

·数字中医药·

本文引用:林景峰,高 强,刘甘露,马华萍,胡文悦,韩振蕴. 基于多组全基因组表达谱的阳虚人群关键候选基因集和通路筛选及生物信息学分析[J]. 湖南中医药大学学报,2021,41(3): 425-430.

基于多组全基因组表达谱的阳虚人群关键候选 基因集和通路筛选及生物信息学分析

林景峰¹,高 强¹,刘甘露¹,马华萍¹,胡文悦¹,韩振蕴^{2*}

(1.北京中医药大学,北京 100029;2.北京中医药大学深圳医院(龙岗),广东 深圳 518100)

[摘要] **目的** 基于多组全基因组表达谱筛选阳虚人群关键候选基因集和通路,探讨不同阳虚人群与对照组差异基因的异同,提出阳虚与基因表达关系的可能结论。**方法** 查找 GEO 基因表达数据库及 PubMed、Embase、CNKI、万方、维普等中英文文献数据库,筛选出其中存在阳虚人群及其对照组的基因表达谱数据或基因表达谱分析结果。应用 R 语言和生物信息学方法筛选差异表达基因,并进行 GO、KEGG 和 GSEA 富集分析,利用韦恩图展现不同阳虚人群的差异表达基因结果的关系,并分析差异基因、富集结果与阳虚之间的关系。**结果** 共得到 2 个数据集(GSE87474、GSE56116)和 4 个基因表达谱分析结果,数据集 GSE56116 中存在 350 个差异表达基因,数据集 GSE87474 中存在 138 个差异表达基因。4 个基因表达谱进行去重与合并后,形成 3 个差异基因集,分别报道了 190 个、66 个和 21 个差异表达基因。对差异表达基因结果取交集,未发现重合的基因。对其中差异表达基因相关的基因集和通路取交集分析,寻找到 8 个共同的基因集和通路。这些通路的功能集中在免疫功能、细胞质囊泡等方面。**结论** 阳虚人群与非阳虚人群的差异基因集和通路主要与能量代谢、细胞质囊泡及免疫调节相关。不同阳虚人群间关键候选基因集和通路存在一定的相似性,但其作用的具体基因存在差异。中医的阳虚概念更可能与一系列基因集和通路存在紧密的联系,而不是单一的基因。

[关键词] 阳虚;全基因组表达;基因集;通路;生物信息学

[中图分类号] R2-0

[文献标志码] A

[文章编号] doi:10.3969/j.issn.1674-070X.2021.03.020

Bioinformatics Analysis Along with Screening on Key Candidate Gene Sets and Pathways in Yang Deficiency Population Based on Genome-wide Expression Profile

LIN Jingfeng¹, GAO Qiang¹, LIU Ganlu¹, MA Huaping¹, HU Wenyue¹, HAN Zhenyun^{2*}

(1. Beijing University of Chinese Medicine, Beijing 100029, China; 2. Shenzhen Hospital of Beijing University of Chinese Medicine (Long Gang), Shenzhen, Guangdong 518100, China)

[Abstract] **Objective** To screen the key candidate gene sets and pathways in the Yang deficiency population based on multiple genome-wide expression profile sets. To explore the differences and similarities between the different genes in the Yang deficiency syndrome population and the control group, and to propose a possible hypothesis on the relationship between Yang deficiency and gene expression. **Methods** Searching for GEO gene expression database and PubMed, Embase, CNKI, Wanfang, Viper databases in language of Chinese and English, and screened out gene expression profile data sets or gene expression profile analysis results of Yang deficiency group and its control group. Differentially expressed genes were screened by R software and bioinformatics methods, and gene ontology (GO) analysis, kyoto encyclopedia of genes and genomes (KEGG) and gene set enrichment analysis (GSEA) were carried out. The Venn diagram was used to show the relationship between the results of differentially expressed genes in different people with Yang deficiency, and the relationship between the results of differentially expressed genes, enrichment and Yang deficiency was analyzed. **Results** Two data sets (GSE87474, GSE56116)

