许　哲,张晨光,张思远,等. 二倍体和四倍体西瓜的转录组初步分析[J]. 江苏农业科学,2020,48(24):53 – 59.
doi:10.15889/j. issn. 1002 – 1302. 2020. 24. 010

二倍体和四倍体西瓜的转录组初步分析

许　哲[1], 张晨光[2], 张思远[2], 王先裕[1], 高建昌[2]
(1. 广西大学,广西南宁 530001; 2. 中国农业科学院蔬菜花卉研究所,北京 100081)

摘要:植物染色体加倍后的表型变化与基因的表达密切相关。本研究对 1 份二倍体和 3 份四倍体西瓜材料进行转录组测序并分析差异表达基因。结果表明,二倍体和四倍体西瓜的差异表达基因显著富集的 GO term 为细胞氮化合物生物合成过程、光系统、半胱氨酸型内肽酶抑制剂活性、氧化还原酶活性以及类囊体等,KEGG 富集分析中发现差异表达基因在光合作用、光合作用天线蛋白、叶绿素和卟啉代谢、乙醛酸和二羧酸代谢等通路较为活跃。二倍体和四倍体西瓜中高水平差异基因大都是上调表达,氧化应激蛋白、脱落酸受体、乙烯响应转录因子等差异表达基因可能是西瓜多倍体形成的关键因子,对相关基因的深入研究有助于解析植物倍性变化的分子机制。

关键词:多倍体;西瓜;转录组;差异表达基因;GO 富集分析;KEGG 富集分析
中图分类号: S651.01　　**文献标志码**: A　　**文章编号**:1002 – 1302(2020)24 – 0053 – 07

西瓜(*Citrullus lanatus*)原产于非洲,属于葫芦科西瓜属,为一年生蔓生藤本植物,是世界上重要的园艺作物。我国是世界上西瓜生产与消费大国,2016 年西瓜播种面积为 189.08 万 hm^2,总产量为 7 940 万 t[1]。

四倍体西瓜是普通二倍体西瓜染色体加倍的结果,在田间其最明显的特征是巨大性,叶片大且肥厚,茎粗且节间短,叶色浓绿,花和果实也较为巨大,综合抗病能力较强。染色体加倍获得四倍体是选育三倍体西瓜(无籽西瓜)的基础。我国从 20 世纪 50 年代开始对多倍体西瓜诱导方法的研究,如今已建立秋水仙素诱导、体细胞杂交、胚乳培养等多种四倍体诱导方法;而且建立了染色体计数、叶肉细胞分析、流式细胞仪计数等染色体倍性检测方法。利用上述方法选育出很多无籽西瓜品种,为我国西瓜产业的发展做出了巨大贡献。但是,对四倍体西瓜表型变化与染色体倍性的关系还缺乏分子层面的认知。

转录组是研究细胞表型和功能的一个重要手段,是指某一特定的生理条件下,细胞、组织或生物体内所有转录产物的集合,即转录后所有 mRNA 的总称[2]。对二倍体和四倍体西瓜进行转录组水平的分析,为解析四倍体表型的成因提供了新的手段。自 2012 年西瓜全基因组测序完成后,西瓜转录组分析亦成为研究热点。Wechter 等利用 832 个表达序列标签研究了果实发育过程中的基因表达情况[3];Guo 等在西瓜果实发育和成熟过程中鉴定出 3 023 个差异表达基因[4];Zhu 等通过比较 2 个红色和黄色果肉的栽培品种,确定了 797 个新基因[5];龙娅丽等通过研究二倍体西瓜及其同源四倍体叶片 sRNA 表达谱,提供了多倍体抗逆性强的理论依据[6]。

多倍体化是植物进化中的一种普遍现象,也是新物种形成的重要途径[7]。目前研究者多集中在对二倍体西瓜的研究,关于二倍体西瓜与四倍体西瓜转录组差异的研究相对较少。本研究利用 Illumina 二代高通量测序平台对二倍体西瓜以及 3 份四倍体西瓜叶片进行转录组测序分析,比较二倍体西瓜和四倍体西瓜中基因表达的差异,以期通过转录组测序(RNA – Seq)技术获得与倍性相关的基因序列,为揭示西瓜的倍性机制提供参考。

1　材料与方法

1.1　试验材料

试验材料为 1 份二倍体,记为 WMA,3 份四倍体,记为 WMB、WMC 和 WMD。WMB 为 WMA 染色

收稿日期:2020 – 03 – 18
基金项目:中国农业科学院科技工程创新项目(编号:CAAS – ASTIPIVFCAAS);农业部园艺作物生物学与种质创制重点实验室项目。
作者简介:许　哲(1994—),男,广西南宁人,硕士研究生,主要从事西瓜遗传育种研究。E – mail:369474330@ qq. com。
通信作者:高建昌,博士,副研究员,主要从事西瓜遗传育种研究。E – mail:gaojianchang@ caas. cn。

