

利用加权基因共表达网络分析方法识别帕金森病进展过程中的关键基因

王树文,樊冬梅,陈方正,谢玉平,吕晓彤,李依璘,林鸿程* (广东医科大学,广东东莞 523808)

摘要: **目的** 利用加权基因共表达网络分析(WGCNA)方法识别帕金森病(PD)相关的关键基因,为研究PD发病机制提供研究靶点。**方法** GEO数据库下载GSE22491表达谱和样本信息数据,基于WGCNA算法,构建基因共表达模块,并计算基因模块与样本信息的相关性,选取与样本信息相关性最强的基因模块进行基因功能(GO)与KEGG富集分析,同时根据模块内基因GS值和MM值及模块内连通性确立PD相关关键基因,并对其进行初步验证。**结果** 通过WGCNA分析构建出了15个共表达基因模块,其中模块Turquoise(MEturquoise)与PD具有强的相关性。对其模块内基因进行筛选,识别出了IFITM5、MYPOP、SIK1、NPAS3、SMARCC2、FBXO17、SLC34A3、AMN、TNRC18、FBRSL1 10个PD关键基因。其中IFITM5和FBRSL1在Genecards数据库检索参与了PD进展过程。**结论** IFITM5、MYPOP、SIK1、NPAS3、SMARCC2、FBXO17、SLC34A3、AMN、TNRC18和FBRSL1基因在PD进展中扮演重要的角色,其具体机制值得进一步研究。

关键词: 帕金森病; 关键基因; 加权基因共表达网络分析

中图分类号: R 749

文献标志码: A

文章编号: 2096-3610(2021)05-0539-06

Identification of key genes in the progression of Parkinson's disease with weighted gene co-expression network analysis

WANG Shu-wen, FAN Dong-mei, CHEN Fang-zheng, XIE Yu-ping, LV Xiao-tong, LI Yi-lin, LIN Hong-cheng* (Guangdong Medical University, Dongguan 523808, China)

Abstract: **Objective** To identify key genes associated with Parkinson's disease (PD) with weighted gene co-expression network analysis (WGCNA) method to provide study target for the study of PD pathogenesis. **Methods** GSE22491 expression profile and sample information data were downloaded from GEO database. A gene co-expression module was constructed based on WGCNA algorithms, and the correlation between gene module and sample information was calculated. Then the gene module the most relevant to the sample information was selected for GO and KEGG enrichment analysis, and the key genes related to PD was determined according to the GS value and MM value of the genes and intra-module connectivity and preliminarily verified. **Results** A total of 15 co-expression gene modules were constructed through WGCNA analysis, of which the Module Turquoise (MEturquoise) has a strong correlation with PD. After screening of genes in the module, 10 key genes related to PD were identified, including IFITM5, MYPOP, SIK1, NPAS3, SMARCC2, FBXO17, SLC34A3, AMN, TNRC18 and FBRSL1, and gene IFITM5 and FBRSL1 were validated involved in PD progress by Genecards database. **Conclusion** IFITM5, MYPOP, SIK1, NPAS3, SMARCC2, FBXO17, SLC34A3, AMN, TNRC18 and FBRSL1 play an important role in the progression of PD, and the specific mechanism deserves further study.

Key words: Parkinson's disease; key genes; weighted gene co-expression network analysis

加权基因共表达网络分析(WGCNA)是探究基因复杂调控网络最常用到的生物信息学分析方法^[1-3]。

帕金森病(PD)作为一种涉及多基因调控的疾病,目前利用WGCNA算法对PD进行研究的文献报道不

基金项目: 广东省医学科研基金项目(No.A2018199),湛江市非资助科技攻关计划项目(No.2017B01123),广东医科大学科研基金项目(No.M2017025),广东省大学生创新创业训练计划项目(No.201910571075, No.201910571090),广东医科大学大学生创新创业训练计划项目(No.GDMU2018075, No.GDMU2018090)

