棉花抗黄萎病共表达网络构建与关键基因筛选

朱宝祺，赵鹏举，吴洪俐，张赛男，刘震*

安阳工学院，安阳市生物信息学重点实验室，河南安阳 455000

摘要： 大丽轮枝菌引起的棉花黄萎病是一种危害非常严重的土传维管束病害，对棉花产业造成了很大的影响。加权基因共表达网络分析是一种利用基因在不同条件下的表达数据对基因进行聚类的数据挖掘方法，该方法可以将性状和基因进行关联，是一种广泛应用的系统生物学技术。本项目使用具有抗黄萎病特性的瑟伯氏棉、具有高抗黄萎病特性的三裂棉和对黄萎病敏感的雷蒙德棉的根、茎、叶感染大丽轮枝菌后的 RNAseq 数据，瑟伯氏棉、三裂棉和雷蒙德棉在冷、盐处理 12 小时候的转录数据，以及雷蒙德棉开花后 3 天胚珠发育的转录数据构建基因共表达网络并筛选核心基因。分析结果获得 1 个与抗黄萎病正相关的模块和 1 个与抗黄萎病负相关的模块。基因功能富集分析结果表明，抗黄萎病性状正相关模块基因主要涉及有机物分解代谢过程、细胞壁组织或生物合成、外部封装结构组织和激素水平的调控等代谢过程。本研究的结果将对棉花黄萎病抗性基因的筛选有重要参考价值。

关键词： 黄萎病；棉花; WGCNA；模块；聚类

DOI: 10.57237/j.life.2023.04.002

Construction of Cotton Verticillium wilt Resistance Co-expression Network and Identification of Hub Genes

Zhu Baoqi, Zhao Pengju, Wu Hongli, Zhang Sainan, Liu Zhen*

Anyang Key Laboratory of Bioinformatics, Anyang Institute of Technology, Anyang 455000, China

Abstract: Cotton verticillium wilt caused by *Verticillium dahliae* is a serious soil transmission vascular bundle disease, which has a great impact on the yield and quality of cotton. Weighted gene co-expression network analysis (WGCNA) is a data mining method that uses gene expression data under different conditions to cluster genes. This method can associate traits with genes, and it is a widely used system biology technology. The transcriptional data of *verticillium dahliae* infected root, stem and leaf of 0h, 12h, 48h of *Gossypium thurberi* with verticillium wilt resistance, *Gossypium trilobum* with high verticillium wilt resistance and *Gossypium raimondii* with verticillium wilt sensitivity were used in this project, transcriptional data of *Gossypium trilobum*, *Gossypium trilobum* and *Gossypium raimondii* under cold and salt treatment for 12 hours respectively and transcriptional data of ovule development 3 days after anthesis in *Gossypium raimondii* were used to construct the gene co

基金项目：大学生创新创业训练计划项目资助 (202311330014).

*通信作者：刘震, liuzhen378@163.com

收稿日期: 2023-10-02; 接受日期: 2023-11-09; 在线出版日期: 2023-11-15

expression network and select hub genes. The results showed that one module was positively correlated with verticillium wilt resistance and the other was negatively correlated with verticillium wilt resistance. The results of gene function enrichment analysis showed that the positive correlation module genes of verticillium wilt resistance mainly involved in organic matter catabolism, cell wall tissue or biosynthesis, external encapsulation structure and hormone level regulation. The results of this study have important value to find verticillium wilt resistance genes in cotton.

Keywords: Vrticillium Wilt; Cotton; WGCNA; Modular; Cluster

1 引言

棉花作为纺织、医药和国防工业的重要原料，在中国国民经济中占有重要地位。然而，棉花生产过程中常因受到各种胁迫而导致减产和品质下降，其中，黄萎病是对棉花生产具有毁灭性的病害之一，直至目前仍然没有防治黄萎病的理想方法。引起棉花黄萎病的致病菌在分类上属于轮枝菌属，该菌属包含约十个种，对棉花及其它农作物危害最为严重的是黑白轮枝菌（*Verticillium alboatrum*）和大丽轮枝菌（*Verticillium dahliae*）[1, 2]。大丽轮枝菌的寄主非常广泛，目前报道的寄主有锦葵科、蔷薇科、茄科、十字花科和豆科等双子叶植物[3, 4]。

