• 药理与临床 •

月腺大戟素 A 抗乳腺癌作用机制转录组分析

覃福礼[1]，赵 亮[2]，周艳卿[2]，周 瑾[2]，须秋萍[2]，李群英[2]，焦 杨[1*]

1. 广西医科大学药学院 药理教研室，广西 南宁 530021
2. 上海市宝山区罗店医院 药剂科，上海 201908

摘 要：目的 利用二代测序技术筛选出具有显著性变化的基因，并探讨月腺大戟素 A 在转录组学层面上抗乳腺癌活性的作用机制。**方法** 从月腺大戟药材中提取乙酰间苯三酚类化合物月腺大戟素 A，对 MCF-7 细胞（乳腺癌细胞的 luminal A 型）进行干扰，观察被干扰后的细胞与正常细胞差异基因表达，采用二代高通量测序平台（Illumina Hi-Seq）测序技术分别对对照组和实验组各 3 个样本进行高通量转录组测序并进行数据分析。**结果** 对照组和实验组分别总获得 123 656 848、123 974 262 个干净序列（clean reads），分别对比到参考基因组上的序列为 119 762 214、119 881 622，各占总数的 96.85%、96.69%；2 组转录组对照可得：差异基因总数为 1 695 个，其中上调基因 770 个，下调基因 925 个，可清楚注释的基因有 3 874 个。应用基因本体论（gene ontology，GO）和京都基因与基因组百科全书（Kyoto encyclopedia of genes and genomes，KEGG）进行生物功能富集分析，GO 分析发现这 3 874 个基因主要涉及生物过程（1 270 个）、细胞组成（1 322 个）与分子功能（1 282 个）3 个大类的 45 个小类，包括细胞的生长发育过程、信号蛋白活性、膜以及基因表达的调控等过程；KEGG 分析发现差异表达基因涉及 263 条信号通路，主要代谢通路为 PI3K-Akt 信号通路、MAPK 信号通路；以及碳水化合物代谢、心肌系统和细胞生殖系统等生物过程。**结论** 利用二代高通量测序平台测序技术一共筛选、鉴定出差异基因 1 695 个，更深入了解了月腺大戟素 A 与 MCF-7 细胞基因之间的相互关系，为乳腺癌治疗提供了一些理论基础。

关键词：月腺大戟素 A；乳腺癌；RNASeq；转录组；功能基因；代谢通路；高通量测序平台

中图分类号：R285.5 **文献标志码：**A **文章编号：**0253 - 2670(2020)04 - 1003 - 13

DOI: 10.7501/j.issn.0253-2670.2020.04.026

Anti-breast cancer mechanism of transcriptome analysis of ebracteolatain A

QIN Fu-li[1], ZHAO Liang[2], ZHOU Yan-qing[2], ZHOU Jin[2], XU Qiu-ping[2], LI Qun-ying[2], JIAO Yang[1]

1. Department of Pharmacology, School of Pharmacy, Guangxi Medical University, Nanning 530021, China
2. Department of Pharmacy, Luodian Hospital of Baoshan District of Shanghai, Shanghai 201908, China

Abstract: Objective To research the mechanism of ebracteolatain A against breast cancer cells, screening the genes with significant changes using second-generation sequencing, and explore the anti-breast cancer mechanism of action of ebracteolatain A at the transcriptomics level. **Methods** The acetyl phloroglucinol compound ebracteolatain A was extracted from *Euphorbia ebracteolata*, interferencing with MCF7 cells (luminal A type of breast cancer cells) to observe differential gene expression between the interfered cells and normal cells. High-throughput transcriptome sequencing and data analysis were performed on three groups of control groups and three experimental groups using Illumina Hi-Seq sequencing technology. **Results** A total of 123 656 848, 123 974 262 available reads were obtained in the control group and experimental group, respectively, the reads on the reference genome were 119 762 214, 119 881 622, respectively, accounting for 96.85% and 96.69% of the total; Two groups of transcriptome controls were available: the total number of differential genes was 1 695, of which 770 were up-regulated, 925 were down-regulated, and 3 874 genes were clearly annotated. Bio-enrichment analysis was carried out by using Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG). GO analysis found that these 3 874 genes mainly involved in biological processes (1 270), cell composition (1 322) and molecular function (1 282), 45 subcategories of three major categories, including cell growth and development, signaling protein activity, membrane and regulation of gene expression. KEGG analysis revealed that the differentially expressed genes involved 263