体加倍获得，为了增加样品的遗传多样性，加入另外 2 份不同来源的四倍体材料 WMC、WMD，材料均来自中国农业科学院蔬菜花卉研究所西瓜育种课题组。本试验未设生物学重复。试验地点为中国农业科学院蔬菜花卉研究所日光温室，2019 年 6 月 1 日将西瓜种子播种于穴盘中，20 d 后长出真叶时取 2 cm^2 大小新鲜叶片，立即放入液氮中保存备用。

1.2　总 RNA 的提取及检测

使用天根生化科技（北京）有限公司 RNA 试剂盒对试验材料进行总 RNA 提取，后通过 Nanodrop 检测 RNA 纯度（$D_{260\ nm/280\ nm}$、$D_{260\ nm/230\ nm}$），通过 Agilent 2100 对 RNA 片段长度进行检测。

1.3　cDNA 文库构建和测序

样品检测合格后，使用带有 Oligo（dT）的磁珠富集 mRNA，之后加入破碎缓冲液将 mRNA 打断成短片段。再用六碱基随机引物（random hexamers）以 mRNA 为模板进行反转录合成一链 cDNA，加入缓冲液、dNTPs 和 DNA 聚合酶Ⅰ合成二链 cDNA。接着，利用 AMPure XP beads 纯化双链 cDNA。对纯化后的双链 cDNA 进行末端修复、加 A（特异碱基）、加接头。通过 AMPure XP beads 核酸纯化试剂盒对双链 cDNA 进行片段大小选择，最后进行 PCR 扩增以构建 cDNA 文库。

文库构建完成后用 Agilent 2100 对文库的插入片段大小进行检测，当片段大小符合预期时，再使用定量 PCR（Q－PCR）方法对文库的有效浓度进行精确定量（文库有效浓度 >4 nmol/L），以达到高质量文库标准，最后通过 Illumina 二代高通量测序平台，采用 PE150 双末端测序策略，完成 RNA－Seq。

1.4　数据质控与基因功能注释

下机数据（raw data）通常会含有少量的接头污染及低质量的 Reads，如果不对其进行过滤处理会对后续分析造成影响，为此我们过滤掉了带有测序接头 adapter 的 Reads、N（不确定碱基）含量比例大于 10% 的 Reads 以及低质量碱基（$Q \leqslant 20$）含量大于 50% 的 Reads。使用 TopHat2 软件将数据过滤后的 Clean Reads 与西瓜参考基因组[8]进行序列比对，之后用 DEGseq[9]进行差异表达分析，再分别与基因本体（GO）、京都基因与基因组百科全书（KEGG）数据库比对，获得相关注释信息。

2　结果与分析

2.1　转录组测序数据质量评估

将二倍体西瓜与四倍体西瓜叶片 RNA 样本进行测序，结果见表 1。二倍体 WMA 和四倍体 WMB、WMC 以及 WMD 有效数据均占原始序列数据的 95.58% 以上，Q20（碱基识别错误率为 1%）碱基百分比均大于 97.87%，Q30（碱基识别错误率为 0.1%）碱基百分比均大于 94.52%，GC 含量均大于 43.40%，能比对到参考序列的 reads 百分数均大于 95.27%，说明测序结果有较高的准确率。

表 1　西瓜样品测序数据评估

样品名称	原始序列数据	有效数据	有效数据占原始数据比例（%）	Q20 碱基百分比（%）	Q30 碱基百分比（%）	GC 含量（%）	比对到参考序列 reads 百分数（%）
WMA	30 645 874	29 398 816	95.93	98.03	94.80	44.80	96.18
WMB	30 843 125	29 479 719	95.58	97.87	94.52	43.64	95.43
WMC	33 292 868	31 819 820	95.58	97.95	94.66	44.39	95.7
WMD	34 011 069	32 566 348	95.75	97.88	94.58	43.40	95.27

2.2　RNA－Seq 相关性检查

为了检测转录组测序的相关性，对每个样品进行相关性分析。相关性分析是基于样品间基因整体的表达水平做的皮尔逊相关系数分析，相关系数越接近 1，表明样品之间表达模式的相似度越高。Encode 计划建议皮尔逊相关系数的平方（r^2）大于 0.92（理想的取样和试验条件下）。而实际项目操作中，要求 r^2 至少要大于 0.8，否则须要对样品做出合理的解释，或重新进行试验。本次试验中 WMA、WMB、WMC 和 WMD 4 个样品之间相关系数 r^2 均大于 0.899，最高达到了 0.958（图 1），表明样品间相关性较好。