棉属有五十多个种，除 7 个四倍体种外，其余均为二倍体种，二倍体棉有 13 条染色体。二倍体棉种的染色体可分为 A、B、C、D、E、F、G 和 K 共 8 种类型[5]。二倍体的亚洲棉（*Gossypium arboreum*，A 组）、草棉（*Gossypium herbaceum*，A 组），四倍体的陆地棉（*Gossypium hirsutum*，AD 组）、海岛棉（*Gossypium barabdense*，AD 组）为栽培棉种，能够产生较好的纤维，其余均为野生棉种。随着测序和组装技术的不断发展，目前，二倍体亚洲棉、草棉、雷蒙德氏棉（*Gossypium raimondii*，D 组）、澳洲棉（*Gossypium australe*，G 组），四倍体陆地棉、海岛棉、黄褐棉（*Gossypium mustelinum*，AD 组）、夏威夷棉（*Gossypium tomentosum*，AD 组）和达尔文棉（*Gossypium darwinii*，AD 组）的基因组已经公布[6~11]。同时，NCBI SRA 数据库和 NCBI GEO 数据库也公布了棉属大量的表达数据，这些都为开展棉属的生物信息学分析提供了绝好的机遇。

栽培棉种经过长期的人工定向选育，在抗病虫害方面存在缺陷，大丽轮枝菌对 4 种栽培棉均有严重的危害。相比之下，野生棉种长期受到自然选择的压力，保留了非常丰富的遗传多样性及优良性状，如抗寒、抗旱、耐盐碱和抗病虫害等，为棉花的遗传改良提供了宝贵资源。二十世纪末，中国科研人员开始关注野生棉在抗黄萎病方面的应用，1993 年顾本康等鉴定了 17 个野生棉对黄萎病的抗性，其中 10 个为免疫、4 个为耐病、3 个为高感[12]。1998 年，梁理民等通过实验分析认为瑟伯氏棉（*Gossypium thurberi*）具有抗黄萎病的特性[13]。此外，赵凤轩等通过对 8 个野生棉的抗病类型进行鉴定分析，认为戴维逊棉（*Gossypium davidsonii*）为感病棉种；瑟伯氏棉、索马里棉（*Gossypium somalense*）、旱地棉（*Gossypium aridum*）、异常棉（*Gossypium anomalum*）、长萼棉（*Gossypium longicalyx*）和澳洲棉为抗病棉种；三裂棉（*Gossypium trilobum*）则为高抗棉种[14]。

加权基因共表达网络分析（Weighted Gene Co-expression Network Analysis，WGCNA）是一种以整体角度进行分析的系统生物学技术，它能够将具有相似表达的基因归类到一个模块，将模块与生物性状进行关联，并从中挖掘出关联度高的关键基因[15]。WGCNA 从整体的角度，而不是关注单个基因，因此更能反映生物分子通过相互作用网络实现生物学功能的情况，同时，这种技术为筛选性状相关基因提供了一种解决方案。

棉花是中国重要的经济作物，黄萎病严重影响了棉花的产量和品质。防治棉花黄萎病的方法主要包括抗性育种、化学防治和生物防治等。目前，抗黄萎病棉花品种的选育虽然取得一定进展，但多数品种仍表现为低耐或感病水平，不能满足生产需求；化学防治方面，主要依赖化学药剂，对环境破坏很大；生物防治主要通过微生物及微生物代谢产物的作用对农作物病虫害进行防治，这种方法目前也不能很好地解决黄萎病的危害。本研究选取对黄萎病具有较好抗性的野生品种和对黄萎病敏感的品种，以黄萎病致病菌侵染不同棉花组织的基因表达数据为基础，筛选表达差异较大基因进行 WGCNA 分析，筛选与棉花黄萎病相关的基因模块，对模块中的基因进行功能富集分析，并进一步从模块中筛选关键基因，从而为进一步研究棉花抗黄萎病提供数据支撑。

2 材料与方法

2.1 实验数据

瑟伯氏棉、三裂棉和雷蒙德棉根、茎、叶感染大丽轮枝菌以及冷、盐处理的转录数据从 NCBI SRA 数据库（https://www.ncbi.nlm.nih.gov/sra）下载，数据编号分别为 PRJNA507768、PRJNA554555、PRJNA79005、SRP166107 和 PRJNA321738。

2.2 差异表达分析

通过 Trimmomatic [16]软件去除转录数据两边的接头和低质量序列。通过 Hisat2 [17]软件将 read 序列比对到雷蒙德棉基因组，再使用 Cufflinks [18, 19]软件计算 FPKM。利用 R 语言的 Rsubread 包[20]和 edgeR 包[21]筛选差异表达基因。由于瑟伯氏棉、三裂棉和雷蒙德氏棉均为二倍体 D 基因组野生棉，具有相似的基因组序列，因此，在本研究中，雷蒙德氏棉的基因组序列将被用作参考基因组计算基因的表达数据。