收稿日期：2019-07-01

基金项目：国家自然科学基金资助项目（81303300）；上海宝山区科技创新基金（19-E-2）

作者简介：覃福礼（1994—），男，壮族，广西南宁人，硕士研究生，专业方向为分子药理学。Tel: 18275748993 E-mail: 13471686168@163.com

***通信作者** 焦 杨，女，教授，主要从事民族药理学研究。Tel: (0771)5358014 E-mail: jiaoyanggx@163.com

signaling pathways; The main metabolic pathways were: PI3K-Akt signaling pathway, MAPK signaling pathway, carbohydrate metabolism, myocardial system and cellular reproductive system and etc. **Conclusion** The results showed that 1 695 differential genes were screened and identified by Illumina Hi-Seq sequencing technology, and the relationship between the genes of ebracteolatain A and MCF7 cells was further understood, which provided some theoretical cornerstones for breast cancer treatment.

Key words: ebracteolatain A; breast cancer; RNASeq; transcriptome; functional gene; metabolic pathway; Illumina Hi-Seq

乳腺癌的发病率和死亡率在中国虽然处于较低的水平，但却有逐年上升的趋势，而且城乡分布差异明显。乳腺癌的发病率在女性恶性肿瘤中排名第 1 位，约有 29%；死亡率占女性恶性肿瘤死亡率的第 2 位，对女性的身心健康造成很大的威胁，乳腺癌是全世界女性死亡最主要和常见的原因[1-6]。目前仍缺乏相关的乳腺癌组织病理学，诊断时期的分子分型等疾病特征的分布[7]，它是一种异质性疾病，分为几种亚型基于组织学标记或各种基因表达谱[8]，乳腺癌根据其分子生物学特征可分为 luminal A、luminal B、人表皮生长因子受体 2（human epidermal growth factor receptor 2，HER2）阳性及三阴型，对乳腺癌的治疗停留于进行外科手术切除结合靶向药物治疗，化学治疗以及放射治疗，尚缺乏理想的药物，这些药物选择性差，对患者的副作用大，肿瘤细胞易对其产生耐药性，大大影响到其临床上的应用[9-10]。

月腺大戟 *Euphorbia ebracteolata* Hayata 为大戟科植物，以根入药，多生长于草地、树下，主要产地为湖北、山东、安徽、河南等地，最初记载于《神农本草经》，其味辛，性平，有大毒，有祛痰和破积杀虫的功效，民间常用于治疗结核和癌症[11-12]，月腺大戟中含有多种化学成分，包括苯乙酮类、萜类、醇类、酯类，以及鞣质、黄酮类等[13-16]。前期从月腺大戟中提取得到月腺大戟素 A，其对 4 种乳腺癌细胞 SUM149（三阴型）、MCF-7（luminal A 型）、ZR-75-1（luminal B 型）、SK-BR-3（HER2 阳性型）增殖均有抑制作用[17]。

转录组测序的研究对象是特定细胞在某一功能状态下所能转录出来的所有 mRNA 的总和，这些技术可以进一步分析目标药物对正常组织和致病过程中的影响，新一代高通量测序技术能够全面快速地获得某一物种特定组织或器官在某一状态下的几乎所有转录本序列信息，从而准确地分析基因表达差异、基因结构变异、筛选分子标记（SNPs 或 SSR）等生命科学重要问题[18-21]。因此，本研究用月腺大戟素 A 干扰乳腺癌 MCF-7 细胞，再采用二代高通量测序平台测序技术分别对 3 组对照组和 3 组实验组进行测序分析，对测序得到的基因结果进行差异表达筛选，筛选出干扰前后 MCF-7 细胞中具有显著性差异的基因，并进一步进行生物信息学分析，探讨月腺大戟素 A 抗乳腺癌活性的作用机制，为月腺大戟素 A 在乳腺癌的进一步深入治疗提供新的思路与理论依据。