[收稿日期] 2020-09-18

[基金项目] 国家重点研发计划项目(2019YFC1710103)。

[作者简介] 林景峰,男,在读博士研究生,研究方向:中医脑病学、人工智能与中医、生物信息学与中医。

[通讯作者] * 韩振蕴,女,教授,主任医师,博士研究生导师,E-mail:tohanzhenyun@sina.com。

and four gene expression profiles were obtained. There were 350 differentially expressed genes in the data set GSE56116 and 138 differentially expressed genes in the data set GSE87474. Four gene expression profiles were selected and combined to form three differentially expressed gene sets. 190, 66 and 21 differentially expressed genes were reported respectively. The results of differentially expressed genes were overlapped and no coincidence genes were found. The gene sets and pathways related to differentially expressed genes were analyzed and 8 common gene sets and pathways were found. The functions of these pathways were concentrated in immune function, cytosolic vesicles and so on. **Conclusion** The different gene sets and pathways between Yang deficiency and non-Yang deficiency populations are mainly related to energy metabolism, cytoplasmic vesicles and immune regulation. The key candidate gene sets and pathways have some similarities among different Yang deficiency populations, but the expressions of specific genes are different. The concept of Yang deficiency in Chinese medicine is more likely to be related to a series of gene sets and pathways rather than a single gene.

[**Keywords**] Yang deficiency; genome-wide expression; gene set; pathway; bioinformatics

阳虚人群包括阳虚证患者和阳虚体质人群。阳虚证、阳虚体质是中医常见的一种证候和体质,通常都有畏寒怕冷、手足不温、喜热饮食等一系列表现^[1]。近年来,中医证候和体质的基因组学、转录组学和蛋白质组学研究逐渐成为分子机制研究的热点,而基因组学整体性及稳定性特点与中医证候特性有异曲同工之处^[2],利用全基因组表达谱对具有阳虚证候群的人群进行生物信息学分析具有可行性。既往亦有不同阳虚人群与对照组基因表达状况的相关研究^[3-7]。但此类研究有样本量较少等不足,同时,不同阳虚人群之间的基因表达差异尚不明确,而其共同差异基因、共同基因集和通路可能与阳虚证候群有较为确切的联系。本研究团队运用生物信息学相关方法对阳虚人群相关基因芯片数据进行分析,结合既往阳虚人群的基因表达研究结果,探讨不同阳虚人群与对照组差异基因、基因集和通路的异同,提出阳虚与基因表达关系的可能结论。

1 资料与方法

1.1 数据获取

查找基因表达数据库(gene expression omnibus, GEO)及 PubMed、Embase、CNKI、万方、维普等中英文文献数据库,筛选出其中存在阳虚证候或体质及其对照组的基因表达谱数据(包括基因芯片与转录组测序数据)或基因表达谱分析结果(差异基因集合)。检索时限均为建库至 2020 年 2 月 1 日。检索采用主题词和自由词相结合的方式,并追溯纳入文献的参考文献,以补充获取相关文献。中文检索词包括:阳虚、基因表达、转录组测序、基因芯片等,英文检索词包括:RNA expression, gene expression, expression profiling, Yang deficiency, TCM syndrome,

humans。未进行灰色文献的人工检索。纳入标准:(1)研究对象为人类。(2)若为分析结果,则报道的差异基因数量需>10 个。(3)试验组为“阳虚体质”“阳虚证”或阳虚相关证型,对照组无阳虚或中医其他证型。得到基因表达数据或基因表达谱分析结果后,对异质性较小的文献结果进行合并,以增加其差异基因数量。