收稿日期: 2021-01-08; **修订日期:** 2021-04-13

作者简介: 王树文(1982-),男,硕士,高级实验师

通信作者: 林鸿程(1980-),男,硕士,高级实验师, E-mail: gdmclhc@139.com

多。本研究通过构建加权基因共表达网络最终识别出PD密切相关的关键基因模块及关键基因,以期为阐明PD发生发展机制及治疗提供新的研究靶点分子。

1 材料和方法

1.1 PD数据集和样本信息的获取

PD相关表达谱芯片数据与样本信息下载于GEO数据库(<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22491>),芯片数据编号:GSE22491,芯片平台:GPL6480(Agilent-014850 Whole Human Genome Microarray 4x44K G4112F),检测细胞类型:peripheral blood mononuclear cells。数据集样本信息主要包括8例正常样本和10例PD样本信息。基于R软件中的“GEOquery”下载数据集矩阵文件和平台注释文件^[4],根据注释文件对下载的矩阵文件进行基因重注释后,过滤掉注释信息缺失的探针和低表达的探针,最后利用R包“limma”中的“normalizeBetweenArrays”函数对芯片表达谱数据进行标准化处理^[5]。另外对矩阵文件中的样本信息单独整理出用于后续WGCNA分析。为保证WGCNA分析过程中运行的顺利性和可靠性,本研究选取了芯片表达谱中方差变化最大的前25%的基因用于下游分析。另外在正式进行WGCNA分析前,还需要通过计算样本间的相关性,绘制样本的分层聚类图,删除离群样本。

1.2 WGCNA分析

WGCNA分析完成主要是基于R软件中的“WGCNA”包实现^[6],具体流程包括:根据表达谱计算出适宜的软阈值,在计算出两两基因之间相互强度的基础上,对其进行加权计算并构建出基于芯片表达谱的邻接矩阵,随后根据TOM重叠计算公式将邻接矩阵转换为拓扑重叠矩阵(TOM),根据聚类树算法和 $1-TOM$ 计算公式,将表达相似基因合并在一起从而构建出多个基因集模块,同一模块内基因往往具有相似和相关的生物学功能。最后计算基因模块的特征值(ME),同时关联样本信息(正常与PD),使用“Pearson”相关系数计算模块特征值与样本性状的相关程度,将与PD相关程度最强的模块定义为PD关键基因模块,随后对该模块内基因进行深入挖掘和功能分析。

1.3 关键基因的筛选

首先计算PD关键基因模块内MM值和GS值,MM值是基因表达水平和ME值之间的相关系数,相关性越高,说明基因和该模块的关联性就越高,GS表

示每个基因与性状的相关程度,参考其他文献我们将PD关键基因模块内基因连接度最大的前10个基因定义为PD关键基因,后续通过文献检索,对PD关键基因进行进一步筛选,识别出目前在PD研究中还未涉及到的关键基因。

1.4 基因功能(GO)和京都基因与基因组百科全书(KEGG)富集分析

利用R软件中的“clusterprofiler”包^[7]对PD关键模块内基因进行GO功能富集分析和KEGG通路富集分析和可视化,分析结果中 $P<0.05$ 表示富集结果的差异具有统计学意义。其富集结果以气泡图形式呈现。

1.5 PD关键基因差异分析与共表达分析

利用R软件中的“limma”包进行筛选出的10个PD关键基因在正常组织样本与PD样本差异性,差异条件设置为:差异倍数绝对值 ≥ 1.2 且校正后的 $P<0.001$ 。最后利用R软件中内置的“cor”函数进行10个PD关键基因的相关性分析,分析方法选用“Pearson”。

1.6 PD相关基因的筛选

基于Genecards数据库(<https://www.genecards.org/>)筛选目前比较明确的PD相关基因,从而对本研究筛选出的10个PD关键基因进行初步验证。

