[文章编号]　1007-3949(2020)28-10-0867-08　·实验研究·

动脉粥样硬化合并糖尿病的加权基因共表达网络构建与分析

孙 琪，徐 莹，李兴江，刘跃光
（牡丹江医学院，黑龙江省牡丹江市 157000）

[关键词]　生物信息学；　加权基因共表达网络分析；　动脉粥样硬化；　糖尿病

[摘　要]　**目的**　利用加权基因共表达网络分析(WGCNA)来识别动脉粥样硬化(As)合并糖尿病有关的功能基因模块。**方法**　从基因表达数据库(GEO)中下载 GSE23304 数据集，其中包含 101 例 As 外周斑块样本(其中 25 例患有糖尿病)的基因表达谱，使用 WGCNA 对数据集进行模块化分析并关联临床表型数据，根据相关系数大小，识别出与 As 最为相关的表型所在的模块，并对模块内基因进行功能注释(DAVID)，最后使用 STRING 进行蛋白互作网络分析。**结果**　使用 WGCNA 分析，最终得到了 33 个模块，其中 As 中识别到的 Darkturquoise 模块与糖尿病最为相关，认为 Darkturquoise 为糖尿病合并 As 的关键基因模块。**结论**　WGCNA 分析方法识别出的 As 关键基因模块 Darkturquoise 在糖尿病中可能起到重要作用。

[中图分类号]　Q33；R5　　[文献标识码]　A

Construction and analysis of weighted gene co-expression network for atherosclerosis with diabetes

SUN Qi, XU Ying, LI Xingjiang, LIU Yueguang
(*Mudanjiang Medical University*, *Mudanjiang*, *Heilongjiang* 157000, *China*)

[**KEY WORDS**]　bioinformatics;　weighted gene co-expression network analysis;　atherosclerosis;　diabetes

[**ABSTRACT**]　**Aim**　To identify functional gene modules related to atherosclerosis (As) with diabetes by weighted gene co-expression network analysis (WGCNA).　**Methods**　GSE23304 data set containing 101 samples of atherosclerotic peripheral plaques (25 of them had diabetes) was downloaded from the gene expression omnibus (GEO), then the gene expression profile was correlated with phenotypic data and analyzed by WGCNA.　According to the correlation coefficient size, the study identified the module which phenotype is the most highly associated with atherosclerosis, with functional annotation (GO) of the genes in the module, and then used STRING for protein interaction network analysis.　**Results**　33 modules were obtained by WGCNA analysis, of which the Darkturquoise module identified by atherosclerosis is the most relevant to diabetes, and Darkturquoise is considered to be a key gene module for diabetes in atherosclerosis.　**Conclusion**　The atherosclerosis key gene module identified by WGCNA analysis may play an important role in diabetes.

动脉粥样硬化(atherosclerosis，As)是心血管系统中常见的慢性血管炎性病变，是许多心脑血管疾病的病理基础。已有调查显示，每年约有 0.3% 的中老年糖尿病患者死于心血管疾病[1]，由此可见心血管疾病合并糖尿病的严重性。另外，As 也是糖尿病致死的主要原因之一，糖尿病患者较未患糖尿病者更易发生 As。在糖尿病前期中，多数患者合并存在血脂代谢异常、血管钙化、胰岛素抵抗等，这些因素既能够导致 As 的发生，同时还可能使得不稳定的 As 斑块发生破裂，导致血栓形成以及血流中断，甚至会危及生命[2]。因此，本研究就 As 合并糖尿病构建加权基因共表达网络，探究在 As 合并糖尿病中共表达的基因模块，希望可以作为 As 合并糖尿病的分子标志物，为后续研究和治疗 As 合并糖尿病提供更多线索。

[收稿日期]　2019-11-01　　[修回日期]　2020-12-30
[基金项目]　黑龙江省自然科学基金重点项目(ZD2019H009)
[作者简介]　孙琪，硕士研究生，E-mail 为 985966015@qq.com。通信作者刘跃光，硕士，教授，硕士研究生导师，主要从事心血管研究，E-mail 为 18704651461@163.com。

1 材料和方法

1.1 数据来源

本课题所研究的 As 组织样本(GSE23304)[3]来自基因表达数据库(gene expression omnibus, GEO)[4],平台号为 GPL4372,该数据集包括 67 例 As 患者的 101 个 As 外周斑块样本的基因表达谱数据,其中有 25 例 As 患者患有糖尿病。本研究所使用的基因表达数据为 101 例样本的全基因组表达原始数据,而非经作者分析后筛选出的特异表达的基因。

1.2 共表达网络的构建和模块的识别

1.2.1 软阈值的选择与模块的初步划分 基因共表达网络应满足无尺度网络原则[5],即服从幂律分布。加权基因共表达网络分析(weighted gene co-expression network analysis, WGCNA)通过对加权系数 β 的选择使得基因共表达网络更加符合无尺度特征。满足无尺度网络的标准:在无尺度网络中,包含连接度为 k 的节点的个数的对数[log(k)]与该节点出现的概率的对数[log(p(k))]要呈现负相关关系,且二者的相关系数要大于 0.85,此系数越高,表明该网络越符合无尺度网络规则。同时,对于每一个模块来说,每一个基因的平均连接度应需较高,这样的模块被检测到才有意义。