1.2 基因表达谱数据

阳虚人群基因表达数据集来源于 GEO 数据库,分别为 GSE87474 和 GSE56116。GSE87474 包含 32 个中国汉族人的外周血单核细胞样本,分别为 12 个阳虚体质、12 个阴虚体质和 8 个平和体质样本。选取其中阳虚体质和平和体质样本进行生物信息学分析。GSE56116 包含 13 份中国绝经后骨质疏松症患者外周血样本,其中辨证为肾阴虚者 4 份、肾阳虚 3 份、无肾虚患者 3 份、健康对照组 3 份。选取其中肾阳虚和无肾虚患者人群进行生物信息学分析。另有数据集 GSE67090 为阳虚四逆证人群与健康人群的比较,但其原始文件与表达矩阵数据均非基因表达值,无法进行二次分析,故剔除。使用 R 语言(版本 3.6.2)进行数据集的生物信息学分析。

1.3 数据集预处理

下载 GSE56116 和 GSE87474 的探针表达矩阵,对其进行 Log₂ 转化和质量控制分析。然后根据 soft 注释文件将表达矩阵与基因名和 entrez ID 进行对应。多个探针对应同一个基因的,取最大表达量探针。

1.4 差异表达基因的筛选

数据集的差异基因需同时满足 $|\text{Log}_2 \text{ fold change} (\text{Log}_2\text{FC})| > 1$ 且 $P < 0.05$ 。利用 pheatmap 包里对前 30 显著差异基因(以 $|\text{Log}_2\text{FC}|$ 为标准)绘制热

图,直观地展示每个差异基因在每个样本中的表达情况。利用 Enhanced Volcano 包对所有基因绘制火山图,直观地展示每个基因在每个样本中的表达情况。同时提取文献中报道的差异基因集合。

1.5 差异表达基因的基因本体论(gene ontology, GO)、通路富集分析(kyoto encyclopedia of genes and genomes, KEGG)和基因富集分析(gene set enrichment analysis, GSEA)

对多组差异基因集合取并集得到差异基因并集。利用 R 语言 clusterProfiler 包对多组差异基因集合及差异基因并集进行 GO 分析、KEGG 分析。对 GSE56116 和 GSE87474 数据集进行 GSEA 分析。设定显著性基因富集的临床值为 $P < 0.05$ 。

1.6 多组数据集的共同差异基因、富集基因集和通路的筛选

利用 Venn Diagram 包绘制多组共同差异基因、基因集和通路的交集。利用 intersect 函数筛选得到共同基因集和通路。

2 结果

2.1 多组数据的获得和数据集差异基因筛选

共获得 2 个基因组表达谱数据集(GSE56116 和 GSE87474)。两组数据集分别有 32 696 个、28 267 个有效基因探针结果,对其进行 Log₂ 转化。清洗未标注基因名的探针、多个探针对应同一个基因取最大表达量探针后,分别得到拥有 19 749 个、17 447 个基因的表达矩阵。R 语言进行差异基因筛选, GSE56116 共筛选出 350 个差异基因,其中 92 个上调基因、258 个下调基因。GSE87474 共筛选出 138 个差异基因,其中 96 个上调基因、42 个下调基因。所有基因表达的火山图见图 1。

选取两个数据集中差异最大的 30 个基因绘制热图,见图 2。

其中 GSE56116 数据集中差异最大的 30 个基因里包括 4 个上调基因和 26 个下调基因。GSE87474 数据集中差异最大的 30 个基因里包括 2 个上调基