2 结果

2.1 芯片表达谱预处理及加权基因共表达网络的构建

整理GSE22491芯片表达谱,过滤其中探针信息注释不全和重复的基因,最终获得16 576个基因,其中包括16 327个编码基因和249个长非编码RNA,选取方差变化最大的前4 082个基因进行后续的WGCNA分析。通过计算样本间的相关性,绘制样本的分层聚类图,将聚类高度设置为90(红线标记),见图1。

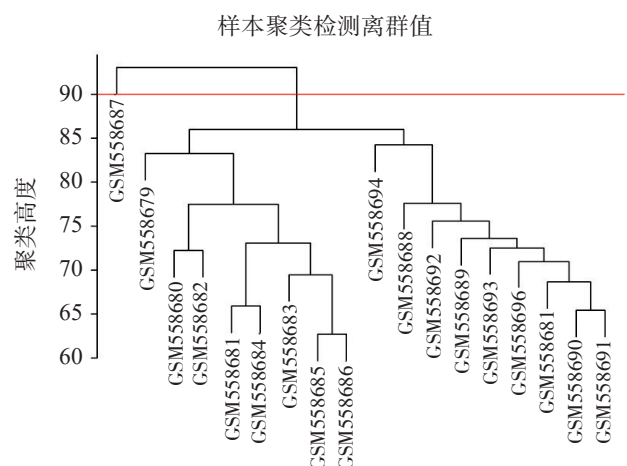


图1 样本聚类并筛选异常样本(红线表示高度=90)

删除异常样本 GSM558687, 剩余 17 例样本信息及相应的 4 082 个基因芯片表达谱数据合并用于后续 WGCNA 构建模块, 见图 2。

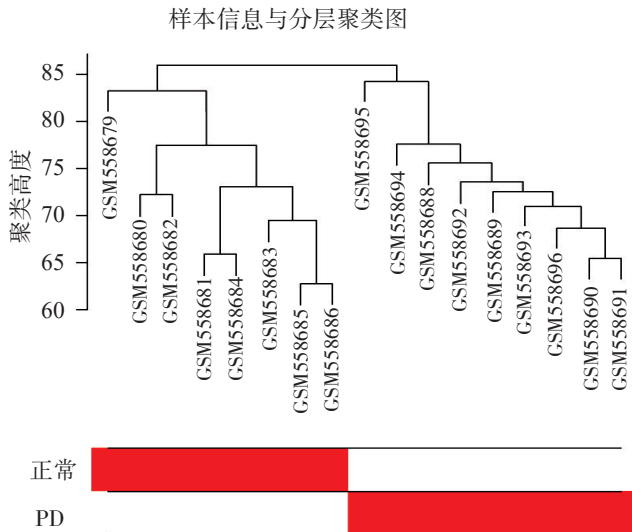


图2 样本聚类与对应的样本信息

2.2 加权基因共表达网络的构建

基于文献中 WGCNA 分析流程, 借助 R 软件中“WGCNA”包实现本研究所需的加权基因共表达网络的构建。在 WGCNA 分析过程中, 软阈值的选取对于构建符合生物学意义的无尺度网络具有决定性的作用, 本研究基于“WGCNA”包中的“pickSoftThreshold”函数进行了软阈值的筛选, 见图 3。当选软阈值即 $\beta=7$ 时, 可以保证本研究构建出的网络既达到无尺度标准, 也使网络的平均连接程度不至于较小而缺乏足够的信息。随后计算基因间的相关性矩阵和邻接矩阵, 根据 TOM 重叠计算公式将邻接矩阵转换为 TOM 重叠矩阵, 根据 $1 - \text{TOM}$ 公式构建基因间分层聚类树, 同时使用动态剪切树的算法把基因分成 15

