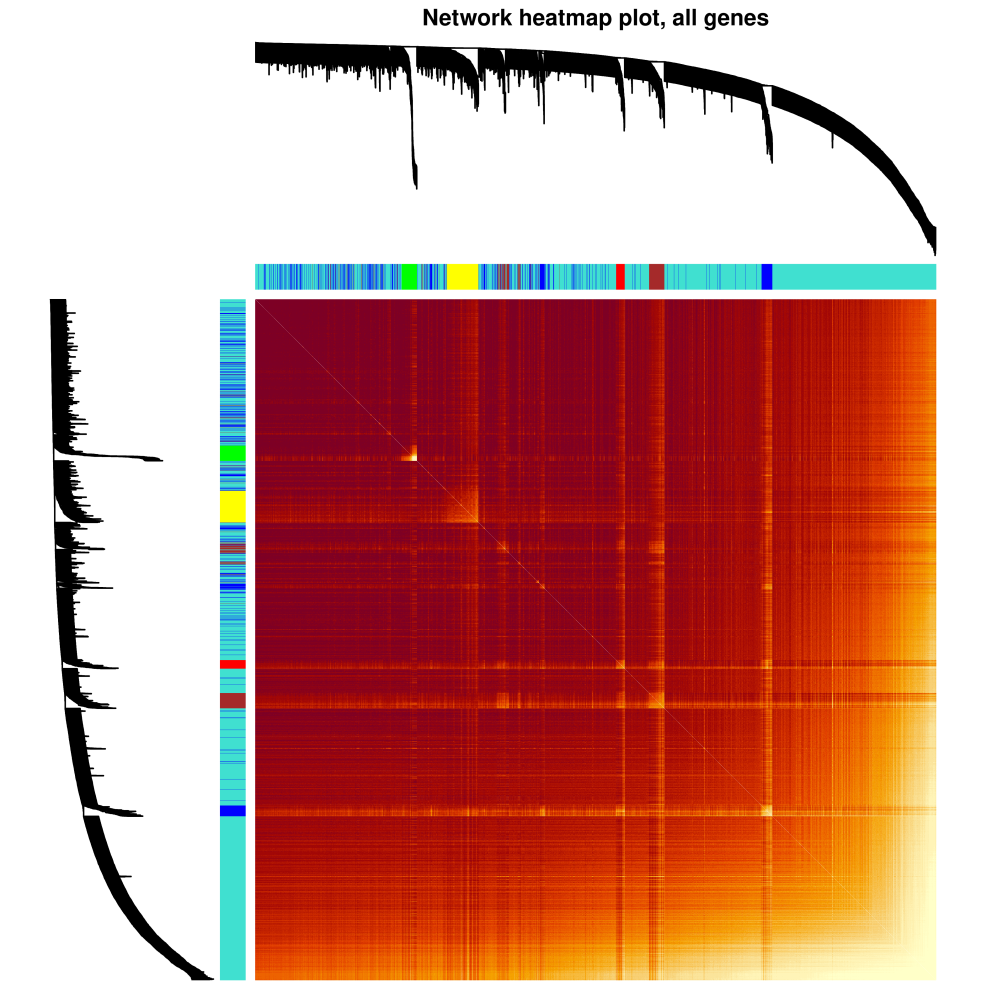


样本聚类及对应分组图



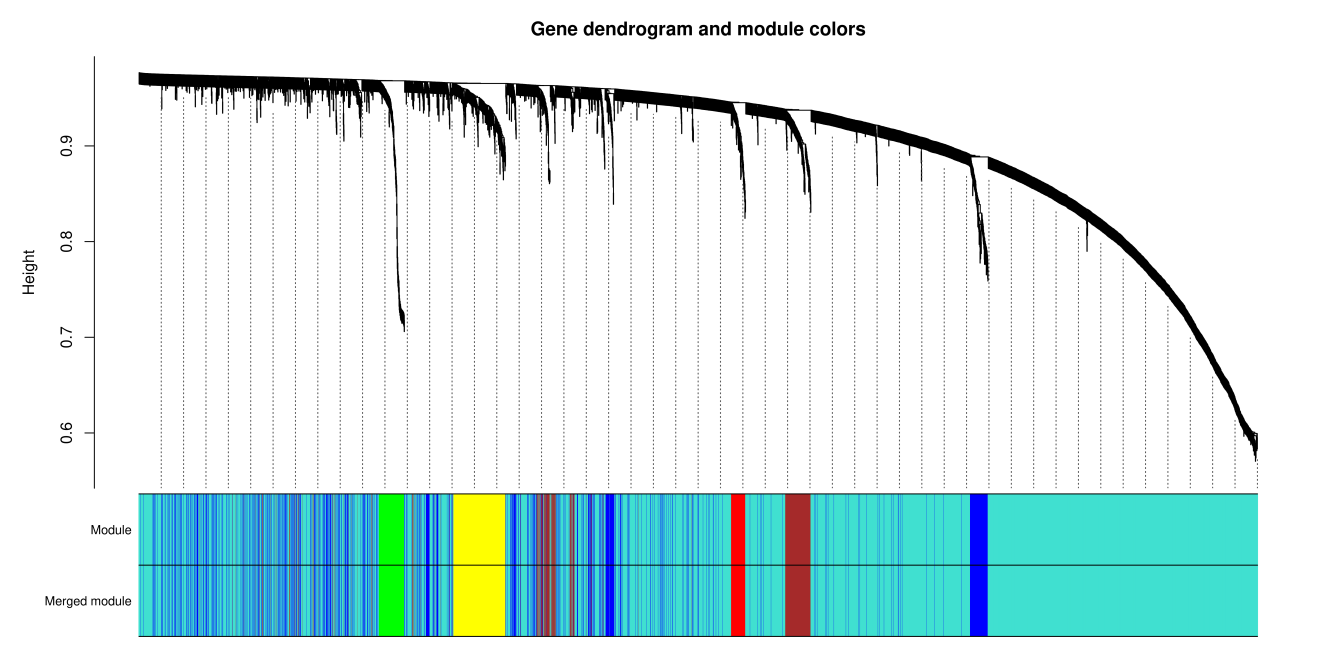
Power值确定图

R2阈值为0.8，这里R2都小于0.8，因此选R2最大Power结果聚类。

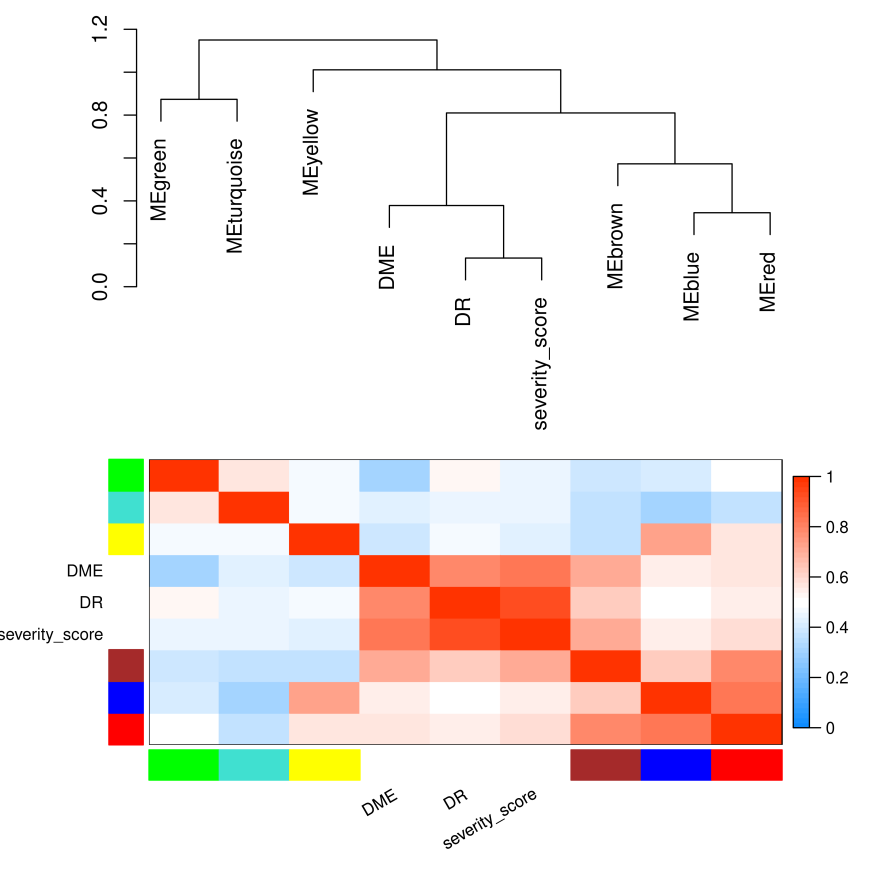