2.3 共表达网络构建

利用 R 语言的 WGCNA 包[22]进行加权基因共表达网络分析。为使共表达网络符合无尺度分布，利用 pickSoftThreshold 函数确定最佳软阈值 β=7，通过 cutreeDynamic 函数实现动态剪切树划分基因模块，设定模块内最少的基因数量为 30，并将相关性大于 0.75 的两个模块合并为一个模块。

2.4 网络可视化及关键基因筛选

通过 Cytoscape [23]实现相互作用网络的可视化及关键基因的筛选。在 Cytoscape 的控制面板中设置节点、边和网络的格式，设计网络布局，并通过 cytoHubb a [24]插件挑选模块的 hub 基因。

2.5 模块功能分析

依据相关系数筛选抗黄萎病相关性最高模块，通过 R 语言的 AnnotationHub 包和 clusterProfiler 包[25, 26]进行 GO（Gene Ontology）和 KEGG（Kyoto Encyclopedia of Genes and Genomes）分析，找出模块中显著富集的生物学功能和代谢通路。

3 结果与分析

3.1 基因共表达网络构建

本研究使用了黄萎病、盐、冷以及胚珠发育相关研究项目的转录组数据。通过动态剪切法划分基因模块，并将相似性高的模块合并为一个模块，最后得到 16 个模块，这些模块用不同的颜色命名（图 1，图 2）。在筛选到网络模块之后，我们将棉花的表型特征与模块进行了关联分析。从结果可以看出，抗黄萎病（Res_Verticillium）性状与 cyan 模块为正相关关系，黄萎病敏感性状与 blue 模块为正相关关系（图 2）。

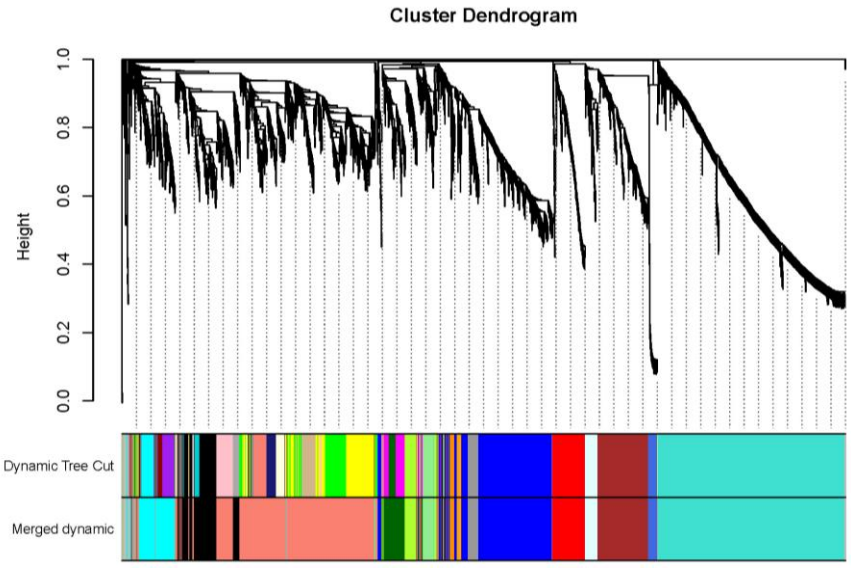

图 1 基因聚类树及模块划分。

Cluster Dendrogram：基于拓扑相异矩阵构建的基因聚类树。Dynamic Tree Cut：使用动态剪切算法得到的基因模块，不同颜色代表不同模块。Merged Dynamic：合并相似模块后的模块划分结果。