个模块, 根据模块相似度大于 0.75 标准合并相似模块 (图 4), 最终基于表达谱数据可以识别出 15 个基因模块, 包括: Black 模块(190 个基因)、Blue 模块(576 个基因)、Brown 模块(430 个基因)、cyan 模块(130 个基因)、Green 模块(257 个基因)、Greenyellow 模块(149 个基因)、Grey 模块(109 个)、Magenta 模块(158 基因)、pink 模块(162 个基因)、purple 模块(150 个基因)、Red 模块(230 个基因)、samon 模块(140 个基因)、tan 模块(149 个基因)、Turquoise 模块(937 个基因)和 Yellow 模块(315 个基因)。鉴于 Grey 模块内包含的基因是指不能归类于其他任何模块的基因, 后续研究中一般将 Grey 模块进行舍弃。

2.3 与 PD 样本表型的相关基因模块识别及该模块内基因功能和通路富集分析

基于 Pearson 相关性分析方法, 通过计算各个基因模块与样本表型之间的关系, 绘制出了基因模块与样本表型热图, 可以看出模块 Turquoise(ME-turquoise)与 PD 样本显著相关($r=0.97, P=3e-10$), 见图 5、6。随后对模块 Turquoise 内 937 个基因进行功能注释和通路富集分析, 结果显示 GO 功能注释 BP 条目显著富集在氯离子跨膜转运过程、氧运输、阴离子转运和中性粒细胞生长及代谢等生物学过程, 见图 7。KEGG 通路富集分析显示 937 个基因显著富集在神经活动配体与受体相关作用和细胞因子-细胞因子受体相互作用等通路途径。

2.4 PD 表型相关关键基因识别

基于上述相关性分析识别出与 PD 显著相关基因模块, 参考其他文献我们将 PD 表型最相关的 Turquoise 基因模块内基因基因连接度最大的前 10 个基因定义为 PD 关键基因, 包括 IFITM5、MYPOP、SIK1、NPAS3、SMARCC2、FBXO17、SLC34A3、AMN、TN-

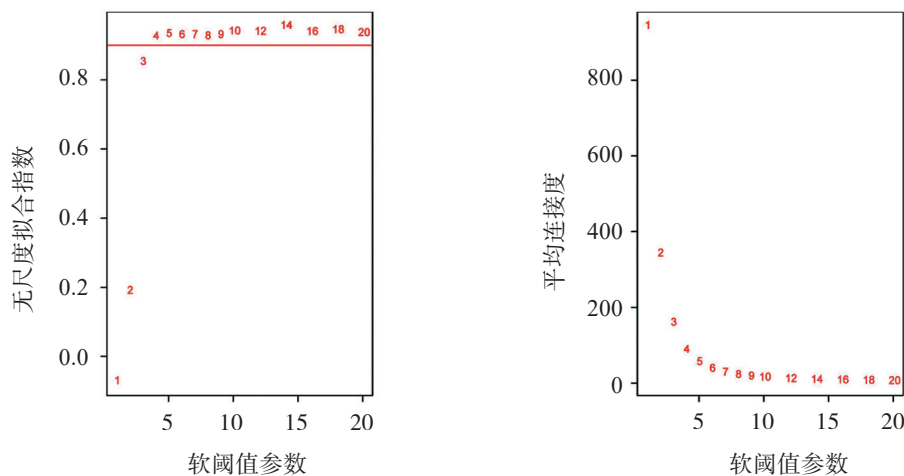


图3 构建 WGCNA 分析过程中软阈值筛选

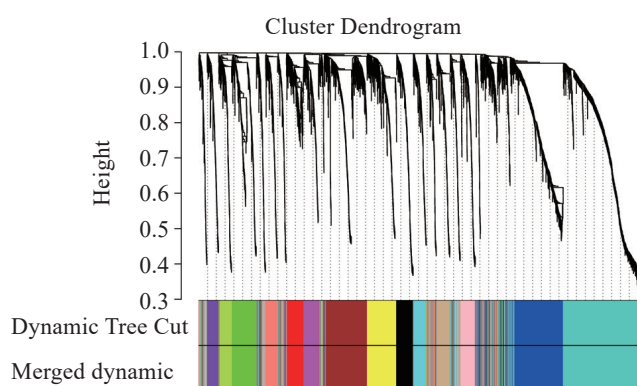


图4 基因共表达网络分层聚类树与共表达模块

RC18 和 FBRSL1, 其中基因 IFITM5 和 FBRSL1 在 Genecards 数据库中已经证实参与 PD 进程, 而剩余 8 个基因与 PD 相关性的研究尚未见文献报道。