2.3　不同倍性西瓜差异表达基因分析

2.3.1　二倍体 WMA 与四倍体 WMB、WMC 和 WMD 差异表达基因筛选　对于无生物学重复的试验，为避免引入试验误差，应该对结果进行严格控制，对差异基因进行筛选的阈值一般为：$|\log_2$（差异倍数）$|>1$ 且 q 值 <0.005[10]。结果（图 2）表明，在 WMA 和 WMB 中共检测到 1 458 条差异表达基因，其中表达上调的基因有 909 条，表达下调的基因

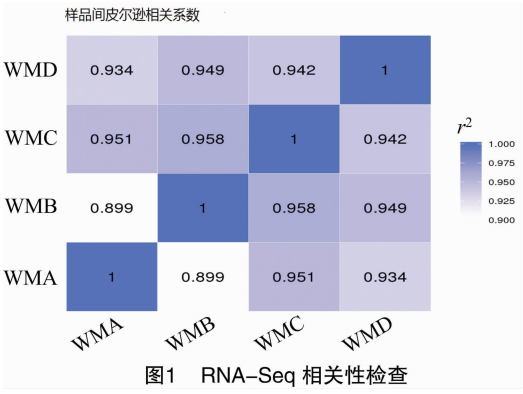
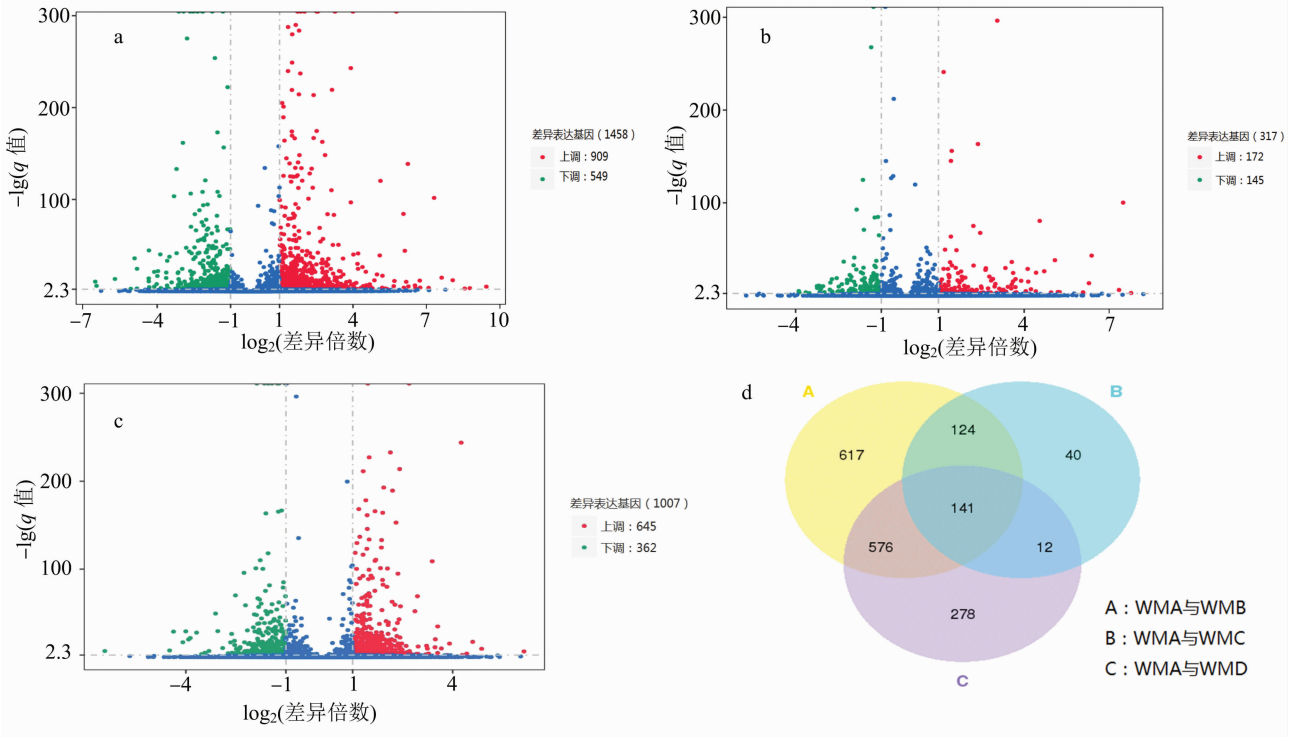