在本课题中,我们使用 R 语言运行 WGCNA 包[6],以此来构建基因共表达网络并识别出模块。首先选择合适的加权系数 β(软阈值)使得构建的基因共表达网络满足上述无尺度网的标准,并使用加权系数将相关矩阵转换为邻接矩阵,进一步转换为拓扑重叠矩阵(topological overlap matrix, TOM),并计算相异度。然后使用相异度作为距离测度,并设定最小模块中包含的基因个数为 30 个基因,之后采用动态剪枝法对模块进行初步划分并绘制基因树状图,其中使用聚类树的分支以及不同的颜色来代表不同的基因模块。

1.2.2 模块的合并与临床信息关联 使用 PCA 主成分分析对每个模块内的所有基因进行降维,选择每个模块中的第一主成分基因作为代表该模块内基因表达的整体水平的特征向量基因,然后根据特征基因的表达量进行聚类,并规定高度小于 0.25 也就是模块特征值之间的皮尔森相关系数大于 0.75 为相似度较高的两个模块,将被合并为一个新的模块,并绘制各模块特征向量基因的相关性图。下一步关联临床信息表型数据,本研究根据 GSE23304 数据集样本临床信息,决定纳入的临床信息有年龄(Age)、性别(G)、吸烟(S)、糖尿病(D)、血管紧张素转换酶抑制剂(ACEI)、血管紧张素受体拮抗剂(ARB)、调血脂治疗(AntiLip)以及钙通道阻滞剂(CCB)。由于只有连续型性状才能进行计算,所以我们将临床信息表转为 0-1 矩阵。其中使用模块的特征向量基因与纳入的临床信息变量的皮尔森相关系数,代表模块与临床信息的相关性,得到相关性最高的模块与表型进行后续分析。其中,本研究计算了每个基因的表达量与临床信息表型的皮尔森相关系数,即基因显著性(gene significance, GS),以及每个基因与其所属模块的相关系数,即模块隶属度(module membership, MM)来衡量该模块内基因与表型、基因与模块之间的相关性。

1.3 模块的功能富集及蛋白互作网络分析

为了解识别到的模块内基因所涉及的生物学功能,本课题将使用 DAVID(https://david-d.ncifcrf.gov/)网站的在线基因本体(gene ontology, GO)富集分析工具及京都基因与基因组百科全书(Kyoto Encyclopedia of Genes and Genomes, KEGG)通路分析工具进行注释[7]。蛋白互作网络分析使用 STRING(https://string-db.org)[8]。

2 结 果

2.1 共表达网络的构建和模块的识别

2.1.1 软阈值的选择与模块的初步划分 为挑选最为符合无尺度网络构建的 β 值,同时需考虑对网络节点(node)平均连接度的适当保留,最终本研究以 log(k)与 log(p(k))的相关系数大于 0.85 的 β=4 作为软阈值来构建共表达网络(图 1)。在本文中我们采用动态剪枝法进行模块的初步划分,最终划分得到 55 个模块(图 2)。

2.1.2 模块的合并与临床信息关联 经过模块的初步划分,合并相似度较高的模块后得到 33 个模块(图 3),通过绘制 33 个模块相关性图验证各模块之间相对独立,可根据划分出的 33 个模块进行后续分析。

将各个划分出的模块与 As 样本的临床信息进行关联分析,得到与临床信息最为相关的模块进行后续研究。结果发现,糖尿病(D)与 Darkturquoise 模块较为相关($r=0.29$, $P=0.03$;图 4)。通过计算 Darkturquoise 模块内基因 GS 与 MM 相关系数($r=0.57$, $P=1.5\times10^{-8}$)(图 5),也进一步验证了 Darkturquoise 模块与糖尿病关联程度最大。