基因间相关性基因聚类热图

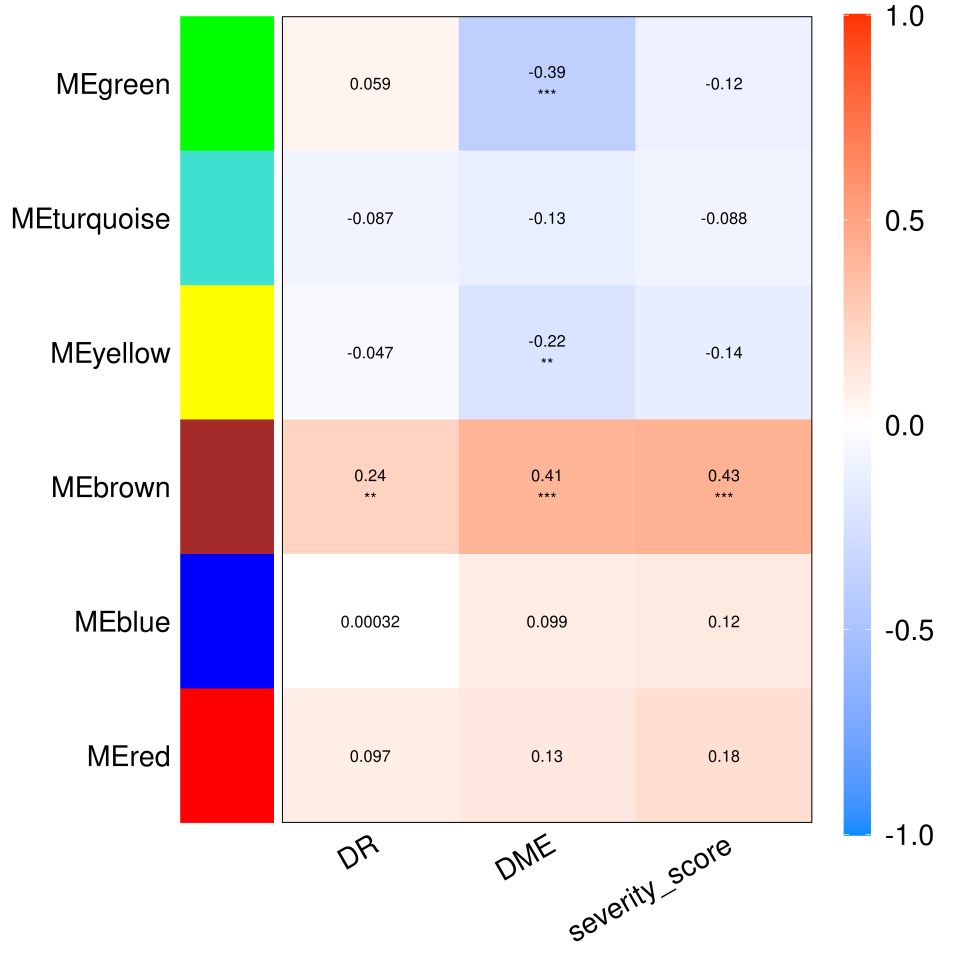
颜色越亮相关性越好



基因聚类及对应共表达模块图



共表达模块聚类及模块间相关性热图



各共表达模块与性状相关性热图

括号内是p值，括号上边是相关系数

**3.1.3、功能富集分析**

这里根据差异表达基因筛选到的（DR vs Ctrl; DM vs Ctrl）组间差异表达基因进行GO、KEGG富集分析。