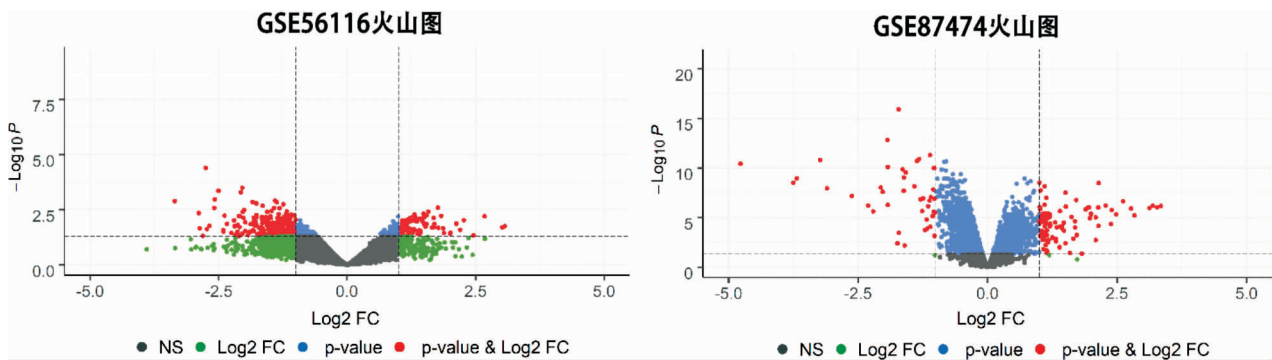


图 1 两组数据集基因表达火山图

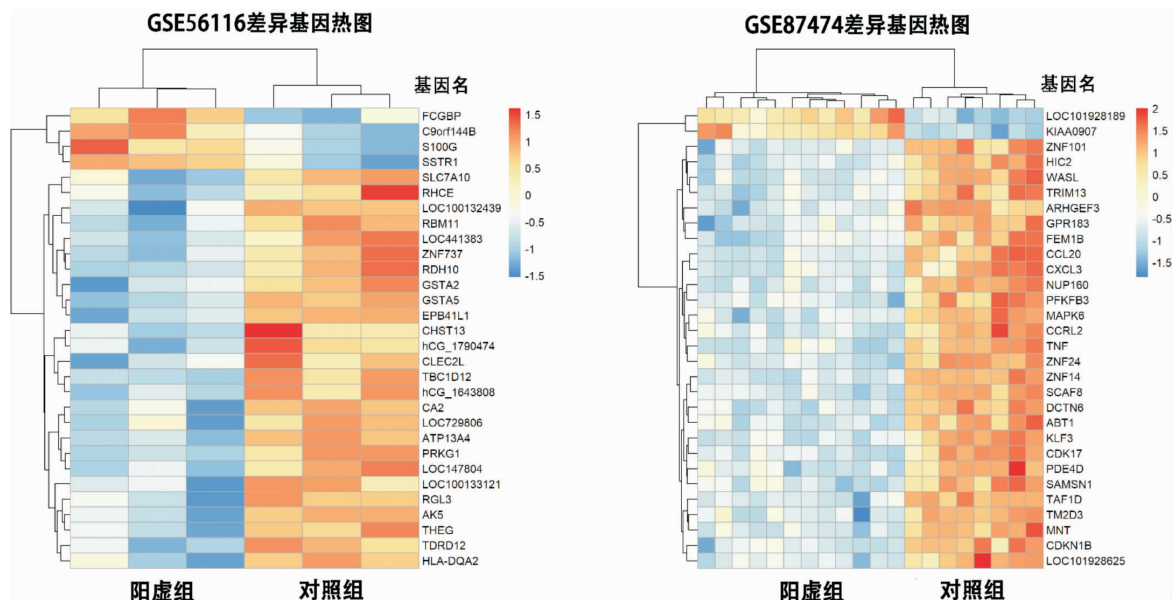


图 2 两组数据集差异基因热图

因和 28 个下调基因。

在文献的基因表达谱分析结果方面,提取到 4 个分析结果。其中,李艳艳^[4]报道了差异最大的 20 个基因,其余 3 个分析结果^[3,5,7]报道了其研究中的所有差异基因。GSE87474 在其注释文件中未对纳入人群的性别和年龄进行说明。Cheng^[7]在论文中未说明纳入受试者的性别和年龄。其中汤朝晖^[3]和李艳艳^[4]的研究为同一机构、相近时间的研究,且纳入试验的受试者信息相近,为增加差异基因总数,便于更好地进行富集分析,将两者差异基因取并集进行合并。6 个差异基因集合进行去重与合并后,形成 5 个差异基因集,分别为 GSE87474 138 个, GSE56116 350 个, Cheng^[7]190 个, 汤朝晖^[3]+李艳艳^[4]66 个和 杨嘉慧^[5]21 个差异表达基因。