1 材料

1.1 细胞

人源乳腺癌 MCF-7 细胞（luminal A 型，由王红阳院士所在的国际肝癌研究中心馈赠）。

1.2 药材

月腺大戟药材购自安徽（产地山东，批号 20140805、20141007），经第二军医大学药学院生药学教研室孙莲娜教授鉴定为月腺大戟 *Euphorbia ebracteolata* Hayata 的根。

1.3 仪器与试剂

AE240 型十万分之一电子天平（梅特勒-托利多瑞士有限公司）；XW-80A 型涡旋混合器（上海医科大学仪器厂）；DJ-04 粉碎机（上海定久仪器设备有限公司）；HH-Sl-2 电热恒温水浴锅（上海跃进医疗器械厂）。乙腈为色谱纯（美国费希尔公司）；乙醇为分析纯（江苏强盛功能化学股份有限公司）；水为超纯水（江苏权坤环保科技有限公司）。

2 方法

2.1 月腺大戟素 A 的制备

采用回流提取、溶剂萃取和吸附色谱法从月腺大戟药材中分离、纯化得到月腺大戟素 A。具体步骤为取 30 kg 干燥月腺大戟药材，用 100 L 80%乙醇回流提取 2 次，每次 2 h，趁热滤过，合并提取液，浓缩至无乙醇残留（浓缩体积至 8 L）。取 2 L 浓缩液加 1 L 水稀释，醋酸乙酯每次萃取体积均为 3 L，减压回收溶剂，得醋酸乙酯部分 160 g。醋酸乙酯部分过正相硅胶色谱柱，以石油醚-醋酸乙酯（20∶1）洗脱，得洗脱物。再过正相硅胶色谱柱，石油醚-醋酸乙酯-甲酸（15∶1∶0.1）洗脱，得单一化合物，通过质谱（MS）和核磁共振（NMR）谱确定其结构，其质量分数≥98%。

2.2 细胞培养

MCF-7 细胞常规复苏后，细胞株接种于培养瓶

中，加入 DMEM 培养液（含 10%胎牛血清、青链霉素各 100 U/mL），置于 37 ℃、5% CO_2 的培养箱中培养。每天观察，细胞呈单层贴壁生长，隔天换 1 次培养液，2～3 d 传代 1 次。传代时先使用 PBS 清洗 3 次，0.25%胰酶消化 5 min 左右，肉眼观察可见细胞完全脱壁消化下来即可，以完全培养液终止消化，并吹打制成细胞悬液后，计数后接种于 6 孔板中，继续置于培养箱中培养，取对数生长期细胞用于实验。

2.3 给药处理

设实验组和对照组，每组设置 3 个样本以做平行对照。通过前期实验室对月腺大戟素 A 的药理学特性研究确定本实验给药浓度为月腺大戟素 A 对 MCF-7 细胞的半数抑制浓度（IC_{50}），即 6.16 μmol/L。实验组 MCF-7 细胞添加含有终浓度为 6.16 μmol/L 月腺大戟素 A 的细胞培养基，对照组 MCF-7 细胞添加等量普通细胞培养基，培养 72 h 后，消化收集 MCF-7 细胞用于下一步实验。