并且根据共表达中相关性最好的模块（brown）中的基

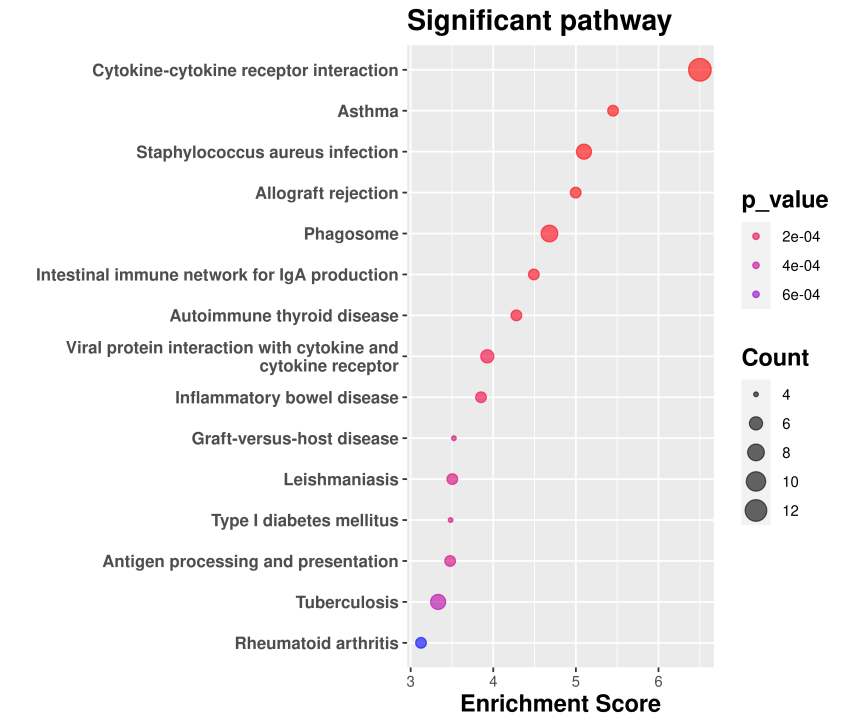
因进行GO和KEGG富集分析。基因功能（GO）分3类：Biological Process (BP)、Molecular Function (MF)、 Cellular Component (CC)。