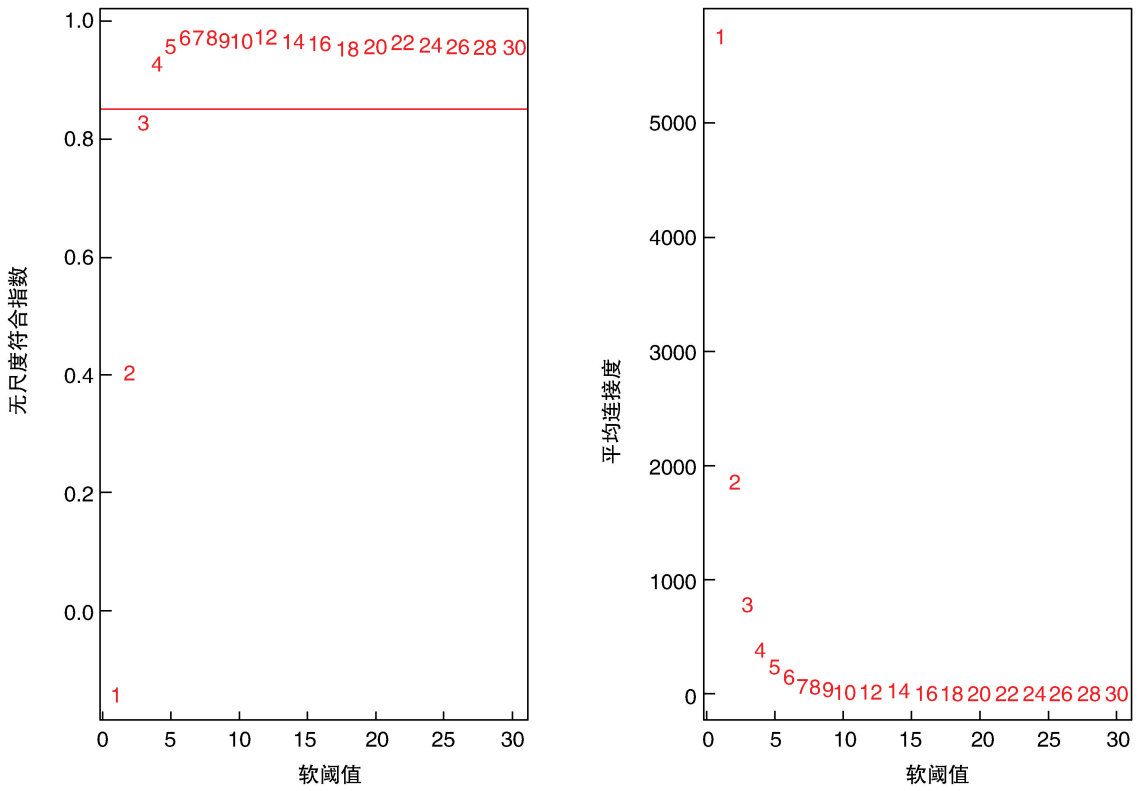

图 1. 加权基因共表达网络分析软阈值(β)的确定　左图为不同 β 下计算的无尺度符合指数,右图为不同 β 下计算的平均连接度。

Figure 1. Determination of soft threshold (β) of weighted gene co-expression network analysis