图1　RNA-Seq 相关性检查

有 549 条;在 WMA 和 WMC 中共检测到 317 条差异表达基因,其中表达上调的基因有 172 条,表达下调的基因有 145 条;在 WMA 和 WMD 中共检测到 1 007 条差异表达基因,其中表达上调的基因有 645 条,表达下调的基因有 362 条。不同组间差异基因维恩图显示,有 141 个不同组间的共同差异表达基因。

2.3.2　差异基因的聚类分析　从图 3 可以看出,WMB 和 WMD 高表达量和低表达量基因模式大体相似。WMA、WMB、WMC 和 WMD 的差异表达聚合



a—WMA和WMB的差异基因分析;b—WMA和WMC的差异基因分析;c—WMA和WMD的差异基因分析;d—差异基因维恩图

图2　不同组合间基因差异表达分析和维恩图

为 5 类。A 组中,WMA 和 WMD 基因表达量以上调为主,WMB 和 WMC 基因表达量以下调为主;B 组中,WMA 基因表达量呈下调趋势,WMB、WMC 和 WMD 基因表达量以下调为主;C 组中 WMA 和 WMC 基因表达量以上调为主,WMB 和 WMD 基因表达量以下调为主;D 组中,WMA 和 WMC 基因表达量以下调为主,WMB 和 WMD 基因表达量以上调为主;E 组中,WMA、WMB、WMC 和 WMD 基因表达量各不一致。二倍体 WMA 和四倍体 WMB、WMD 之间的差异表达基因主要分布在 C 和 D 组中,这部分可能调控西瓜倍性的差异基因。

2.3.3　差异基因 GO 富集分析　GO 是一个用于描述生物体中基因和蛋白质的功能分类体系。GO 分为分子功能(molecular function)、生物过程(biological process)、和细胞组成(cellular component)等 3 个方面。GO 的基本单位为 term,每个 term 对应一个功能或属性。在分析中采用的软件为 GOseq[11],前 30 个富集最显著的 GO term 见图 4,如果不足 30 个,则全部展示。结果表明,WMA 和 WMB 之间富集的差异表达基因与生物过程、细胞组成和分子功能相关的分别有 21、3、6 个 GO term。说明大部分差异基因与生物过程相关,其中与生物过程相关富集最显著、基因数量最多的 term 为小分子代谢过程(small molecule metabolic process)和细胞氮化合物生物合成过程(cellular nitrogen compound biosynthetic process);与细胞组成相关,富

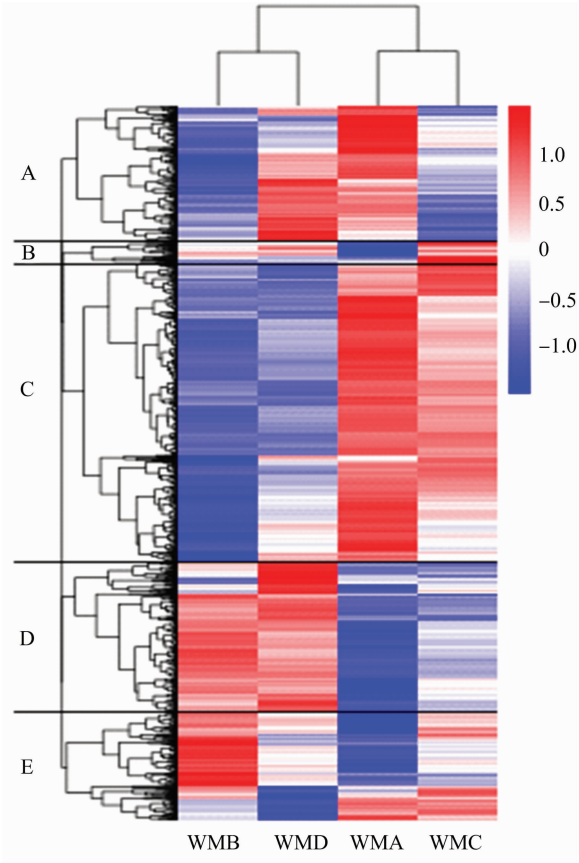


图3　不同倍性西瓜叶片中基因表达层次聚类

集最显著的 term 为光系统Ⅰ(photosystem Ⅰ),富集基因数量最多的类别为光系统(photosystem);与分子功能相关富集最显著的 term 为半胱氨酸型内肽酶抑制剂活性(cysteine-type endopeptidase inhibitor activity),富集基因数量最多的 term 为氧化还原酶活性(oxidoreductase activity)。