2.5 PD 表型相关关键基因差异分析及相关性分析

利用 R 软件中的“limma”包对上述筛选出的 10 个 PD 关键基因在正常组织样本与 PD 样本差异性进行分析, 结果显示, 与正常组织相比, 筛选出的 10 个 PD 关键基因在 PD 样本显著上升(差异倍数大于 1.2 且校正后 $P < 0.001$)。相关性分析显示 10 个 PD 关键基因具有强的相关程度($r > 0.9$), 见图 8。

3 讨论

PD 是 1 种多发于中老年人、以运动障碍为主要表现的神经系统变性疾病^[8]。随着社会老龄化的日益加剧, PD 已成为我国仅次于阿尔茨海默病的神经系统性疾病^[9]。目前, PD 的发病机制仍存在诸多未知

且临床上缺乏有效的治疗药物。因此, 开展 PD 分子机制的相关研究对改善目前 PD 现状具有积极意义。

PD 最主要的病理改变比较明确, 首先由于中脑黑质多巴胺(dopamine, DA)能神经元的变性死亡, 继而引起纹状体 DA 含量显著性减少而导致患者致病^[10]。目前关于 PD 临床特征和发病机制并未明了, 积极探索 PD 发生发展机制对于人们掌握其主要临床特征, 筛选有效的药物治疗靶点具有重要意义。

多基因参与了疾病的发生发展过程, PD 也不例外。近年来出现的高通量测序技术为人们研究多基因在疾病进程中的作用提供了必要条件, 相较于传统的单个分子的生物学研究, 基于高通量的生物信息学研究从系统生物学的角度实现了对疾病相关多个分子靶点及其功能的探究。WGCNA 分析算法是目前进行基因共表达网络研究最常用的生物信息学分析方法, 广泛用于癌症研究领域, 其稳定性和有效性在多项研究中得到证实, 该算法的优势是在计算两两基因相互关系的基础上引入加权值, 从而构建出了更符合生物学意义的基因共表达网络。在本研究中, 我们首先将 WGCNA 算法用于 PD 转录组基因分析, 试图利用先进的生物信息学分析方法, 通过对 PD 高通量芯片数据的挖掘, 为人们深入探究 PD 的发病机制提供新的研究思路和必要的候选靶点分子。

通过构建加权基因共表达网络, 我们识别出 Turquoise 模块内基因与 PD 具有非常强的相关性, 通过对模块内基因进行功能注释和通路富集分析, 结果显示该模块内基因显著富集在氯离子跨膜转运过程、氧运输、阴离子转运和中性粒细胞生长及代谢相关生物