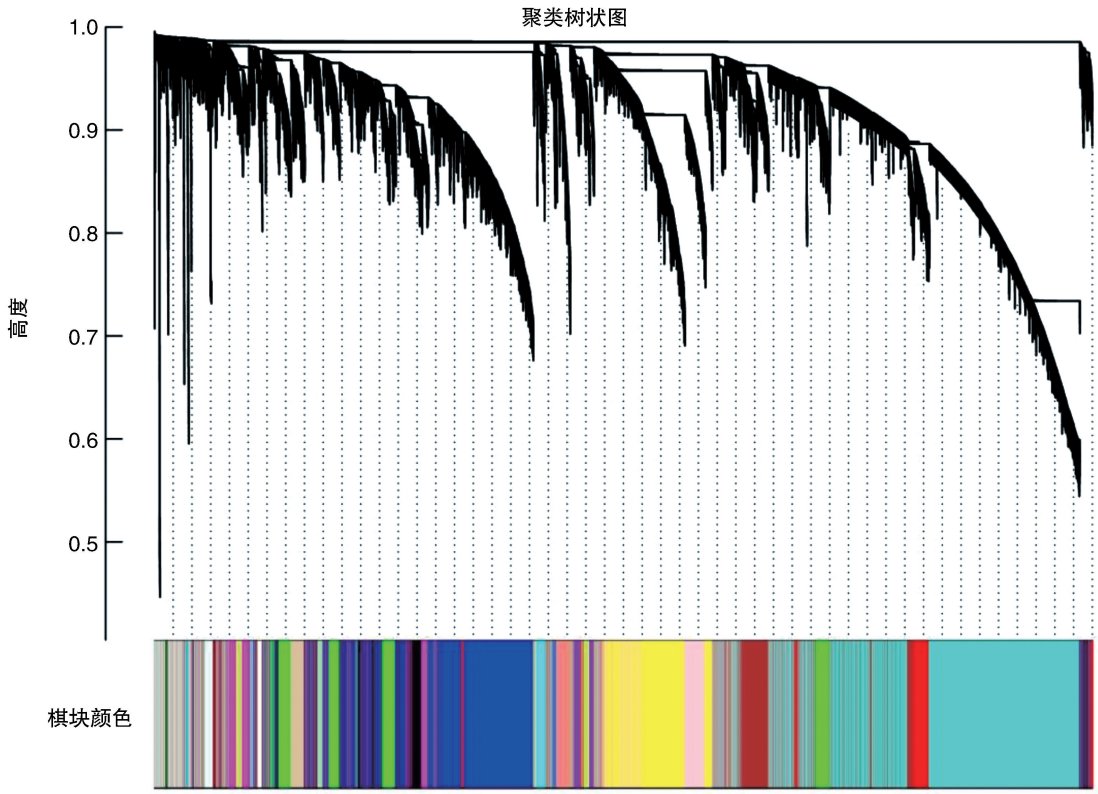

图 2. 基因聚类树状图　基于邻接相似性的层次聚类获得的基因聚类树(树状图),树状图下方的彩色行表示通过动态树切割方法识别的模块。

Figure 2. Clustering dendrograms of genes

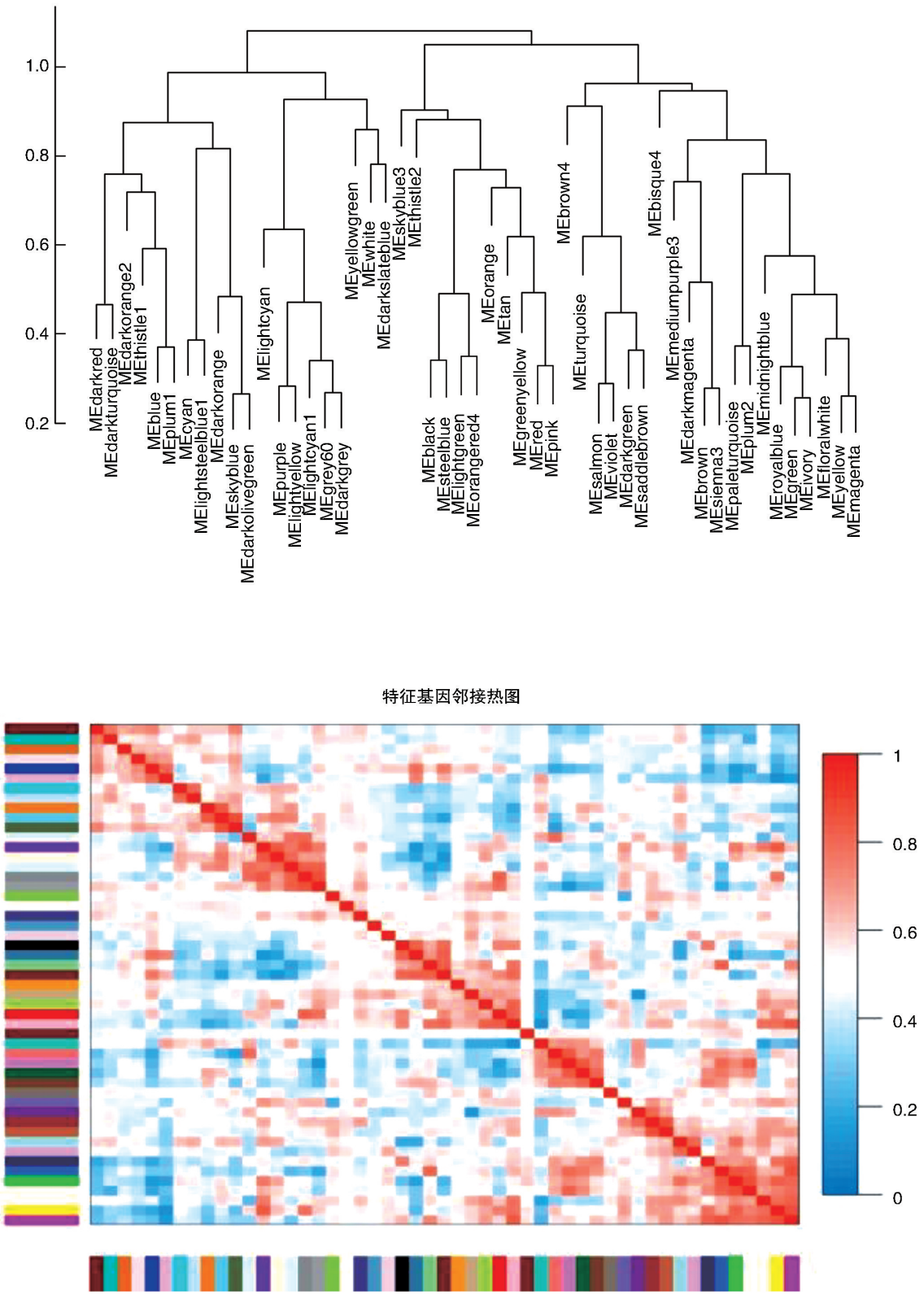