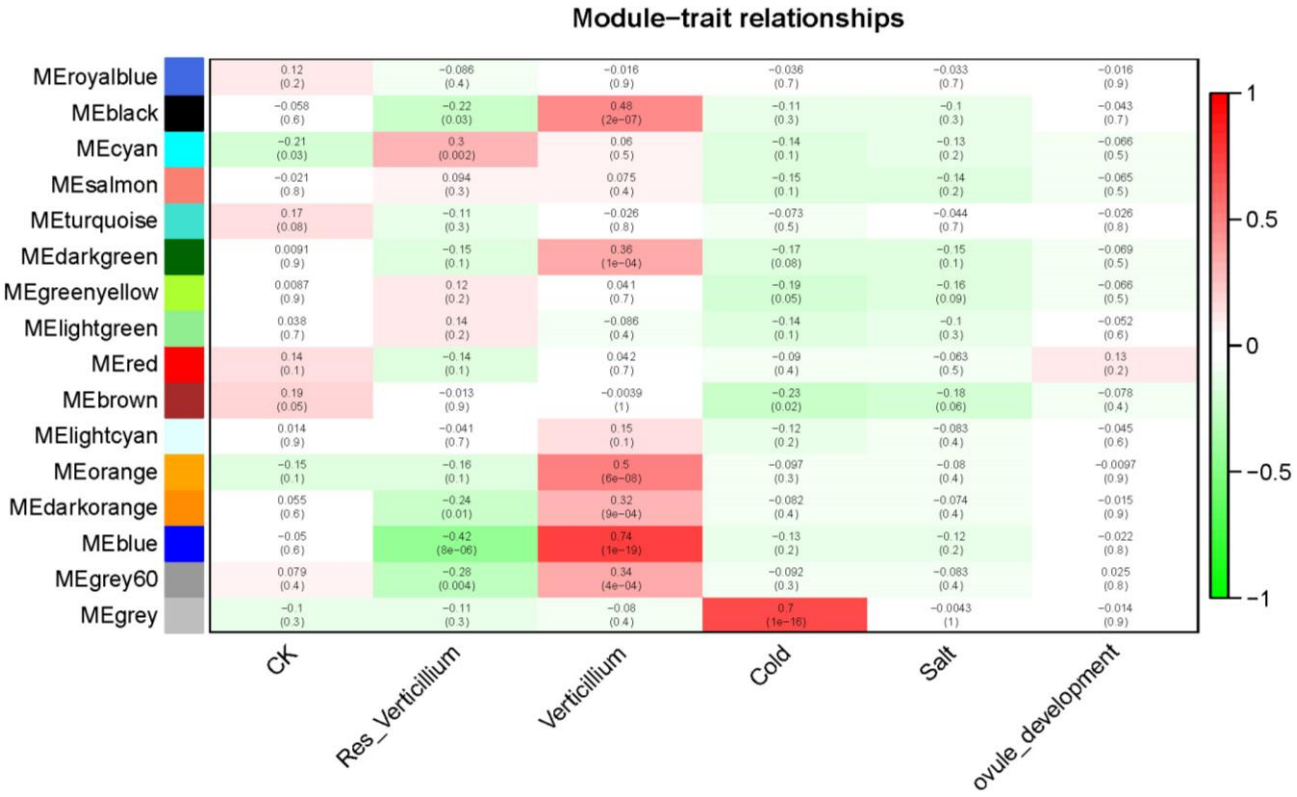

图 2 模块与性状相关性热图

每行代表一个模块，每列代表一种性状。矩形框里的数字代表模块与性状之间的相关系数及相应的 p 值。

3.2 模块分析

Cyan 模块与抗黄萎病性状之间为正相关，该模块的 267 个基因形成 30364 个相互作用关系对，从这些相互作用关系中选择权重最高的 300 对，构建相互作用网络（图 3），并从网络中选出最核心的 10 个基因（表 1）。LOC105778767（蛋白：XP_012457982.1）为 MYB 转录因子，该转录因子是植物重要的转录因子家族之一，其结构域包含 52 个氨基酸，参与植物非生物胁迫反应、外界环境因素响应、激素应答以及多种次生代谢等[27, 28]。LOC105783379（XP_012464255.1）包含一个植物必须膜蛋白结构域 Mlo，Mlo 有 7 个跨膜区，位于细胞质膜。有研究表明，缺失 Mlo 的突变体会对细胞死亡控制失调，对非生物刺激有自发的细胞死亡反应[29]。

表 1 Cyan 模块的核心基因

序号	基因	蛋白
1	LOC105779263	XP_012458483.1
2	LOC105778767	XP_012457982.1
3	LOC105767770	XP_012442791.1
4	LOC105767657	XP_012442681.1
5	LOC105783379	XP_012464255.1
6	LOC105782454	XP_012462654.1
7	LOC105770016	XP_012446501.1
8	LOC105770279	XP_012446858.1
9	LOC105793839	XP_012478152.1
10	LOC105762693	XP_012435984.1