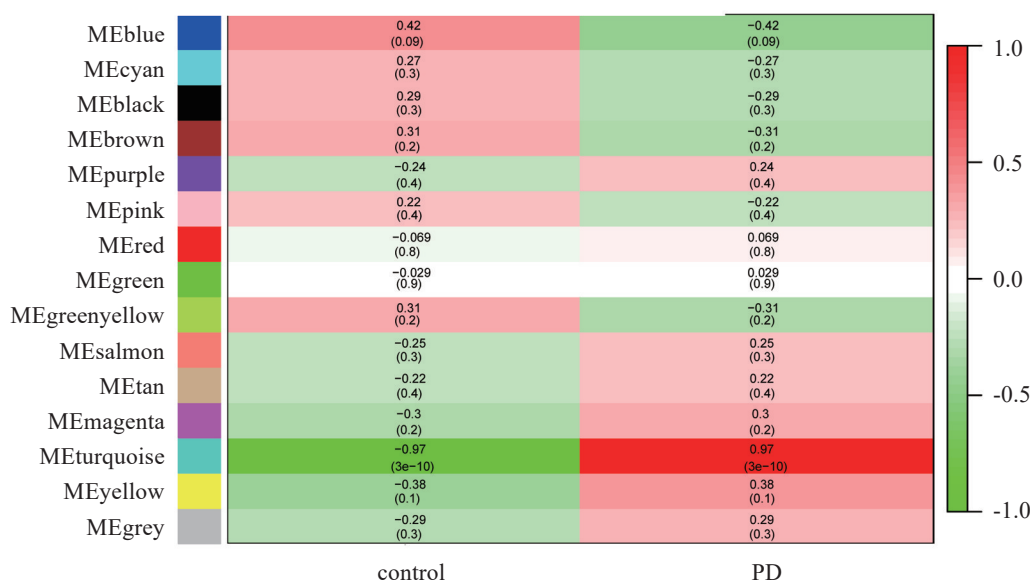


图5 WGCNA 构建模块(ME 值)与样本信息相关热图

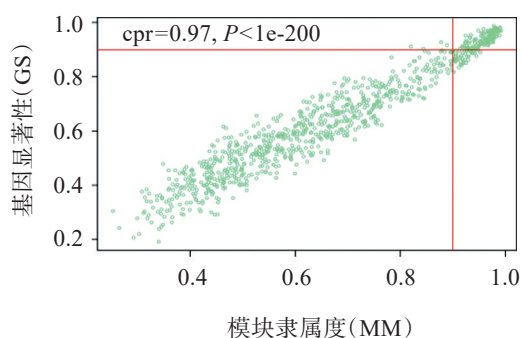


图6 turquoise模块内基因MM与GS散点图(红线表示GS=0.9,MM=0.9)

学过程和神经活动配体与受体相关作用与细胞因子-细胞因子受体相互作用等通路途径,其中已有文献报道^[11-12],功能和通路中相关分子参与了PD进展的过

程,本研究结论与前述文献报道较一致。我们随后对Turquoise模块内基因进行挖掘,最终识别出10个基因,包括IFITM5、MYPOP、SIK1、NPAS3、SMARCC2、FBXO17、SLC34A3、AMN、TNRC18和FBRSL1,在Turquoise模块内具有MM值和GS均大于0.9,且网络连接度较高的特征。我们将上述10个基因定义为PD相关的关键基因。与正常样本相比,10个基因在PD组织中的表达均显著升高。进一步文献检索发现,此10个基因目前在与PD相关的研究中均较少出现。利用genecards数据库检索此10个基因与PD的相关性得知,IFITM5基因和FBRSL1在genecards数据库与PD均具有一定的相关性,其中IFITM5基因即干扰素诱导跨膜蛋白5,作为干扰素诱导跨膜蛋白家