图 3. 模块相关性图 X 和 Y 轴上的不同颜色表示不同的模块，红色或蓝色的亮度表示不同模块的相关性，颜色越红表示相关性越强。

Figure 3. Correlation diagram between modules

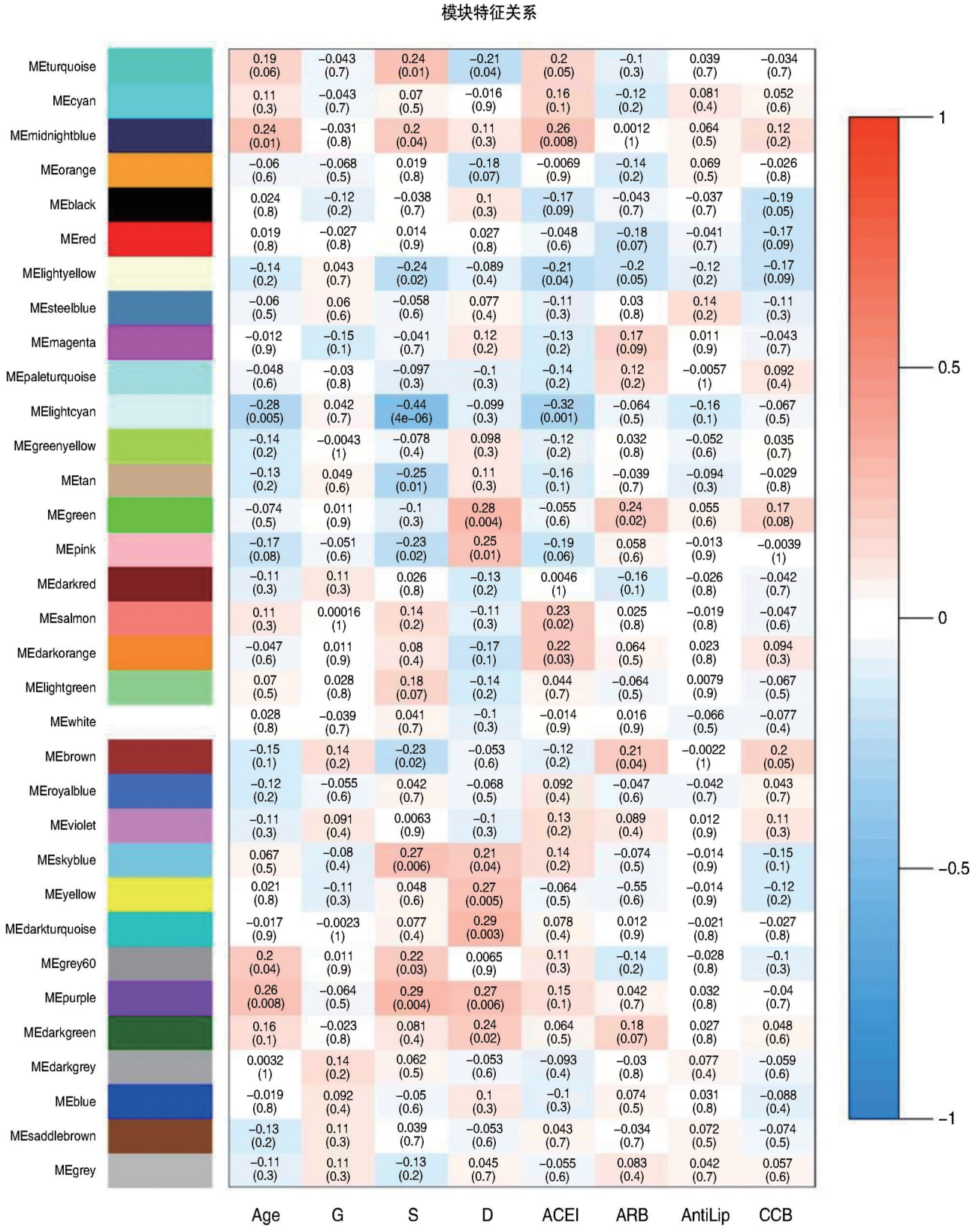

图 4. 基因模块与临床信息关联性　每行代表模块特征基因，每列代表临床特征信息；每个单元格内第一行为相应的相关性值，第二行为 *P* 值；单元格颜色越红越正相关，越蓝越负相关。