结果文件夹：functions

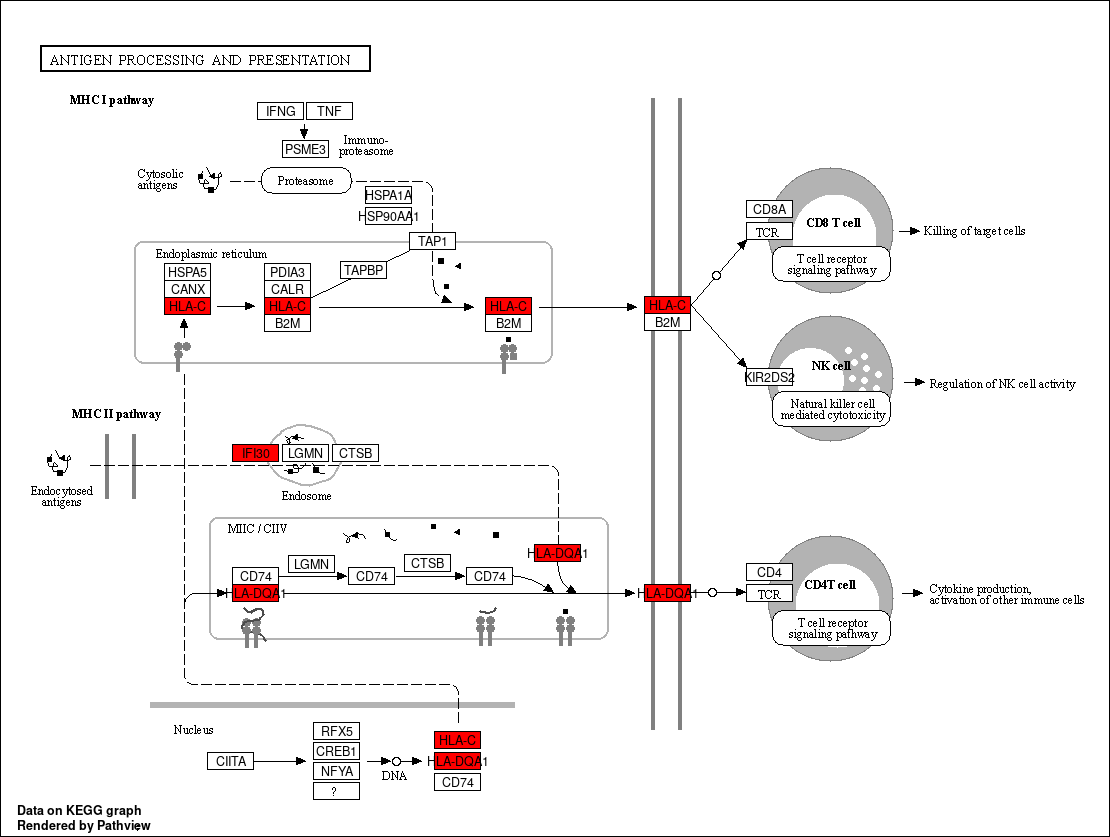
注：这里只以共表达中相关性最好的模块（brown）中的基因KEGG图例富集结果进行展示，其他结果见相应文件。

KEGG富集部分结果表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | **Term** | **p\_value** | **fdr** | **Enrichment\_Score** |
| hsa04060 | Cytokine-cytokine\_receptor\_interaction | 3.15E-07 | 3.85E-05 | 6.501264959 |
| hsa05310 | Asthma | 3.55E-06 | 0.000216413 | 5.450046496 |
| hsa05150 | Staphylococcus\_aureus\_infection | 7.99E-06 | 0.000306542 | 5.097562105 |
| hsa05330 | Allograft\_rejection | 1.01E-05 | 0.000306542 | 4.997810368 |
| hsa04145 | Phagosome | 2.09E-05 | 0.000509596 | 4.68016383 |



富集结果图气泡图



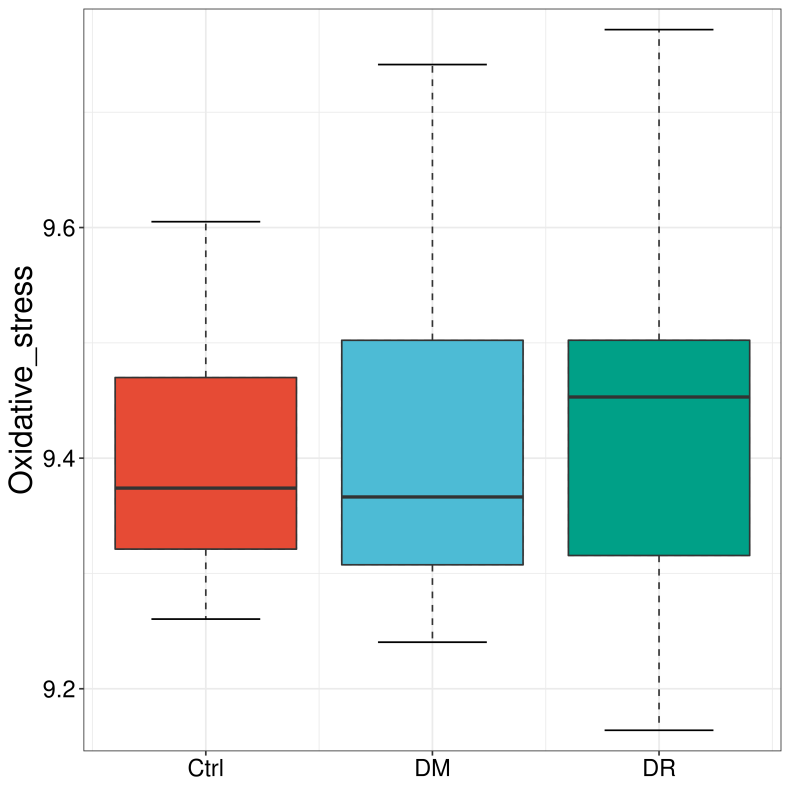
基因在KEGG通路中的位置图（在EMT vs CTR中红色是上调基因，绿色下调）