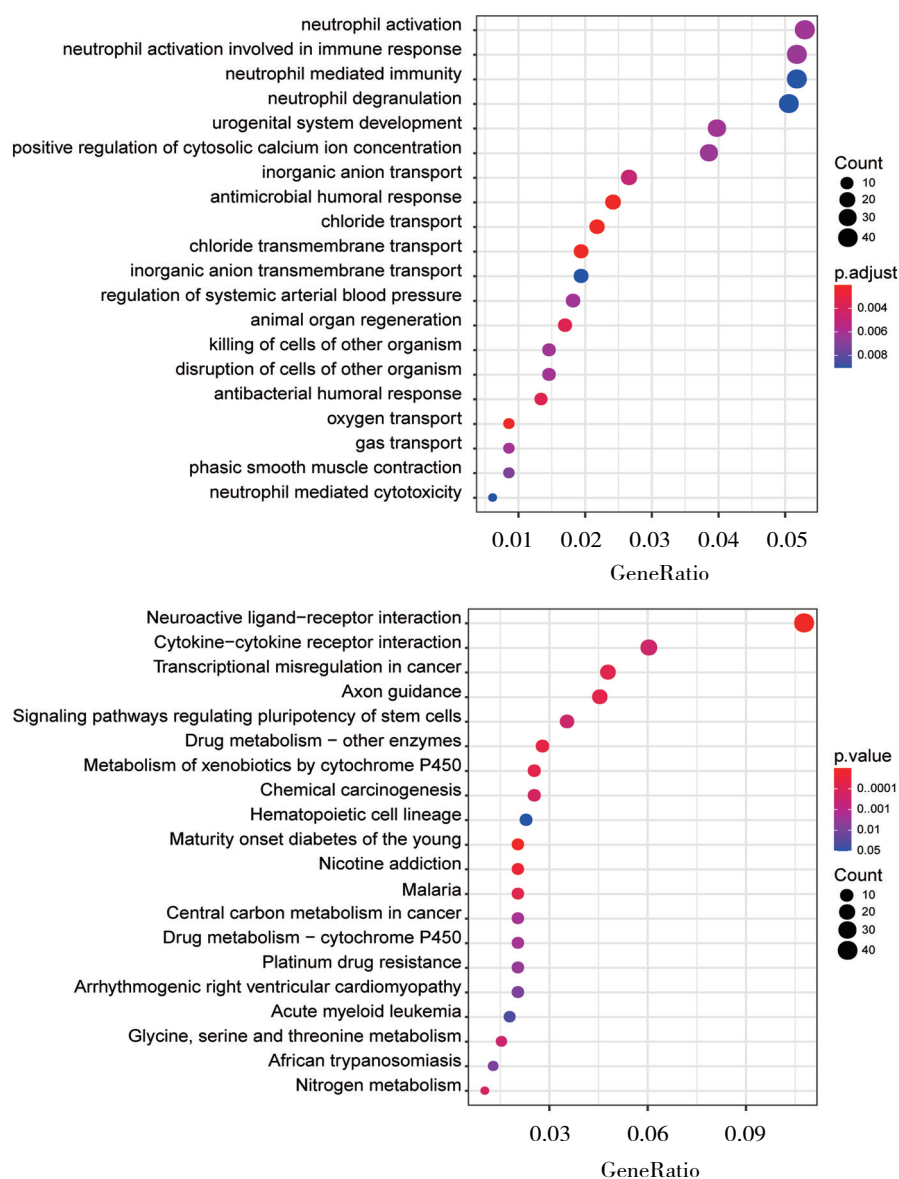


图7 turquoise模块内基因GO生物学过程和KEGG通路分析

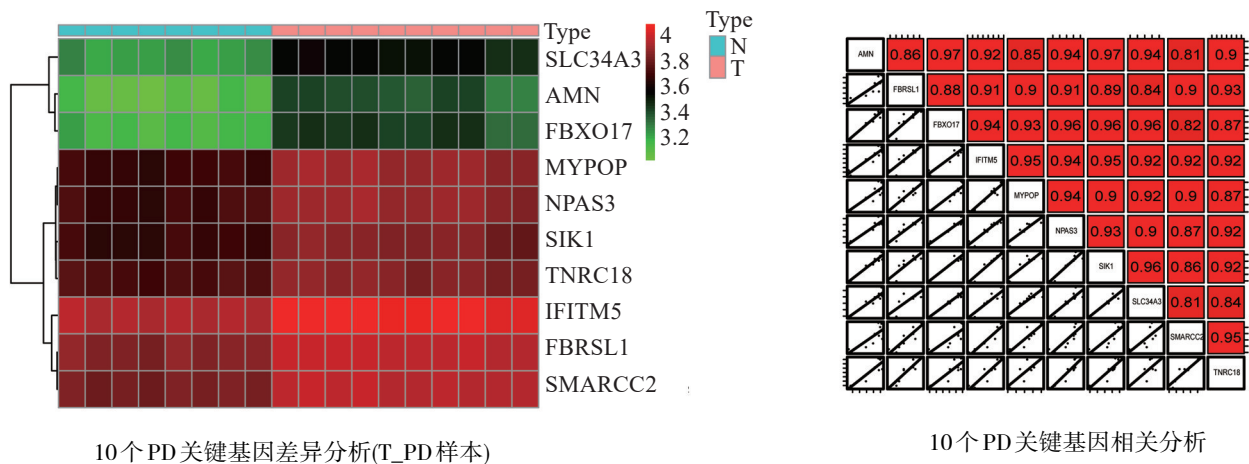


图8 10个PD相关关键基因差异分析和共表达分析

族成员之一,在成骨细胞矿化过程中的具有重要作用。鉴于有学者进行骨髓基质细胞向神经细胞分化及对PD的治疗研究^[13],我们推测 IFITM5 可能通过影响骨髓基质细胞向成骨细胞和神经细胞的转化过程环节参与了PD的发生、发展过程,尽管在 genecards 数据库中检索到 FBRSL1 基因属于PD相关基因,但其具体机制并不明确。考虑到 WGCNA 构建同一模块内基因具有相似性的思路,我们对此10个基因进行了相关性分析,结果显示,9个基因间均与 IFITM5 具有非常强的相关性,其参与PD进程的具体机制值得进一步探索。

综上所述,本研究利用 WGCNA 分析方法识别出了与PD具有强相关性的基因模块。通过该模块内基因进行挖掘,识别出10个与PD强相关性的关键基因,但目前相关研究仍非常缺乏。因此,本研究的完成将会为进行PD致病机理的相关研究及PD药物靶点研究提供重要的候选靶点分子。

参考文献:

- [1] 李一,熊莹,张远.利用加权基因共表达网络挖掘乳腺癌相关疾病靶标[J].第二军医大学学报,2019,40(9):1001-1009.
- [2] 肖观发,石胜军,胡万里.共表达网络分析筛选肾透明细胞癌进展和预后相关基因[J].现代泌尿生殖肿瘤杂志,2019,11(4):227-232,237.
- [3] WANG H,LIU J,LI J,et al.Identification of gene modules and hub genes in colon adenocarcinoma associated with pathological stage based on WGCNA analysis[J]. Cancer Genetics, 2020, 242:1-7.
- [4] LIN P I, SHU H, MERSHA T B. Comparing DNA methylation profiles across different tissues associated with the diagnosis of pediatric asthma[J]. Sci Rep, 2020, 10(1):151.