Figure 4. Module-feature associations

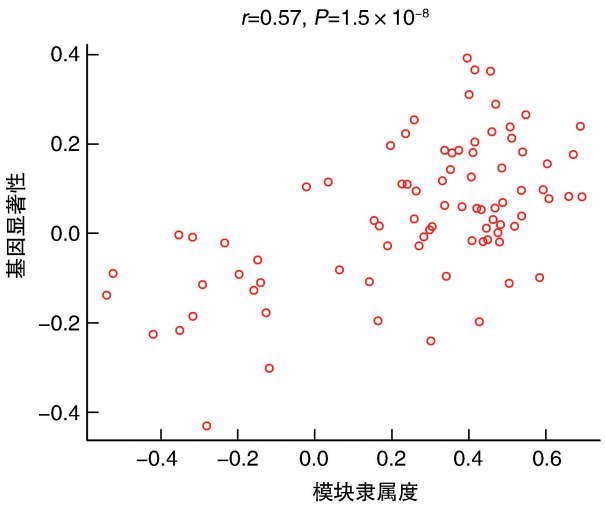

图 5. 基因显著性(GS)与模块隶属度(MM)相关性

Figure 5. Correlation between GS and MM

2.2 模块的功能富集及蛋白互作网络分析

使用 DAVID 对识别出来的 Darkturquoise 模块内的所有基因进行富集分析,结果发现,模块内的基因大多数参与了信号转导与通信等生物学过程,正如我们所知,糖尿病与许多胰岛素等信号转导通路相关,这也部分解释了 Darkturquoise 模块与糖尿病之间的关联(表 1 及图 6)。在分子功能上,模块内基因与同一蛋白结合紧密相关。另外,在细胞组分方面,与细胞质部分、细胞内以及细胞内细胞器等组分相关。除此之外,本文还研究了识别到的 Darkturquoise 模块中蛋白质相互作用情况(图 7)。

3 讨 论

As 合并糖尿病发病机制尤为复杂,涉及到多种基因[9],以往的对于单个基因的研究分析方法已经不能够对 As 以及糖尿病进行深入的研究。因此,本研究选用了 WGCNA 方法,此方法构建了一个复杂的基因共表达网络,并在该高通量数据中划分并筛选出功能相关的模块。通过该生物信息方法,研究人员可以得到模块内基因的相关性,以及模块与模块间基因的相关性,还可以与临床信息相关联,得到与疾病相关的临床信息,并进行进一步的研究。以往有研究报道,He 等[10]学者应用 WGCNA 方法鉴别出与乳头状肾细胞癌相关的生物标志物,可见 WGCNA 方法更加能够从整体来分析基因的功能以及基因间的内在联系[11]。

本研究应用 WGCNA 方法,对纳入的 GSE23304 数据集中的 101 例 As 外周斑块样本展开分析,最终

表 1. Darkturquoise 模块富集分析结果

Table 1. The result of enrichment analysis about Darkturquoise module

模块	功能	数量	*P* 值
GO_MF	同一蛋白质结合	15	0.003014
GO_MF	热休克蛋白结合	4	0.007648
GO_MF	p53 结合	3	0.037778
GO_MF	阴离子结合	5	0.040753
GO_MF	乙醇结合	3	0.045222
GO_BP	单体分解代谢过程	11	0.003079
GO_BP	信号转导调控	21	0.010022
GO_BP	细胞应激反应	15	0.018307
GO_BP	细胞组件组装	19	0.027348
GO_BP	细胞通信调节	21	0.028496
GO_BP	蛋白质折叠调控	2	0.029863
GO_BP	高分子复合物亚基组织	18	0.030879
GO_BP	细胞定位的建立	15	0.046015
GO_CC	细胞质	61	0.000424
GO_CC	细胞质部分	50	0.000565
GO_CC	细胞内细胞器部分	49	0.001974
GO_CC	细胞外	70	0.002184
GO_CC	内质网	16	0.003844
GO_CC	细胞内膜结合细胞器	59	0.005068
GO_CC	细胞内	70	0.006958
GO_CC	内膜系统	27	0.008077
GO_CC	细胞内细胞器	62	0.010698
GO_CC	液泡膜	8	0.014932
GO_CC	内质网膜	10	0.02546
GO_CC	内质网部分	11	0.026903
GO_CC	核外膜-内质网膜网络	10	0.028773
GO_CC	液泡部分	8	0.032759
GO_CC	包涵体	3	0.038617
GO_CC	ESCRT Ⅲ复合体	2	0.049862

划分出了 33 个功能模块。本研究发现,由 As 样本基因表达谱识别出的含有 84 个基因的 Darkturquoise 模块与糖尿病最为相关,并通过富集分析发现,该模块在 GO 分类上与糖尿病相关进程有很大关联:如信号转导的调节[12]、细胞通信的调节[13]等。同时,STRING 蛋白互作网络显示,从 Darkturquoise 模块中识别到多种已经有文献报道的与糖尿病相关的“著名”基因,如 APOE[14]、OGDH[15]等。这意味着 Darkturquoise 模块中的基因可能存在作为 As 合并糖尿病分子标志物的潜能,并