**3.2、探究氧化应激表达异质性并挖掘疾病诊断标志物**

**3.2.1、ssGSEA分析**

为了进一步说明疾病发生是和氧化应激是相关的,基于氧化应激相关基因，计算 GSE160306 数据集中每个样本的氧化应激活性评分，再展示氧化应激评分在组间差异。利用GSVA包中的ssgsea法对每个样本进行打分。然后用wilcox.test计算组间得分差异。

结果文件夹：ssGSEA

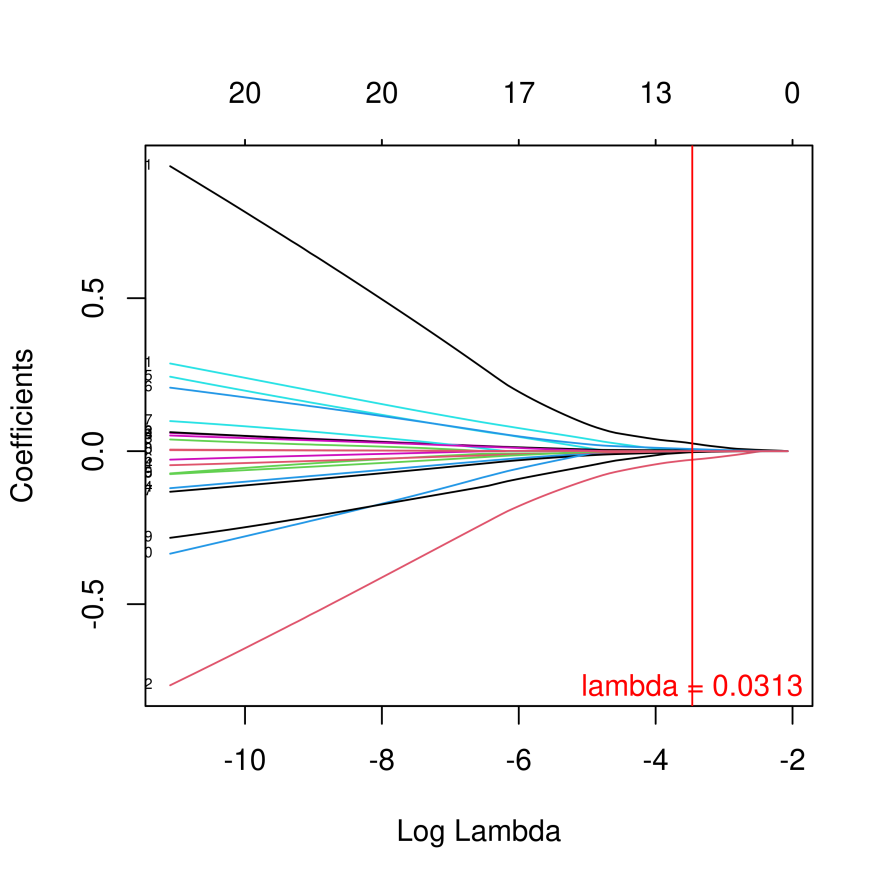


组间得分差异结果图

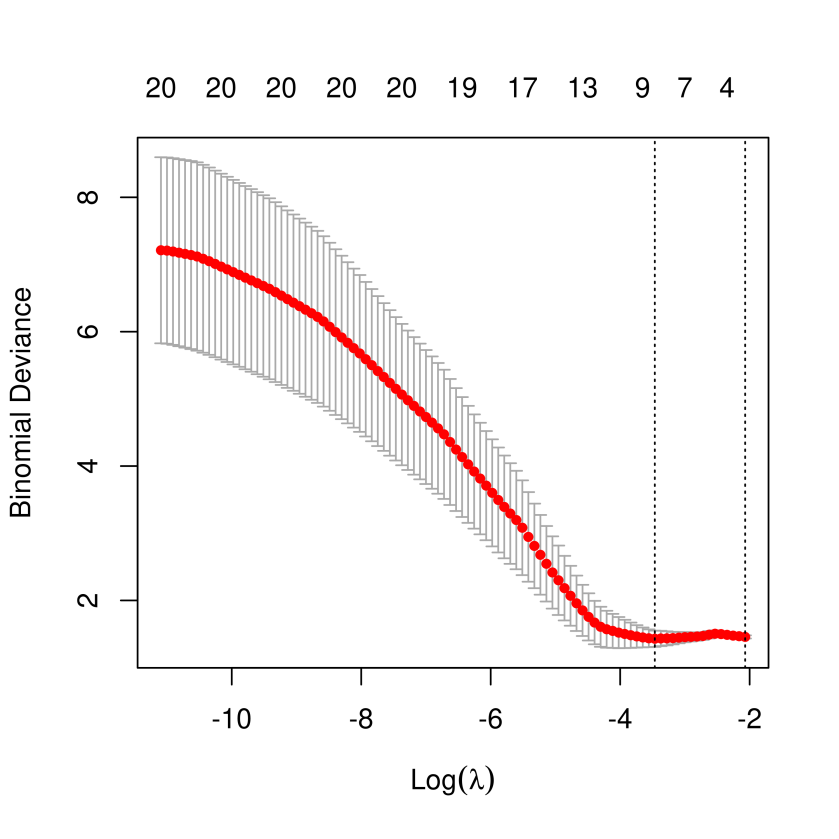
## 3.2.2、Lasso回归模型

基于 GSE160306 数据集差异分析结果筛选出差异表达氧化应激相关基因，再通过 lasso 等算法筛选出诊断标志物。利用glmnet构建诊断模型。