2.4 RNA 提取及质控

首先在每盘细胞中加 TRIZOL 试剂（Gibco 公司）1 mL，摇匀，消化 3～5 min 后，将消化好的细胞裂解液吸到 DEPC 水处理过的 1.5 mL EP 管中，加氯仿 0.2 mL，轻摇 15 s。室温静置 2～3 min 后，12 000 r/min，4 ℃，离心 15 min，然后取上清无色水相到 EP 管（DEPC 水处理过）中，加 0.5 mL 异丙醇，室温下静置 10 min，12 000 r/min，4 ℃离心 10 min。观察总 RNA 在管底的白色沉淀，弃上清，75%乙醇 1.0 mL 洗涤后，7 500 r/min，4 ℃离心 5 min。弃上清，将 RNA 沉淀进行冻干处理，加入 DEPC 水 20～30 μL，涡旋混匀，55～60 ℃水浴 10 min 溶解总 RNA，测吸光度值，最后进行电泳。电泳跑胶质控结果显示，本批细胞样本提取的 RNA 合格，样本总量及质量均满足后续建库要求。

2.5 文库构建

样本检测合格后，进行文库构建，用带有多聚胸腺嘧啶、T 重复寡核苷酸 Oligo（dT）的磁珠富集 mRNA，加入破碎缓冲液将 mRNA 进行随机打断。以 mRNA 为模板，用六碱基随机引物合成第 1 条 cDNA 链，然后加入缓冲液、dNTPs、核糖核酸酶 H（RNase H）和 DNA 聚合酶 I 合成第 2 条 cDNA 链，利用 DNA 纯化磁珠（AMPure XP beads）纯化 cDNA；纯化的双链 cDNA 再进行末端修复、加 A 尾并连接测序接头，然后用 DNA 纯化磁珠（AMPure XP beads）进行片段大小选择；最后通过 PCR 富集得到 cDNA 文库，从而完成整个文库构建工作。文库构建完成后，对文库质量进行检测，检测结果达到要求后才能对样品进行上机测序。

2.6 上机测序

文库检测合格后，不同文库按照目标下机数量进行汇集，选择二代高通量测序平台（Illumina HiSeq2000）进行测序分析。

2.7 数据分析

经过测序平台测序所得原始下机序列（raw reads），通过去低质量序列、去接头污染等过程完成数据处理得到干净序列（clean reads，通过删除质量不好的 reads 后得到的数据），后续所有的分析都是基于 clean reads。转录组测序数据分析流程主要分为 3 部分：测序数据质控、数据比对分析和转录组深层分析。其中，测序数据质控包括过滤测序所得序列、评估测序数据质量以及计算序列长度分布等；数据比对分析主要是针对比对到基因组中的序列，根据不同的基因组注释信息依次进行分类和特征分析，并计算相对应的表达量；转录组深层分析包括差异表达分析、可变剪接分析、新转录本预测和变异分析等。

2.7.1 质量控制 测序中有些质量很差，有些会有接头污染，甚至会有其他物种的污染情况，运用 Fast-QC（http://www.bioinformatics.babraham.ac.uk/projects/fastqc/）软件对测序数据的质量进行整体评估，包括碱基的质量值分布，质量值的位置分布，GC 含量，PCR 重复含量，K-mer（K-mer 指将 1 条 read 连续切割，挨个碱基划动得到的一序列长度为 K 的核苷酸序列）的频率（frequency）等，对质控后的基因 GC 浓度进行线性分析，其含量分布情况，实际线条分布跟正态分布接近，结果较好，可以用于后续分析；采用 HISAT2 软件对 RNA-seq 数据进行比对，参考基因组版本：GRCh38；将过滤后的 clean reads 进行绘图，统计测序数据的绘图率、基因在染色体上的分布和基因组结构的分析；再根据测序得到的序列总数进行 reads 在染色体上的分布统计。

2.7.2 差异基因的筛选 采用 FPKM 法计算基因的原始表达量，FPKM 法能消除基因长度和测序量差异对计算基因表达的影响，计算得到的基因表达量可直接用于比较不同样品间的基因表达差异。如果一个基因存在多个转录本，则用该基因的最长转录本计算其测序覆盖度和表达量。根据数据，进行差异基因表达分析，用国际公认算法 DESeq2 对基因表