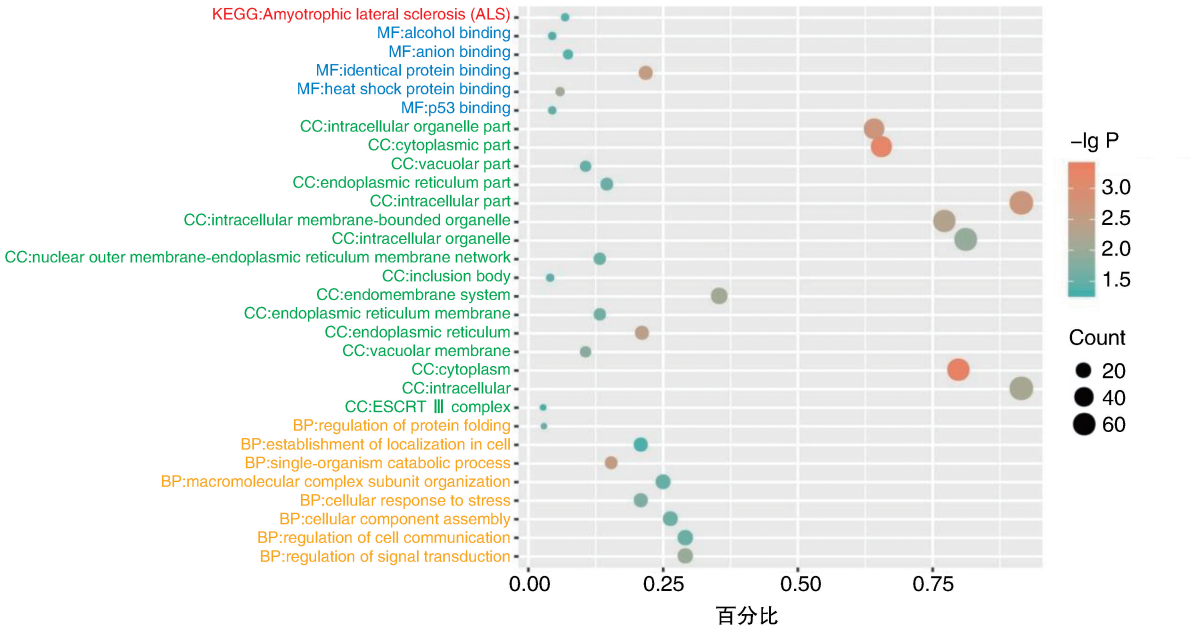

图 6. Darkturquoise 模块的 GO 功能与 KEGG 通路富集分析结果　点的大小表示基因数量，Y 轴表示 GO 和 KEGG 的功能；图例颜色是根据−lg P 的值来标注的。

Figure 6. GO functional and KEGG pathway enrichment analysis for genes in the Darkturquoise module

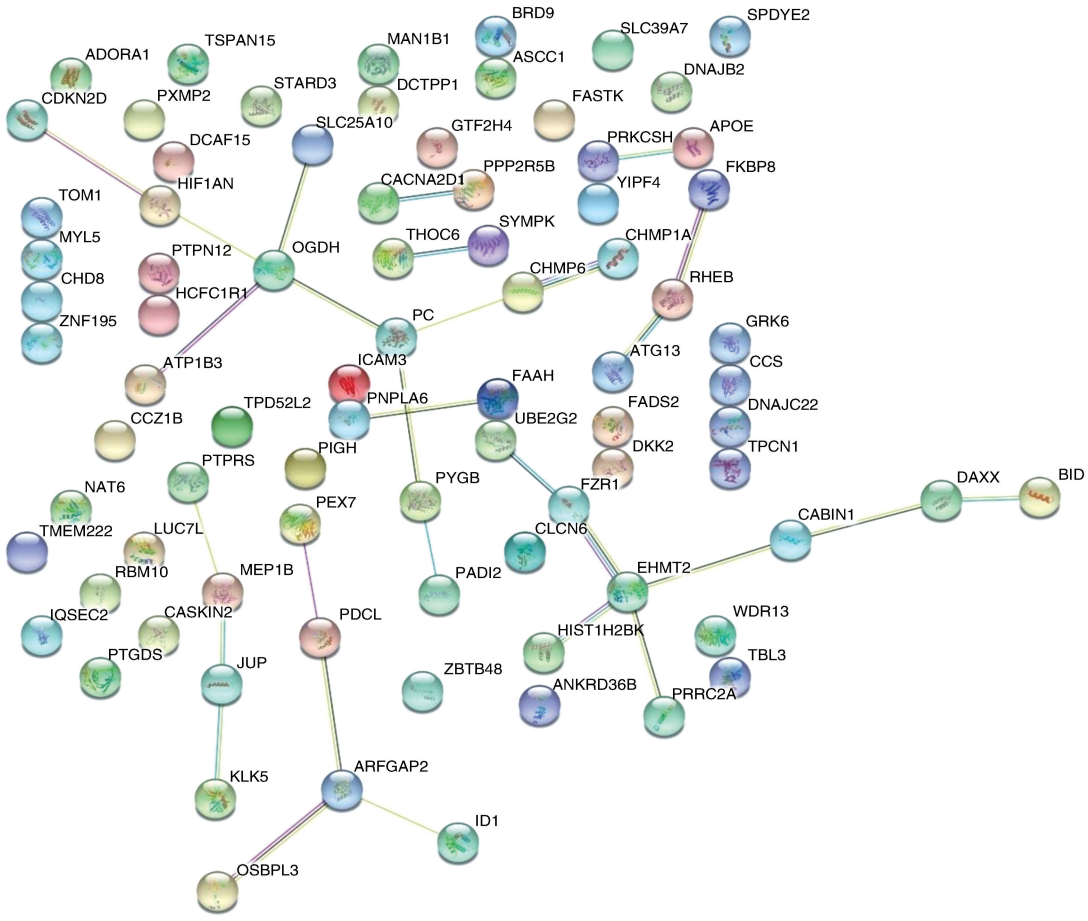

图 7. 模块内蛋白互作网络　每个节点代表一种蛋白质，节点之间的线表示该对蛋白质具有相互作用关系。天蓝色和紫色线表示已知的相互作用；绿色、红色、蓝色代表预测的相互作用；其他颜色表示其他交互，如黑色代表共表达。