## 结果文件夹：lasso

****

LASSO回归结果图

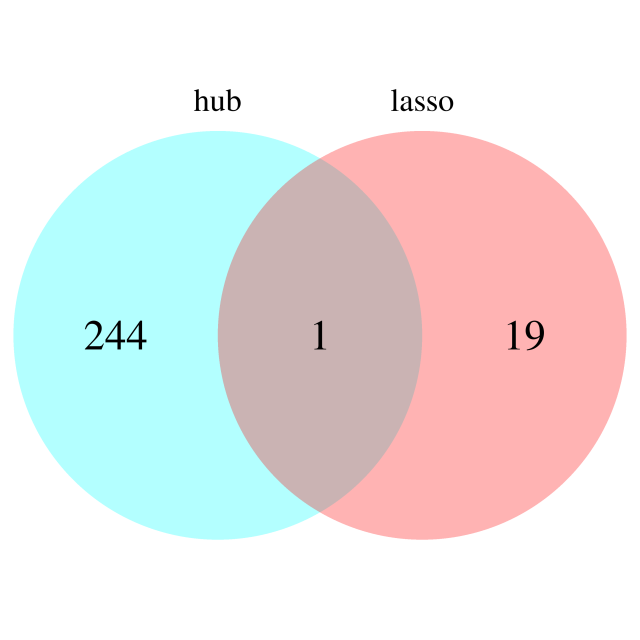
****

LASSO回归结果基因确定图

**3.2.3、差异交集基因**

基于 WGCNA 分析筛选到的hub基因与筛选到的诊断标志物取交集即为目标基因(LTF)。

结果文件夹：venn

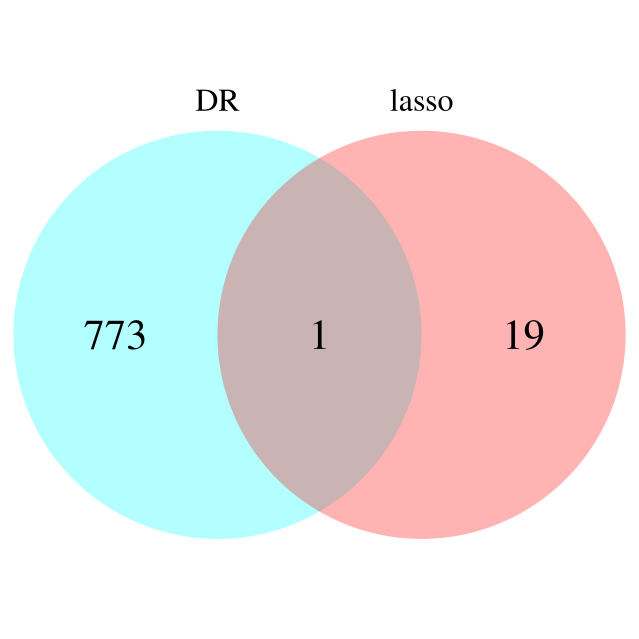


交集基因韦恩图

**3.3、探究目标基因调控机制及作为诊断标志物性能**

**3.3.1、验证目标基因差异表达**

基于 GSE185011 数据集展示筛选到的诊断标志物在组间表达差异，交集也为目标基因(DYSF)。