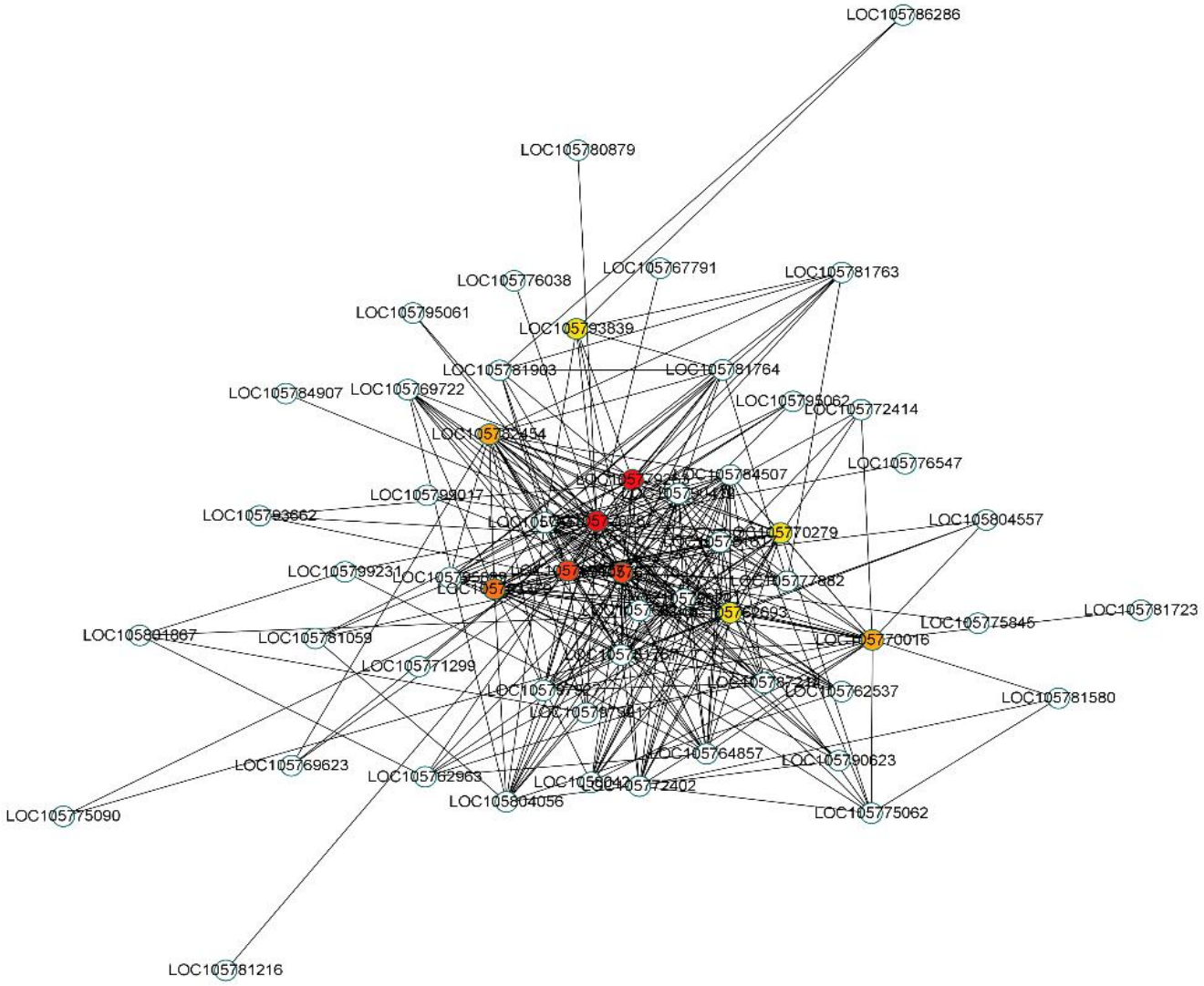

图 3 Cyan 模块的分子相互作用网络

最核心的 10 个基因为彩色

表 2 Blue 模块的核心基因

序号	基因	蛋白
1	LOC105789099	XR_008193763.1
2	LOC105779228	XR_001128923.2
3	LOC105771376	XR_001126485.2
4	LOC105787212	XR_001131481.2
5	LOC105790826	XP_012474067.1
6	LOC105763983	XR_001124495.1
7	LOC105772236	XR_001126655.2
8	LOC105798547	XP_012484120.1
9	LOC105771733	XP_012448588.1
10	LOC105784196	uncharacterized