2.2 多组差异基因的交集分析及差异基因并集的建立

对获取的 5 组差异基因取交集进行分析。结果显示,没有基因同时存在于 5 组、4 组及 3 组的差异基因集中。

2.3 GO 功能富集

将 5 组差异基因集取并集后共得到 742 个差异基因。GO 功能富集分析包括生物过程 (biological process, BP)、细胞组成 (cellular component, CC) 和分子功能 (molecular function, MF) 3 个方面。利用 clusterProfiler 包对差异基因并集进行 GO 功能富集分析,选取前 10 位的功能富集类别,见图 3。富集功能主要涉及中性粒细胞介导免疫、粒细胞趋化性、趋化因子介导的信号通路、细胞质囊泡腔等。

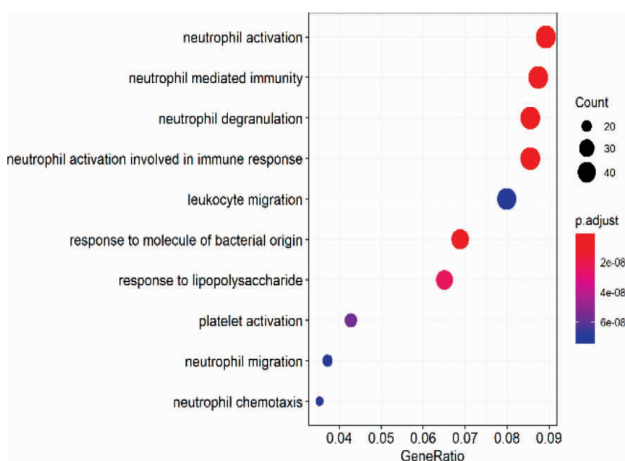


图 3 GO 功能富集结果 (前 10 位)

2.4 KEGG 通路富集

对差异基因并集分别进行 KEGG 通路富集分析,选取前 10 位的通路,见图 4。基因通路主要涉及造血细胞谱系、移植物抗宿主病、抗原处理及呈递等。

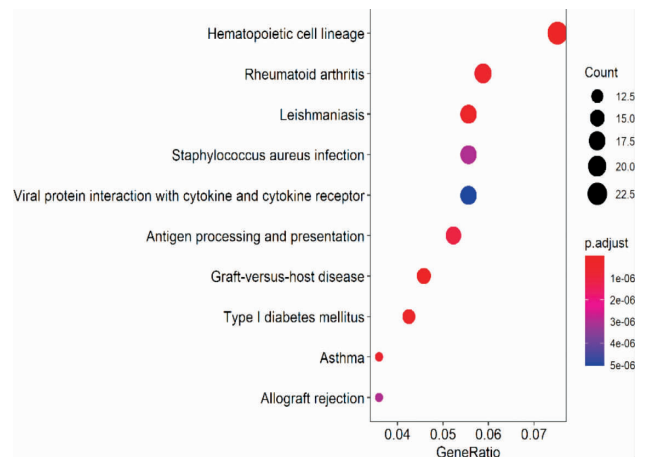


图 4 KEGG 基因通路富集结果 (前 10 位)

2.5 GSEA 基因富集分析

对 GSE87474 和 GSE56116 进行 GSEA 富集分析。选取 $P < 0.05$ 的富集通路,其中 GSE87474 存在 1 个富集通路, GSE56116 存在 2 个富集通路。

在 GSEA 分析中,从 2 个数据集中筛选出了 3 个基因富集功能,无共有富集功能。相比于 GO 基因富集分析和 KEGG 基因通路分析来说,其结果数量太少,不具有代表性。