- [5] RITCHIE M E, PHIPSON B, WU D, et al. Iimma powers differential expression analyses for RNA-sequencing and microarray studies[J]. Nucleic Acids Res, 2015, 43(7):e47.
- [6] LANGFELDER P, HORVATH S.WGCNA: An R package for weighted correlation network analysis[J]. BMC Bioinformatics, 2008, 9:559.
- [7] YU G, WANG L G, HAN Y, et al.Cluster Profiler:An R package for comparing biological themes among gene clusters[J]. OMICS, 2012, 16(5):284-287.
- [8] SCHNEIDER S A, OBESO J A. Clinical and pathological features of parkinson's disease[J]. Curr Top Behav Neurosci, 2015, 22:205-220.
- [9] SERGE PRZEDBORSKI. The two-century journey of Parkinson disease research[J]. Nat Rev Neurosci,2017,18(4):251-259.
- [10] 钟花,王埃.脑源性神经营养因子在帕金森病治疗中的研究进展[J].海南医学,2018,29(19):2781-2784.
- [11] 杨依航. TaClo 介导的氧化应激和炎症反应在诱导帕金森病中的机制研究[D]. 济南:山东大学,2019.
- [12] 杨小英,杨新玲,徐军伟.中性粒细胞/淋巴细胞比值与帕金森合并周围神经病变的相关性分析[J].中风与神经疾病杂志,2019,36(6):546-548.
- [13] 单立冬.骨髓基质细胞向神经细胞分化及对帕金森病的治疗[D].苏州:苏州大学,2004.