在 WMA 和 WMC 之间,富集的差异表达基因与生物过程、细胞组成和分子功能相关的分别有 15、0、15 个 GO term。其中与生物过程相关富集最显著的 term 为发病机制(pathogenesis),富集基因数量最多的 term 为胺生物合成过程(amine biosynthetic process);与分子功能相关富集最显著、基因数量最多的 term 为内肽酶抑制剂活性(endopeptidase inhibitor activity)、内肽酶调节剂活性(endopeptidase regulator activity)、肽酶抑制剂活性(peptidase inhibitor activity)、肽酶调节剂活性和酶抑制剂活性(enzyme inhibitor activity)。

在 WMA 和 WMD 之间,富集的差异表达基因与生物过程、细胞组成和分子功能相关的分别有 19、5、6 个 GO term。其中与生物过程相关富集最显著、基因数量最多的 term 为电子转运(electron transport),其次为细胞氮化合物生物合成过程(cellular nitrogen compound biosynthetic process);与细胞组成相关,富集最显著、基因数量最多的 term 为类囊体(thylakoid)和类囊体组分(thylakoid part);与分子功能相关富集最显著、基因数量最多的 term 为氧化还原酶活性(oxidoreductase activity)。

2.3.4　差异基因的 KEGG 注释　KEGG 是一个整合了基因组、化学和系统功能信息的数据库,是系统分析基因产物在细胞中的代谢途径以及基因产物功能的数据库[12]。本研究以 KEGG 代谢途径数据库为依据[13],差异表达基因显著富集的前 20 条通路见图 5。在 WMA 和 WMB 之间,差异表达基因富集最显著的途径有光合作用、光合作用天线蛋白、乙醛酸和二羧酸代谢、卟啉和叶绿素代谢等;在 WMA 和 WMC 之间,差异表达基因富集最显著的途径有异喹啉生物碱生物合成、牛磺酸和亚牛磺酸代谢、托烷和哌啶、吡啶生物碱的生物合成等;在 WMA 和 WMD 之间,差异表达基因富集最显著的途径有光合作用天线蛋白、光合作用、卟啉和叶绿素代谢、乙醛酸和二羧酸代谢等。

3　结论与讨论

多倍体化后,随着染色体的加倍,染色体上每个位点等位基因的数量也会发生增倍,并可能由此导致基因表达质和量的变化[7]。本试验中,通过对二倍体西瓜和四倍体西瓜的差异表达基因进行 GO 功能注释,主要 GO term 为细胞氮化合物生物合成过程、光系统、半胱氨酸型内肽酶抑制剂活性、氧化还原酶活性以及类囊体。Compton 等认为叶长与叶宽是西瓜倍性水平很好的标志,四倍体子房直径是二倍体子房直径的 1.4 倍[14]。蔡力研究发现,同一发育时期四倍体紫锥菊叶绿素 a、叶绿素 b 和总叶绿素含量均高于二倍体紫锥菊[15]。本次试验发现的差异表达基因大都与植株的光合作用有关,一定程度上验证、解释了前人的研究,说明四倍体植株形成需要调控光合作用的相关基因有较高的表达量。

通过 KEGG 显著性富集分析发现差异表达基因主要在光合作用、光合作用天线蛋白、叶绿素和卟啉代谢、乙醛酸和二羧酸代谢等通路中较为活跃,说明它们在四倍体植株形成和生长过程中发挥重要作用。