达进行差异筛选分析。其中筛选的条件为 $\log_2$FC＞1 或＜−1，错误检出率（FDR）＜0.05。并绘制聚类图和火山图。

2.7.3 差异表达基因的功能分析及其关系网络构建 将分析得到的差异基因基于数据库分别从生物过程（biological process，BP）、分子功能（molecular function，MF）、细胞组成（cellular component，CC）这 3 个层面进行基因本体（gene ontology，GO）注释，得到基因参与的所有 GO，采用 GO Term Fisher 软件[22]精确检验得到显著性水平（*P* 值），通过 BH 对于多重假设检验的结果进行校正，确定 GO 的 FDR，从而筛选出差异基因富集的显著性 GO。通过 GO 功能显著性富集推测差异表达基因行使的主要生物学功能。实验中基因同时参与了很多显著性 GO，基于 GO 的层次结构，将所有 GO 之间的相互调控及从属关系整理成数据库，通过构建功能关系网络，总结实验影响的功能群体，以及显著性功能的内在从属关系。以差异基因所做 GO 分析中的显著性 GO-Term（*P*＜0.01）为研究对象进行功能调控分析，构建功能调控网络。

2.7.4 KEGG 分析及通路作用关系构建 通路分析是基于基因注释数据库，检测差异基因显著通路的手段。将筛选出的差异基因于基因组百科全书（Kyoto encyclopedia of gene and genomes，KEGG）[23]信号通路进行通路注释分析，得到差异基因参与的所有 pathway term，采用 Fisher 软件检验计算通路的显著性水平（*P* 值），以确保整体错误发生率控制在一定水平范围[24]。从而筛选出差异基因富集的显著性 pathway term。选用显著性通路中的 pathway term，根据 KEGG 数据库中通路之间的关联关系整合成显著性通路之间的信号转导网络。在此以差异基因所做的路径分析（path-analysis）中的显著性 path-term（*P* 值＜0.05）进行信号通路相互关系网络构建。利用 Pathway 分析可进一步了解基因所参与的代谢通路及其具体的生物学意义。KEGG 是有关通路的主要公共数据库。

3 结果与分析

3.1 测序数据质控评估

碱基质量值（quality score）是碱基识别（base calling）出错的概率的整数映射。通常使用 Phred 质量评估公式。剪辑质量值越高表明碱基识别越可靠，碱基测错的可能性越小。表 1 对于碱基质量值为 Q20 的碱基识别，100 个碱基中有 1 个会识别出错；对于碱基质量值为 Q30 的碱基识别，1 000 个碱基中有 1 个会识别出错；对于碱基质量值为 Q40 的碱基识别，10 000 个碱基中有 1 个会识别出错。以测序循环为单位，对单个样本所有序列平行测序的剪辑质量值做分布图，可以查看单个样本各个测序循环及整体的测序质量。图 1 为转录组碱基质量分布图，图中碱基质量值大于 20 表示碱基测序准确度大于 99%，可用于后续分析。

表 1 碱基质量值与碱基识别出错的概率对应关系

Table 1 Correspondence between base quality value and probability of base recognition error

碱基质量值	碱基识别出错的概率	碱基测序准确度/%
Q10	1/10	90.00
Q20	1/100	99.00
Q30	1/1 000	99.90
Q40	1/10 000	99.99

GC 含量分布检查用于检测有无 AT、GC 分离现象，而这种现象可能是测序或者建库所带来的，并会影响后续的定量分析。在二代高通量测序平台的转录组测序中，反转录成 cDNA 时所用的 6 bp 的随机引物会引起前几个位置的核苷酸组成存在一定的偏好性。而这种偏好性与测序的物种和实验室环境无关，但会影响转录测序的均一性程度。此外，理论上普通文库的 G 和 C 碱基及 A 和 T 碱基含量每个测序循环上应分别相等，且整个测序过程稳定不变，呈水平线。图 2 为转录组每条序列的 GC 含量，蓝线是理论 GC 含量分布情况，红线是实际 GC 含量分布情况，每个样本的左右小图分别为质控筛选前后的 GC 含量分布，左右两图中的实际线条分布跟正态分布接近，结果较好，可用于后续分析。