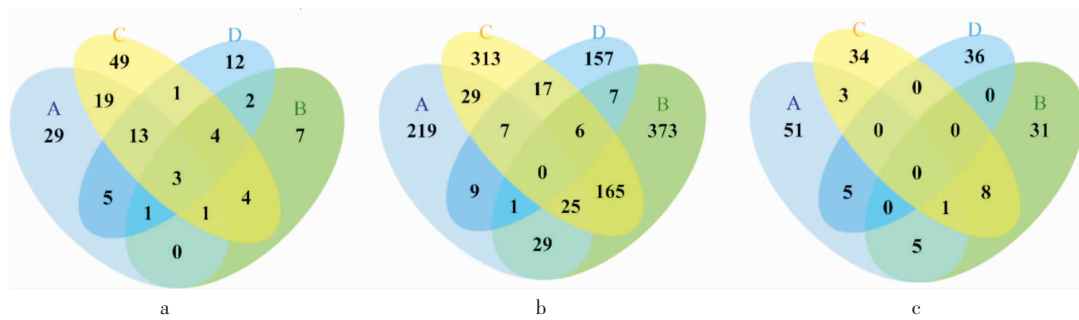
2.6 多组数据集的共同富集基因集和通路的筛选

对各组的差异基因分别进行 GO 功能 3 个方面的富集分析。由于杨嘉慧^[5]发表的差异基因数量过少,不适于进行基因富集和通路分析,对其进行剔除。选取 $P < 0.05$ 的剩余 4 组 GO 功能富集结果,并绘制韦恩图,见图 5。

在有关细胞组成的 GO 功能中可见有 3 个 GO 功能集合为 4 个组所共有。这 3 个功能集合分别为 GO:0060205、GO:0031983、GO:0034774;其功能分别为细胞质囊泡腔、囊泡腔、分泌颗粒腔。

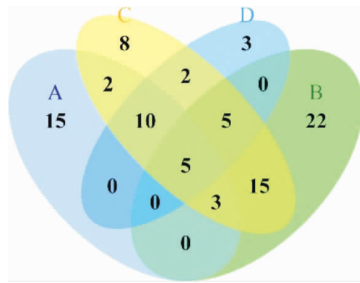
对剩余 4 个组的差异基因分别进行 KEGG 通路富集分析。选取 $P < 0.05$ 的剩余 4 组 KEGG 通路富集分析结果,并绘制韦恩图,见图 6。

图中可见有 5 个通路为 4 组所共有。这 5 个通路分别为 hsa04640、hsa05332、hsa05140、hsa04940、hsa05321;其功能分别为造血细胞系 (免疫功能相关)、移植物抗宿主病 (免疫功能相关)、利什曼病 (免



注:A.GSE87474;B. GSE56116;C.Cheng;D.汤朝晖+李艳艳;a.细胞组成;b.生物过程;c.分子功能

图5 不同阳虚人群 GO 富集分析韦恩图



注:A.GSE87474;B. GSE56116;C.Cheng;D.汤朝晖+李艳艳

图6 不同阳虚人群 KEGG 通路分析韦恩图

疫与炎症反应相关)、I型糖尿病、炎症性肠病(炎症反应相关)。

对超过3个组所共有的差异基因、GO富集基因集、KEGG富集通路进行统计。共有的差异基因数目为0,共有的GO富集集合(细胞组成)数为22个,共有的GO富集集合(生物过程)数为39个,共有的GO富集集合(分子功能)数为1个,共有的KEGG通路为23条。利用 χ^2 检验比较其间的差异,发现差异有统计学意义($P < 0.05$)。结果表明:不同阳虚人群之间更有可能也更容易存在共有的基因集和通路,而非单个差异基因。

3 讨论

中医阳虚概念与全基因组表达的关系目前尚不明确。本研究选取了GSE87474、GSE56116两组GEO数据集中阳虚体质与正常体质,绝经后骨质疏松症肾阳虚患者与无肾阳虚患者样本进行分析,将阳虚证候或体质作为变量,对其进行生物信息学分析,以探讨阳虚证候或体质与其基因表达的关系。本研究发现,阳虚人群与非阳虚人群的差异基因集和通路主要与能量代谢、细胞质囊泡及免疫调节相关,且不同阳虚人群间关键候选基因集和通路存在一定的相似性。各个纳入分析的差异基因集的通路分析结果具有较好的一致性。与相关研究相比较,研究结果亦具有一致性。在GO基因富集分析中,研