多倍体化不仅导致植物基因组大小以及结构

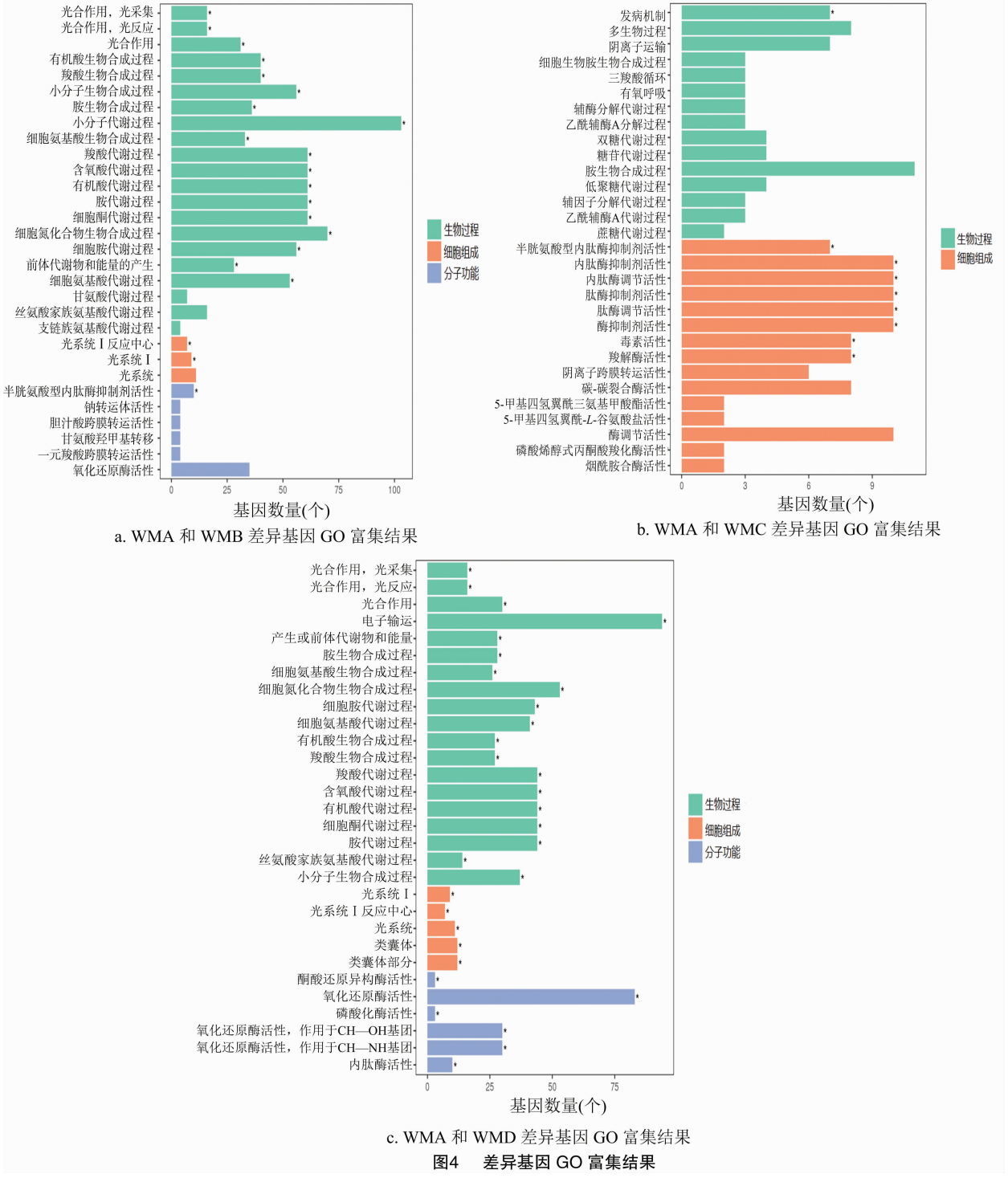


图4　差异基因 GO 富集结果

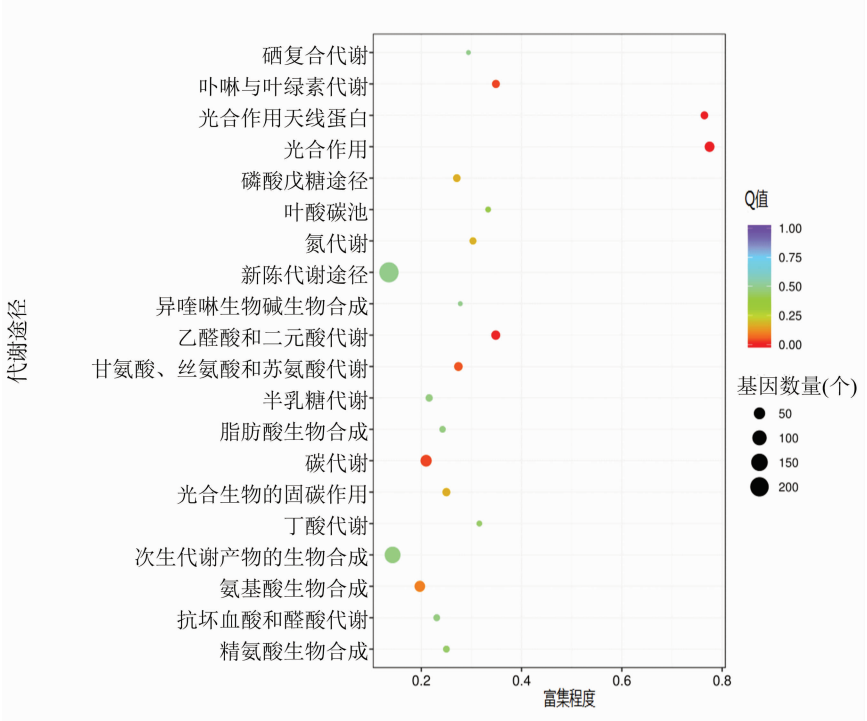
发生改变，还影响了基因的表达[16]。为了挖掘在二倍体和四倍体西瓜中与倍性相关的基因，笔者查找了几个显著富集的代谢通路下的差异表达基因，结合基因定量 FPKM，发现一些在四倍体中表达量明显高于二倍体的基因：*Cla97C10G205730*、*Cla97C10G187010*、*Cla97C01G019450*、*Cla97C02G026280*、*Cla97C01G004920* 和 *Cla97C07G140200*，并对它们进行了 GO 功能注释。