表 2 为实验组和对照组样本的转录组测序后的统计结果，2 组样品分别获得 139 145 182、139 426 050 条可用 reads，总碱基数分别是 21 008 909 801 bp 和 21 051 017 061 bp，其中识别准确率高的碱基总数为 18 632 072 837 bp 和 18 679 920 240 bp，结果表明其筛选后的基因识别率都达 88.0%以上，GC 含量（GC 含量指碱基 G 和 C 的数量总和占总的碱基数量的百分比）在 50%左右，说明结果测序质量好，保证数据研究的可靠性。

测序中常常产生数亿的结果序列，不可避免地会出现低质量的测序结果，实验对数据进行筛选过滤处理（表 3），2 组样本经过滤后分别得到 123 656 848、

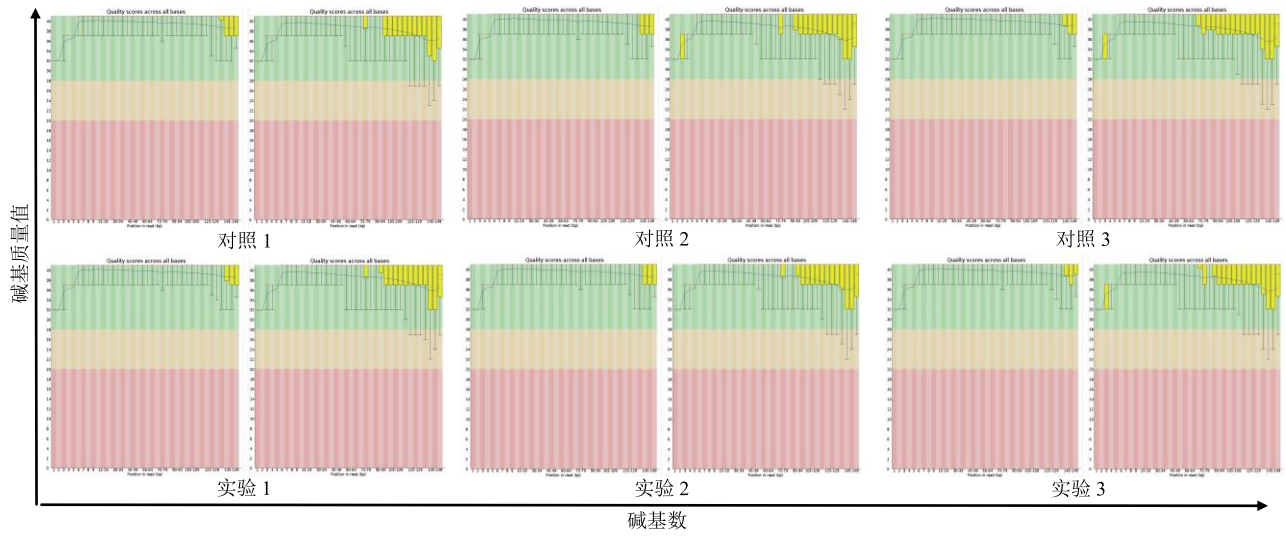


图 1 转录组碱基质量分布

Fig. 1 Distribution map of transcriptome base mass

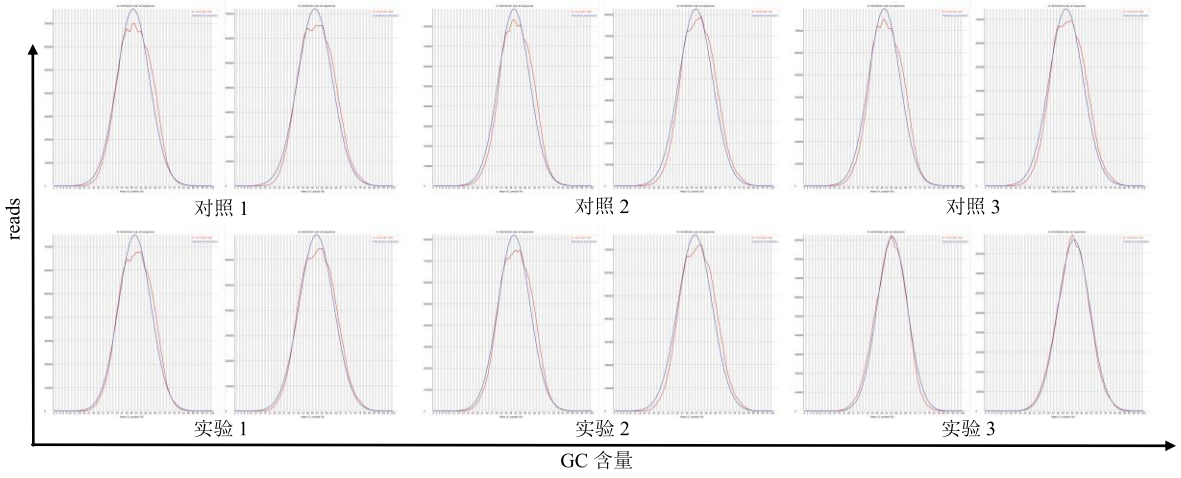


图 2 转录组每条序列的 GC 含量

Fig. 2 Per sequence GC content of transcriptome

表 2 转录组测序数据评价分析

Table 2 Evaluation of transcriptome sequencing data

样本	筛选前总 reads	筛选前总碱基数	筛选后总 reads	筛选后总碱基数	reads 识别率/%	碱基识别率/%	筛选后 GC 含量/%
对照	139 145 182	21 008 909 801	123 656 848	18 632 072 837	88.87	88.68	51.33
实验	139 426 050	21 051 017 061	123 974 262	18 679 920 240	88.91	88.73	52.17