究团队选出了3个共同GO富集集合:“GO:0060205”“GO:0031983”“GO:0034774”,主要涉及细胞质囊泡。在KEGG基因通路分析中选出了5个共同基因通路:“hsa04640”“hsa04940”“hsa05240”“hsa05321”“hsa05332”,主要与免疫调节等相关。既往研究中,GSE87474与GSE56116数据集下尚无数据集上传者发表的论文,李艳艳^[4]报道了“hsa04640”“hsa04940”,汤朝晖^[3]报道了“hsa04940”,杨嘉慧^[5]报道了囊泡运输途径和“hsa04640”。其他相关研究中,TANG等^[8]利用四逆汤治疗肾阳虚证模型大鼠肾上腺基因表达变化寻找相关基因集与通路,发现四逆汤治疗肾阳虚证涉及能量代谢与线粒体功能相关通路。XU等^[9]利用GSE57273和GSE56116数据集探讨六味地黄丸治疗绝经后骨质疏松症肾阴虚证的靶点与基因通路,得到的差异基因、基因集和通路主要涉及免疫功能等方面。LIU等^[10]对气虚血瘀型和阴虚血瘀型缺血性脑卒中雄性大鼠模型的基因表达谱进行分析,发现两组间存在一定数量的相同差异基因,这些差异基因可能与气虚、血瘀或阴虚有关,功能主要集中在能量代谢方面,并利用免疫印迹法验证了富集基因和通路的产物。但其实验对象全部为大鼠,组间差异小,难以体现出人类人群间的多样性对实验结果的影响。GUAN等^[11]对肾阴虚大鼠基因表达情况进行分析,发现其差异基因功能亦主要集中于能量代谢等方面。此外,多项研究^[12-19]亦从中医药与针灸治疗角度分析了治疗前后基因表达变化,但对证型间的差异尚无涉及。尽管不同证候、不同物种的基因表达谱应有差异^[6],但本研究团队总结既往类似研究结果,综合本研究结果,认为阳虚人群与非阳虚人群间的基因表达差异可能多与能量代谢、免疫调节相关基因集和通路有关。

从差异基因的筛选结果来看,GSE56116共筛选出350个差异基因,GSE87474共筛选出138个

差异基因,疾病状态下的差异基因数量亦大于健康人群中的差异基因数量。本研究团队认为部分差异基因的表达差异可能由疾病或肾虚造成,从而导致筛选出差异基因数量增多。

对多组差异基因取交集后发现并无共同差异基因。而进行 GO 功能富集和 KEGG 基因通路分析后,发现其存在多个共同基因集和通路。见图 5-6。对共同差异基因或基因集合数量进行 χ^2 检验,发现差异有统计学意义($P<0.05$)。本研究团队认为阳虚人群在基因组表达层面更有可能与一些基因集和基因通路相关,而非单一的基因。

本研究的不足之处:(1)本数据集数据量偏少,且各个数据集使用的平台并不完全一致,纳入受试者的年龄、性别亦存在一定的差异。本研究的结果需要更大、更多元、相同平台的中医阳虚人群基因表达数据来进行证实。(2)纳入的两个 GEO 数据集存在一定的异质性,此异质性可能会对分析结果形成影响。

本研究表明:阳虚人群与非阳虚人群的差异基因集和通路主要与能量代谢、细胞质囊泡及免疫调节相关,不同阳虚人群间关键候选基因集和通路存在一定的相似性,且本研究的结果与既往相关研究具有较强一致性。疾病状态下的阳虚证候、体质人群的基因表达与健康人的阳虚体质基因表达存在一定的差异,今后开展类似研究可能需考虑此方面的差异。同时,不同阳虚人群的基因表达与基因集和通路的关系可能比单个基因更为密切,此观点亦需大数据层面的验证。最后,如何更好地从全基因组差异表达层面阐释阳虚概念乃至中医证候、体质、证型的形成机制,需在大数据支持下进一步探索。

参考文献

[1] 黄瑞聪,李美红,陈梦华,等.膏肓灸对阳虚质人人质量表积分及红外热图的影响研究[J].中医临床研究,2019,11(30):70-72.