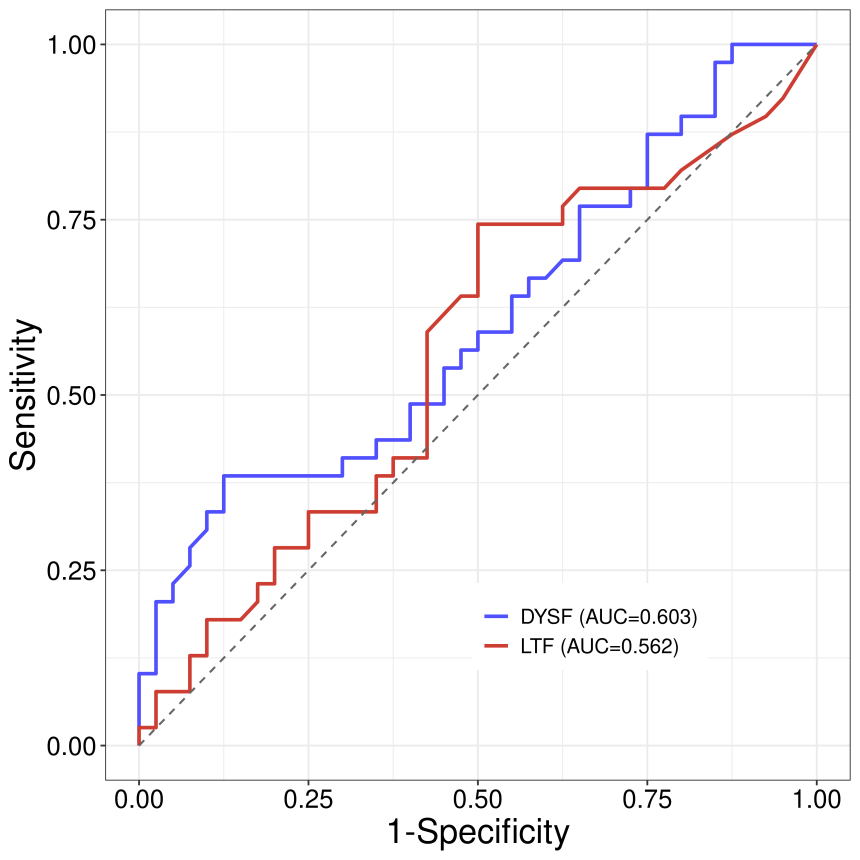
交集基因韦恩图

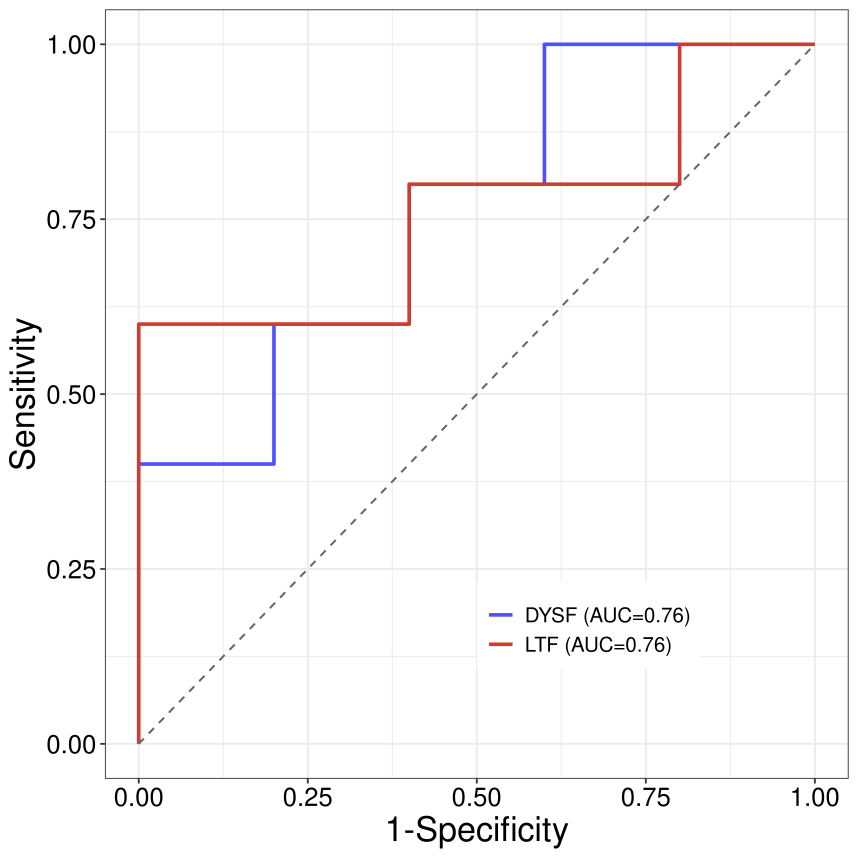
**3.3.2、ROC 分析**

本研究根据上述得到的两个差异基因交集结果，结合各样本的分组信息。利用pROC包分别进行ROC分析。

结果文件夹：ROC

部分ROC结果表





ROC结果图（上：GSE160306，下：GSE185011）

**3.3.3、miRNA预测**

这里根据miRNA与mRNA结合数据库（miRDB、TargetScan、miRanda、miTarBase）找出目标基因对应的miRNA。然后筛选至少4个数据库中同时存在的关系对，作为最终的miRNA-Gene关系对。