表 3 转录组数据的 Mapping 结果统计

Table 3 Mapping statistics of transcriptome sequencing data

样本	不映射碱基数	映射碱基数	映射率/%	精准映射碱基数	精准映射率/%
对照	3 894 634	119 762 214	96.85	110 464 860	89.33
实验	4 092 640	119 881 622	96.69	110 358 595	89.01

123 974 262 个 clean reads，本实验使用指定的基因组作为参考进行序列比对及后续分析，对比到参考基因组上的分别有 119 762 214 个与 119 881 622 个，是总数的 96.85%和 96.69%。比对效率是指绘图序列占可读序列的百分比，是转录组数据利用率最直观的体现。通过比对效率，可以评估所选基因组组装是否能满足生物信息学分析的要求。表 3 结果表明对比结果较好，可以准确对比到功能的效率分别是 89.33%和 89.01%，表明测序结果丰富，有效性高。

3.2　基因组结构分析

过滤后的可读序列进行制图（mapping），统计测序数据的 mapping 率、基因基因组结构的分析。人类基因结构主要分布在 4 个区域：①编码区，包括外显子与内含子；②前导区，位于编码区上游，相当于 RNA 5’末端非编码区（非翻译区）；③尾部区，位于 RNA 3’编码区下游，相当于末端非编码区（非翻译区）；④调控区，包括启动子和增强子等。图 3 为对照组和实验组细胞的基因结构分布；由图可知，每组的基因分布规律为 Exon（外显子）＞CDS（mRNA 序列中编码蛋白质的那部分序列）＞UTR3（3’端非编码片段）＞Intron（内含子）＞InterGenic（基因间隔片段）＞UTR5（5’端非编码片段）＞Tss（转录起始位点，启动子）＞Tes（转录终点），所有样品基因分布规律一致，且绝大多数测序得到的基因都位于编码区，可以继续进行后续分析研究。