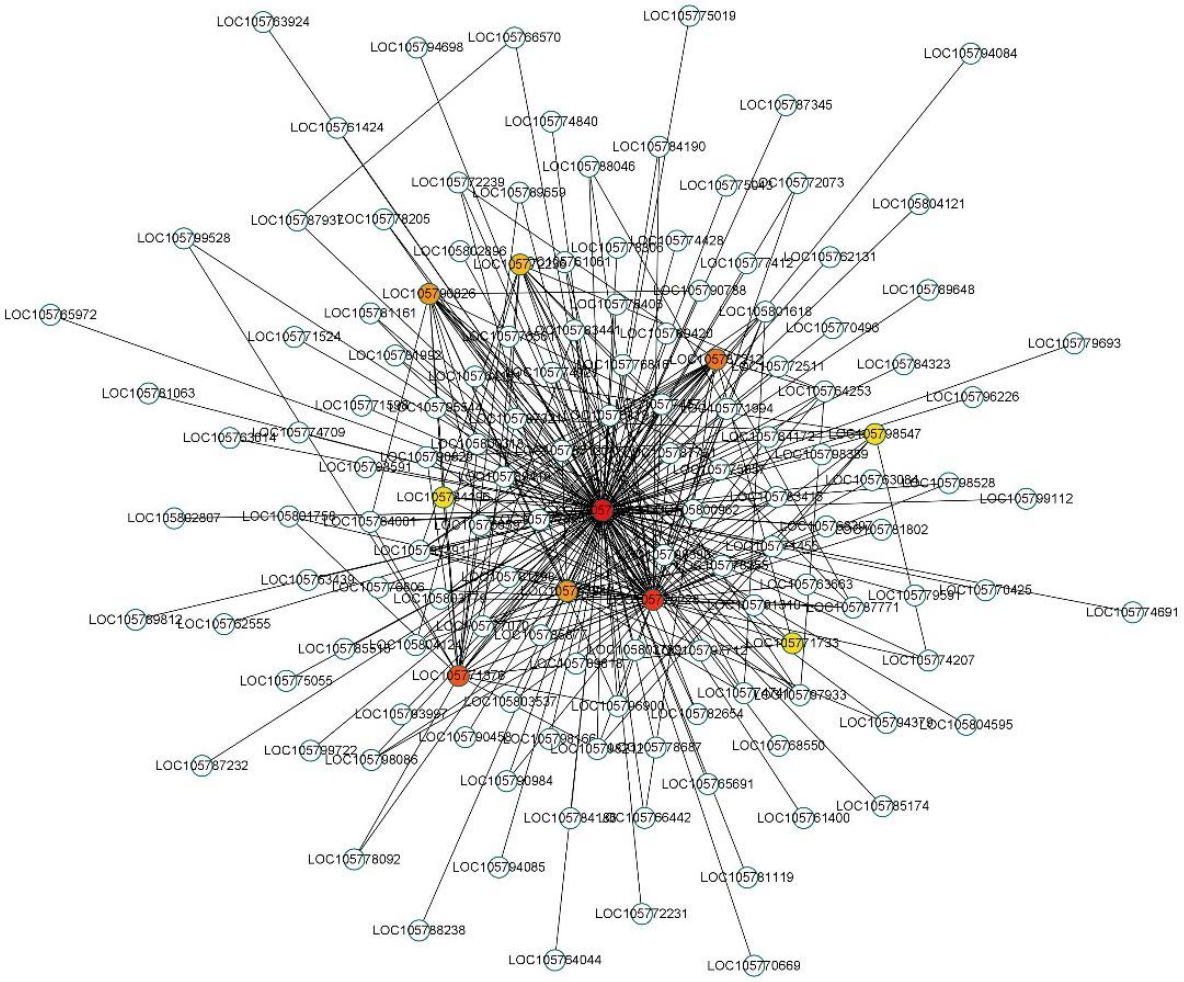

图 4 Blue 模块的分子相互作用网络

最核心的 10 个基因为彩色

Blue 模块与抗黄萎病性状之间为负相关。Blue 模块的 645 个基因形成 182549 个相互作用关系对，从这些相互作用关系中选择权重最高的 300 对，构建相互作用网络（图 4），并从中网络中选出最核心的 10 个基因（表 2）。Blast 搜索结果表明 LOC105790826（XP_012474067.1）是一个 DNA 修复蛋白。

3.3 富集分析

Cyan 模块显著富集到有机物分解代谢过程（organic substance catabolic process）、外部封装结构组织（external encapsulating structure organization）、细胞壁组织或生物合成（cell wall organization or biogenesis）、激素水平的调控（regulation of hormone levels）、细胞激素代谢过程（hormone metabolic process）和多元醇分解代谢过程（polyol catabolic process）等（图 5）。