结果文件夹：miRNA

目标基因miRNA预测部分结果表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **miRNA** | **Target\_Gene** | **miRDB** | **TargetScan** | **miRanda** | **miRMap** | **miTarBase** | **Count** |
| hsa-miR-214-3p | LTF | 1 | 1 | 1 | 1 | 1 | 5 |
| hsa-miR-1275 | DYSF | 1 | 1 | 1 | 1 | 0 | 4 |
| hsa-miR-196b-3p | LTF | 1 | 1 | 1 | 1 | 0 | 4 |
| hsa-miR-654-3p | DYSF | 1 | 1 | 1 | 1 | 0 | 4 |
| hsa-miR-761 | LTF | 1 | 1 | 1 | 1 | 0 | 4 |

**3.3.4、TF预测**

这里根据数据库（ChEA、ENCODE、hTFtarget、TRANSFAC和TRRUST）中基因与转录因子（TF）的关系找出目标基因对应的转录因子（至少3个数据库中同时存在的关系对）。

结果文件夹：TF

目标基因TF预测部分结果表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **TF** | **Target** | **ChEA** | **ENCODE** | **hTFtarget** | **TRANSFAC** | **TRRUST** | **Count** |
| JUN | DYSF | 0 | 1 | 1 | 1 | 0 | 3 |
| EP300 | DYSF | 1 | 1 | 1 | 0 | 0 | 3 |
| GATA1 | DYSF | 1 | 1 | 1 | 0 | 0 | 3 |
| GATA3 | DYSF | 1 | 1 | 1 | 0 | 0 | 3 |

**3.3.5、lncRNA预测**

这里根据ENCORI（https://starbase.sysu.edu.cn/）结合数据库找出候选基因对应的lncRNA。

结果文件夹：lncRNA

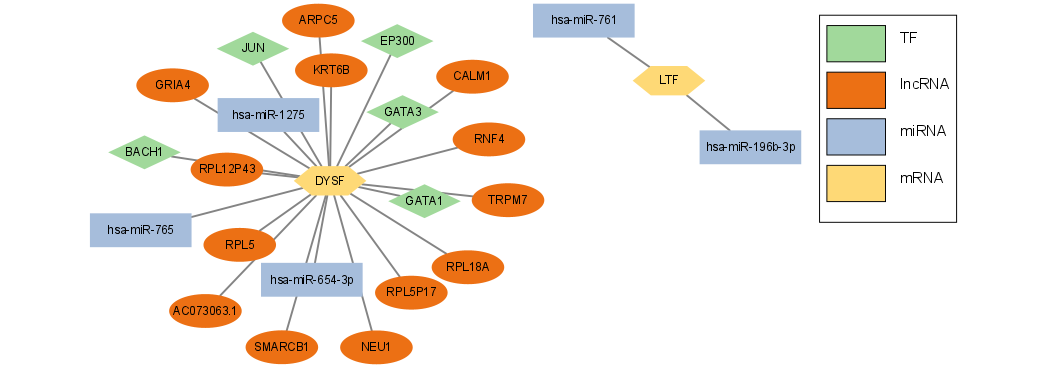
目标基因lncRNA预测部分结果表

|  |  |  |  |
| --- | --- | --- | --- |
| **geneID** | **geneName** | **pairGeneID** | **pairGeneName** |
| ENSG00000135636 | DYSF | ENSG00000092439 | TRPM7 |
| ENSG00000135636 | DYSF | ENSG00000099956 | SMARCB1 |
| ENSG00000135636 | DYSF | ENSG00000063978 | RNF4 |
| ENSG00000135636 | DYSF | ENSG00000105640 | RPL18A |
| ENSG00000135636 | DYSF | ENSG00000122406 | RPL5 |
| ENSG00000135636 | DYSF | ENSG00000185479 | KRT6B |

**3.3.6、调控网络**

这里根据上述找到的目标基因对应的TF,miRNA和lncRNA，用cytoscape构建网络图。

结果文件夹：network



调控网络图

**参考文献**

1. ROBINSON M D, MCCARTHY D J, SMYTH G K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data [J]. Bioinformatics (Oxford, England), 2010, 26(1): 139-40.
2. Langfelder, P., Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008) doi:10.1186/1471-2105-9-559。
3. Ashburner, M., et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000. 25(1): p. 25-9.
4. Kanehisa, M. and S. Goto, KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 2000. 28(1): p. 27-30.
5. Friedman J, Hastie T, Tibshirani R (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1–22. <http://www.jstatsoft.org/v33/i01/.>
6. VEJNAR C E, BLUM M, ZDOBNOV E M. miRmap web: Comprehensive microRNA target prediction online [J]. Nucleic acids research, 2013, 41(Web Server issue): W165-W8.
7. MIRANDA K C, HUYNH T, TAY Y, et al. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes [J]. Cell, 2006, 126(6): 1203-17.
8. WONG N, WANG X. miRDB: an online resource for microRNA target prediction and functional annotations [J]. Nucleic acids research, 2015, 43(Database issue): D146-D52.

[10] NIELSEN C B, SHOMRON N, SANDBERG R, et al. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs [J]. RNA, 2007, 13(11): 1894-910.

[11] HUANG H-Y, LIN Y-C-D, LI J, et al. miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database [J]. Nucleic Acids Research, 2019, 48(D1): D148-D54.