Cla97C10G187010、*Cla97C02G026280*、*Cla97C01G004920*、*Cla97C07G140200* 是调控细胞核成分的基因，控制多倍体西瓜细胞遗传与代谢。研究发现西瓜四倍体叶片比二倍体的叶片大而且厚，颜色也较深[17]，*Cla97C10G205730* 和 *Cla97C01G019450* 是调控西瓜细胞壁、液泡膜、内质网、高尔基体等膜结构的基因，它们在四倍体西瓜中的表达量上调较为明显，

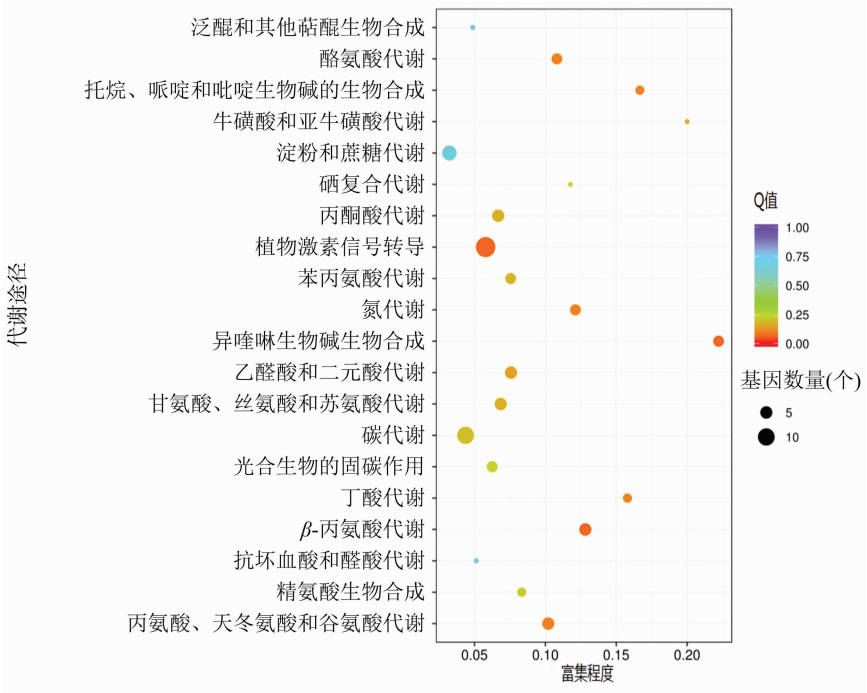
促进了细胞中膜结构的发育，可能是多倍体西瓜叶片大而厚的主要原因。

另外，对 3 组对比中差异倍数达到 2 倍以上的差异表达基因进行了注释分析。结果表明二倍体和四倍体西瓜中高水平差异基因大都是上调表达，主要包括氧化应激蛋白、脱落酸受体、乙烯响应转录因子等。这些差异表达基因可能是西瓜多倍体形成的关键因子。

总之，通过转录组初步分析，我们较为全面地了解了西瓜二倍体与四倍体的转录水平变化，发现了一些与倍性相关的基因，进一步深入研究这些基因的生理功能，有助于揭示多倍体形成的分子机制，