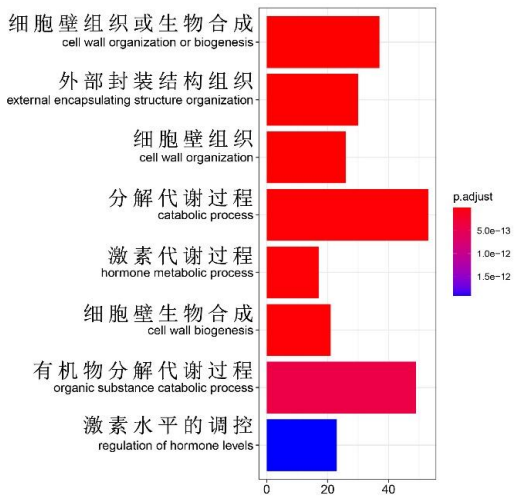

图 5 Cyan 模块的 GO 富集分析

Cyan 模块的 KEGG 富集主要涉及苯丙烷生物合成

（Phenylpropanoid biosynthesis）、植物激素信号转导（Plant hormone signal transduction）、倍半萜和三萜生物合成（Sesquiterpenoid and triterpenoid biosynthesis）、和 MAPK 信号通路（MAPK signaling pathway）等代谢途径（图 6）。

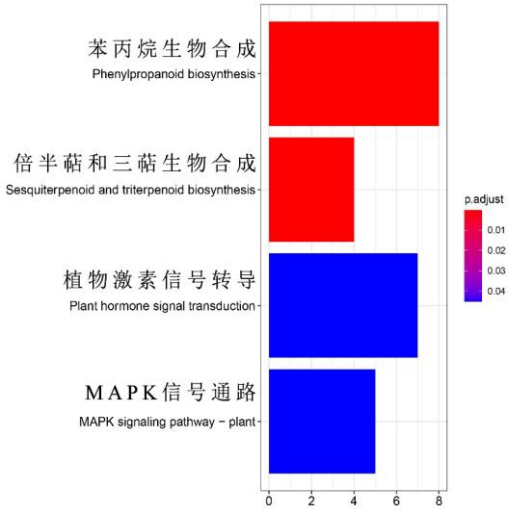

图 6 Cyan 模块的 KEGG 富集分析

Blue 模块的富集分析显著富集到防御反应（defense response）、生殖结构发育（reproductive structure development）、生殖系统发育（reproductive system development）、对激素反应（response to hormone）、对内源性刺激反应（response to endogenous stimulus）、细胞发育过程（cellular developmental process）、生殖芽系统发育（reproductive shoot system development）和细胞分化（cell differentiation）等（图 7），然而，Blue 模块没有显著富集到任何 KEGG 代谢通路。

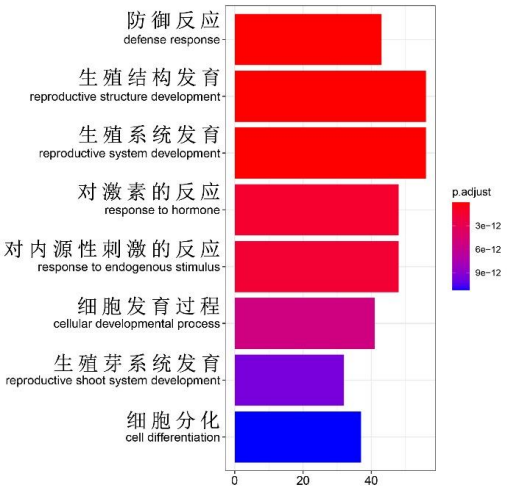

图 7 Blue 模块的 GO 富集分析

4 结论

黄萎病对棉花生产造成了严重影响，随着测序技术的发展，公共数据库积累了大量的基因表达数据，WGCNA 技术可以利用基因在不同生理状态下的表达数据，将基因进行聚类，并将聚类的基因模块与性状相关联，这种方法已被证明是一种高效的数据挖掘方法，为筛选棉花抗黄萎病相关基因提供了新的方案。

本项目筛选到了 1 个与抗黄萎病性状正相关的模块（Cyan）和 1 个与抗黄萎病性状负相关的模块（Blue），通过 GO 和 KEGG 分析，大致了解了这些模块涉及的生物学功能和参与的代谢过程。此外，本研究也依据模块节点的关联度从模块中筛选了关键基因。这些结果将为棉花抗黄萎病相关研究提供数据支撑。

参考文献

[1] 朱荷琴, 冯自力, 李志芳, 等. 分离自棉花的轮枝菌“种”的鉴定[J]. 中国农业科学, 2013, 46(10): 2032-2040.