Figure 7. Protein-protein interaction network

且能够对二者起到调控作用,为后续研究 As 合并糖尿病提供了更多的线索。

本研究的不足之处在于未能从 Darkturquoise 模块中识别到与 As 合并糖尿病相关的枢纽(hub)基因,原因在于该模块内基因连接度均差不多,没有能够识别到连接度较大的节点,从而未能展示几个较为相关的枢纽基因。但同时,由于涉及 As 合并糖尿病共同关联的基因的研究还不多,又是使用 WGCNA 方法从生物功能整体来识别模块,所以本研究认为该模块中的基因均存在研究的价值,希望能够为研究人员的深入分析提供一些线索。

综上,随着生物信息学的快速发展,生物信息学现已逐步的应用于医学中。一方面,符合"精准医疗"的要求[16];另一方面,生物信息能够从高通量的角度来分析和研究基础以及临床问题[17],为医学领域提供了新的视角。随着算法的不断更新,生物信息将在医学领域发挥着更加重要的作用。

[参考文献]

[1] Cho NH, Shaw JE, Karuranga S, et al. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045 [J]. Diabetes Res Clin Pract, 2018, 138: 271-281.

[2] Low Wang CC, Hess CN, Hiatt WR, et al. Clinical update: cardiovascular disease in diabetes mellitus: atherosclerotic cardiovascular disease and heart failure in type 2 diabetes mellitus - mechanisms, management, and clinical considerations [J]. Circulation, 2016, 133(24): 2459-2502.

[3] Puig O, Yuan J, Stepaniants S, et al. A gene expression signature that classifies human atherosclerotic plaque by relative inflammation status [J]. Circ Cardiovasc Genet, 2011, 4(6): 595-604.

[4] Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets--update [J]. Nucleic Acids Res, 2013, 41(Database issue): D991-995.

[5] Dennis G, Sherman BT, Hosack DA, et al. DAVID: Database for annotation, visualization, and integrated discovery [J]. Genome Biol, 2003, 4(5): P3.

[6] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis [J]. BMC Bioinformatics, 2008, 9: 559.

[7] Zhao W, Langfelder P, Fuller T, et al. Weighted gene coexpression network analysis: state of the art [J]. J Biopharm Stat, 2010, 20(2): 281-300.

[8] Szklarczyk D, Franceschini A, Wyder S, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life [J]. Nucleic Acids Res, 2015, 43(Database issue): D447-452.

[9] Bhupathiraju SN, Hu FB. Epidemiology of obesity and diabetes and their cardiovascular complications [J]. Circ Res, 2016, 118(11): 1723-1735.

[10] He Z, Sun M, Ke Y, et al. Identifying biomarkers of papillary renal cell carcinoma associated with pathological stage by weighted gene co-expression network analysis [J]. Oncotarget, 2017, 8(17): 27904-27914.

[11] Zhai X, Xue Q, Liu Q, et al. Colon cancer recurrenceassociated genes revealed by WGCNA coexpression network analysis [J]. Mol Med Rep, 2017, 16(5): 6499-6505.

[12] Zhang WS, Pan A, Zhang X, et al. Inactivation of NF-kappaB2 (p52) restrains hepatic glucagon response via preserving PDE4B induction [J]. Nat Commun, 2019, 10(1): 4303.

[13] David JA, Rifkin WJ, Rabbani PS, et al. The Nrf2/Keap1/ARE pathway and oxidative stress as a therapeutic target in type Ⅱ diabetes mellitus [J]. J Diabetes Res, 2017, 2017: 4826724.

[14] Liu S, Liu J, Weng R, et al. Apolipoprotein E gene polymorphism and the risk of cardiovascular disease and type 2 diabetes [J]. BMC Cardiovasc Disord, 2019, 19(1): 213.

[15] Yu Q, Liu B, Ruan D, et al. A novel targeted proteomics method for identification and relative quantitation of difference in nitration degree of OGDH between healthy and diabetic mouse [J]. Proteomics, 2014, 14(21-22): 2417-2426.

[16] Konig IR, Fuchs O, Hansen G, et al. What is precision medicine? [J]. Eur Respir J, 2017, 50(4). DOI: 10.1183/13993003.00391-2017.

[17] Doring Y, Noels H, Weber C. The use of high-throughput technologies to investigate vascular inflammation and atherosclerosis [J]. Arterioscler Thromb Vasc Biol, 2012, 32(2): 182-195.

(此文编辑 许雪梅)