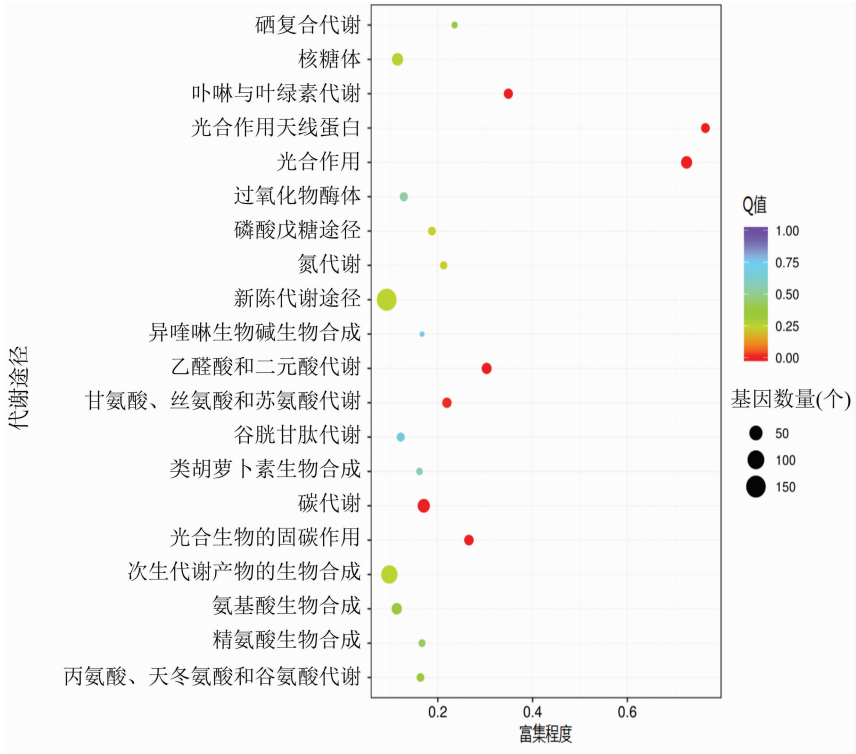


a. WMA 和 WMB 差异基因 KEGG 富集散点图



b. WMA 和 WMC 差异基因 KEGG 富集散点图

图5　差异基因 KEGG 富集散点图



c. WMA 和 WMD 差异基因 KEGG 富集散点图

图5(续)

丰富人类对植物多倍性的认知。

参考文献：

[1]中华人民共和国农业部. 中国农业统计资料 2016[M]. 北京:中国农业出版社,2017.

[2]姚　娜,刘秀明,董园园,等. 转录组的测序方法及应用研究概述[J]. 北方园艺,2017(12):192－198.

[3] Wechter W P, Levi A, Harris K R, et al. Gene expression in developing watermelon fruit[J]. BMC Genomics,2008,9:275.

[4]Guo S G,Liu J A,Zheng Y,et al. Characterization of transcriptome dynamics during watermelon fruit development－sequencing, assembly, annotation and gene expression profiles [J]. BMC Genomics,2011,12(1):454.

[5]Zhu Q L,Gao P,Liu S,et al. Comparative transcriptome analysis of two contrasting watermelon genotypes during fruit development and ripening[J]. BMC Genomics,2017,18:3.

[6]龙娅丽,江雪飞,周　鹏,等. 二倍体西瓜及其同源四倍体叶片 sRNA 表达谱分析[J]. 热带作物学报,2018,39(4):661－668.

[7]王　涛,陈孟龙,刘　玲,等. 植物多倍体化中基因组和基因表达的变化[J]. 植物学报,2015,50(4):504－515.

[8]Guo S G, Zhang J G, Sun H H, et al . The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions[J]. Nat Genet,2013,45:51－58.

[9] Anders S, Huber W. Differential expression analysis for sequence count data[J]. Genome Biol,2010,11:1－12.

[10] Liu Y, Wei H B, Ma M D, et al. Arabidopsis *FHY3* and *FAR1* regulate the balance between growth and defense responses under shade conditions[J]. The Plant Cell,2019,31(9):2089－2106.

[11] Young Matthew D, Wakefield M J, Smyth G K, et al. Gene ontology analysis for RNA－seq:accounting for selection bias[J]. Genome Biology,2010,11:R14.

[12]张少平,洪建基,邱珊莲,等. 紫背天葵高通量转录组测序分析[J]. 园艺学报,2016,43(5):935－946.

[13]Kanehisa M, Arak M, Goto S, et al. KEGG for linking genomes to life and the environment[J]. Nucleic Acids Research,2008,36:480－484.

[14]Compton M E, Barnett N, Gray D J. Use of fluorescein diacetate (FDA) to determimne ploidy of *in vitro* watermelon shoots[J]. Plant Cell, Tissue and Organ Culture,1999,58:199－203.

[15]蔡　力. 二倍体和四倍体紫锥菊中叶形态结构及其光合效率的比较研究[D]. 广东:华南农业大学,2016:35.

[16]王家利,王　芳,郭小丽,等. 同源多倍体化效应研究进展[J]. 中国农学通报,2013,29(12):22－29.

[17]施先锋,彭金光,李煜华,等. 西瓜多倍体鉴定方法的研究[J]. 浙江农业科学,2010(2):273－274.