[2] 张绪振, 张树琴, 陈吉棣, 等. 我国棉花黄萎病菌“种”的鉴定 [J]. 植物病理学报, 1981(03): 15-20.

[3] 丁晓华. 棉花枯、黄萎病的症状及防治方法 [J]. 河北农业, 2018(07): 30-31.

[4] 伊静. 棉花黄萎病的发生与防治技术 [J]. 现代农业研究, 2018(06): 87-88.

[5] 王坤波, 刘旭. 棉属多倍化研究进展 [J]. 中国农业科技导报, 2013, 15(02): 20-27.

[6] Cai Y, Cai X, Wang Q, et al. Genome sequencing of the Australian wild diploid species *Gossypium australe* highlights disease resistance and delayed gland morphogenesis [J]. Plant Biotechnol J, 2020, 18(3): 814-828.

[7] Wang K, Wang Z, Li F, et al. The draft genome of a diploid cotton *Gossypium raimondii* [J]. Nat Genet, 2012, 44(10): 1098-1103.

[8] Hu Y, Chen J, Fang L, et al. *Gossypium barbadense* and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton [J]. Nat Genet, 2019, 51(4): 739-748.

[9] Li F, Fan G, Lu C, et al. Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution [J]. Nat Biotechnol, 2015, 33(5): 524-530.

[10] Chen Z J, Sreedasyam A, Ando A, et al. Genomic diversifications of five *Gossypium allopolyploid* species and their impact on cotton improvement [J]. Nat Genet, 2020, 52(5): 525-533.

[11] Li F, Fan G, Wang K, et al. Genome sequence of the cultivated cotton Gossypium arboretum [J]. Nat Genet, 2014, 46(6): 567-572.

[12] 顾本康, 李经仪, 顾萍, 等. 棉属野生种枯萎病黄萎病抗性鉴定初报 [J]. 江苏农业科学, 1993(05): 36-37.

[13] 梁理民, 刘有良, 王增信, 等. 陆地棉×斯特提棉种间杂交创造抗枯黄萎病新种质 [J]. 西北农业学报, 2002(04): 16-18.

[14] 赵凤轩, 戴小枫. 棉花黄萎病菌的侵染过程 [J]. 基因组学与应用生物学, 2009, 28(04): 786-792.

[15] Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis [J]. Stat Appl Genet Mol Biol, 2005, 4: e17.

[16] Bolger A M, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data [J]. Bioinformatics, 2014, 30(15): 2114-2120.

[17] Kim D, Langmead B, Salzberg S L. HISAT: a fast spliced aligner with low memory requirements [J]. Nat Methods, 2015, 12(4): 357-360.

[18] Ghosh S, Chan C K. Analysis of RNA-Seq Data Using TopHat and Cufflinks [J]. Methods Mol Biol, 2016, 1374: 339-361.

[19] Pollier J, Rombauts S, Goossens A. Analysis of RNA-Seq data with TopHat and Cufflinks for genome-wide expression analysis of jasmonate-treated plants and plant cultures [J]. Methods Mol Biol, 2013, 1011: 305-315.

[20] Liao Y, Smyth G K, Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads [J]. Nucleic Acids Res, 2019, 47(8): e47.

[21] Robinson M D, McCarthy D J, Smyth G K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data [J]. Bioinformatics, 2010, 26(1): 139-140.

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

[23] Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks [J]. Genome Res, 2003, 13(11): 2498-2504.

[24] Chin C H, Chen S H, Wu H H, et al. cytoHubba: identifying hub objects and sub-networks from complex interactome [J]. BMC Syst Biol, 2014, 8 Suppl 4(Suppl 4): S11.

[25] Wu T, Hu E, Xu S, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data [J]. Innovation (Camb), 2021, 2(3): 100141.

[26] Yu G, Wang L G, Han Y, et al. clusterProfiler: an R package for comparing biological themes among gene clusters [J]. OMICS, 2012, 16(5): 284-287.

[27] Du H, Zhang L, Liu L, et al. Biochemical and molecular characterization of plant MYB transcription factor family [J]. Biochemistry (Mosc), 2009, 74(1): 1-11.

[28] Millard P S, Kragelund B B, Burow M. R2R3 MYB Transcription Factors - Functions outside the DNA-Binding Domain [J]. Trends Plant Sci, 2019, 24(10): 934-946.

[29] Devoto A, Piffanelli P, Nilsson I, et al. Topology, subcellular localization, and sequence diversity of the Mlo family in plants [J]. J Biol Chem, 1999, 274(49): 34993-35004.