[2] 曾召琼,易帆,李萍,等.基因组学技术在中医药研究中的应用[J].国际检验医学杂志,2018,39(24):3089-3092.

[3] 汤朝晖.老龄肾阳虚证候的诊断及其差异表达基因谱研究[D].成都:成都中医药大学,2008.

[4] 李艳艳.老年肾阳虚证的差异基因表达研究[D].成都:成都中医药大学,2009.

[5] 杨嘉慧.肾阳虚证排卵障碍性不孕的差异基因表达谱研究[D].成都:成都中医药大学,2012.

[6] 谭从娥.肾阳虚证膝骨关节炎的差异基因表达谱研究[D].成都:成都中医药大学,2006.

[7] CHENG H T, CHEN C R, LI C Y, et al. The classification of sini decoction pattern in traditional Chinese medicine by gene expression profiling[J]. Evidence-Based Complementary and Alternative Medicine: ECAM, 2016, 2016: 8239817.

[8] TANG N, LIU L H, QIU H, et al. Analysis of gene expression and functional changes of adrenal gland in a rat model of kidney Yang deficiency syndrome treated with Sini decoction[J]. Experimental and Therapeutic Medicine, 2018, 16 (4): 3107-3115.

[9] XU F, GAO F. Liuwei Dihuang pill cures postmenopausal osteoporosis with kidney-Yin deficiency[J]. Medicine, 2018, 97(31): e11659.

[10] LIU T L, LIU M N, XU X L, et al. Differential gene expression profiles between two subtypes of ischemic stroke with blood stasis syndromes[J]. Oncotarget, 2017, 8(67): 111608-111622.

[11] GUAN W, LIU Y, LI X M, et al. iTRAQ-based proteomics to reveal the mechanism of hypothalamus in kidney-Yin deficiency rats induced by levothyroxine[J]. Evidence-Based Complementary and Alternative Medicine, 2019, 2019: 3703596.

[12] 周 岚,李 杨,汪 典,等.利用基因芯片分析活血中药、破血中药对 ApoE 基因敲除小鼠动脉粥样硬化模型的差异表达基因[J].疑难病杂志,2017,16(1):18-22.

[13] 郑 玫,连 方,孙振高.补肾中药对肾虚不孕患者卵巢颗粒细胞基因表达干预的研究[J].辽宁中医杂志,2017,44(1):2-15,221.

[14] 魏歆然,魏高文,郑雪娜,等.不同经穴组合针刺对失眠大鼠下丘脑生物钟基因 Clock 和 Bmal 1 表达的影响[J].针刺研究,2017,42(5):429-433.

[15] 王 俊,虞彬艳,周斯斯,等.粗针神道穴平刺对周围性面瘫大鼠面神经基因表达谱的影响[J].中华中医药杂志,2016,31(1):287-291.

[16] 贾文睿.针刺对应激性高血压前期大鼠心脏基因表达谱的影响[D].北京:北京中医药大学,2017.

[17] WU P, HUANG R, XIONG Y L, et al. Protective effects of curcumin against liver fibrosis through modulating DNA methylation[J]. Chinese Journal of Natural Medicines, 2016, 14(4): 255-264.

[18] GUO Y, XIE X J, GUO C Q, et al. Effect of electro-acupuncture on gene expression in heart of rats with stress-induced pre-hypertension based on gene chip technology[J]. Journal of Traditional Chinese Medicine, 2015, 35(3): 285-294.

[19] HUANG Y L, WAN M Y, LIANG X S, et al. Effect of acupuncture along affected meridian on the MME gene expression of migraine patients without aura of Gan-Yang hyperactivity syndrome[J]. Chinese Journal of Integrated Traditional and Western Medicine, 2015, 35(3): 294-298.