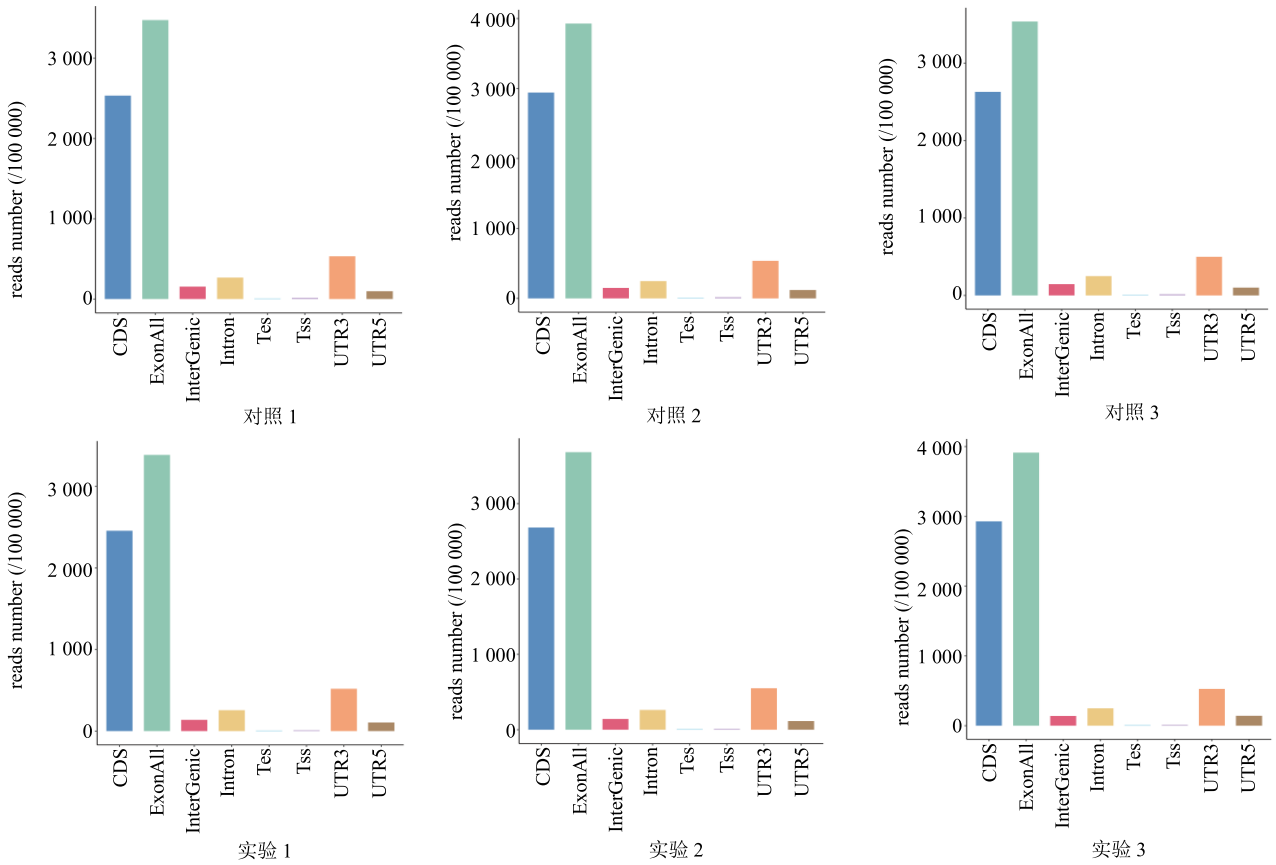


图 3　基因组结构分析

Fig. 3　Genomic structure analysis

3.3　基因表达量

基因表达水平的直接体现就是其转录本的密度情况，转录本密度越高，则基因表达量越高（图 4 左上图）。通过对基因表达量进行主成分分析（PCA），结果如图 4 左下图所示，实验组与对照组在基因表达量上能够明显区分；在 RNA-seq 分析中，可以通过定位到基因组区域或基因外显子区的测序序列（reads）的计数来估计基因的表达水平。Reads 计数除了与基因的表达量呈正比外，还与基因的长度和测序深度呈正相关。为了使不同基因、不同实验室估计的基因表达量具有可比性，人们引入 RPKM 的概念，RPKM（reads per kilobase of per

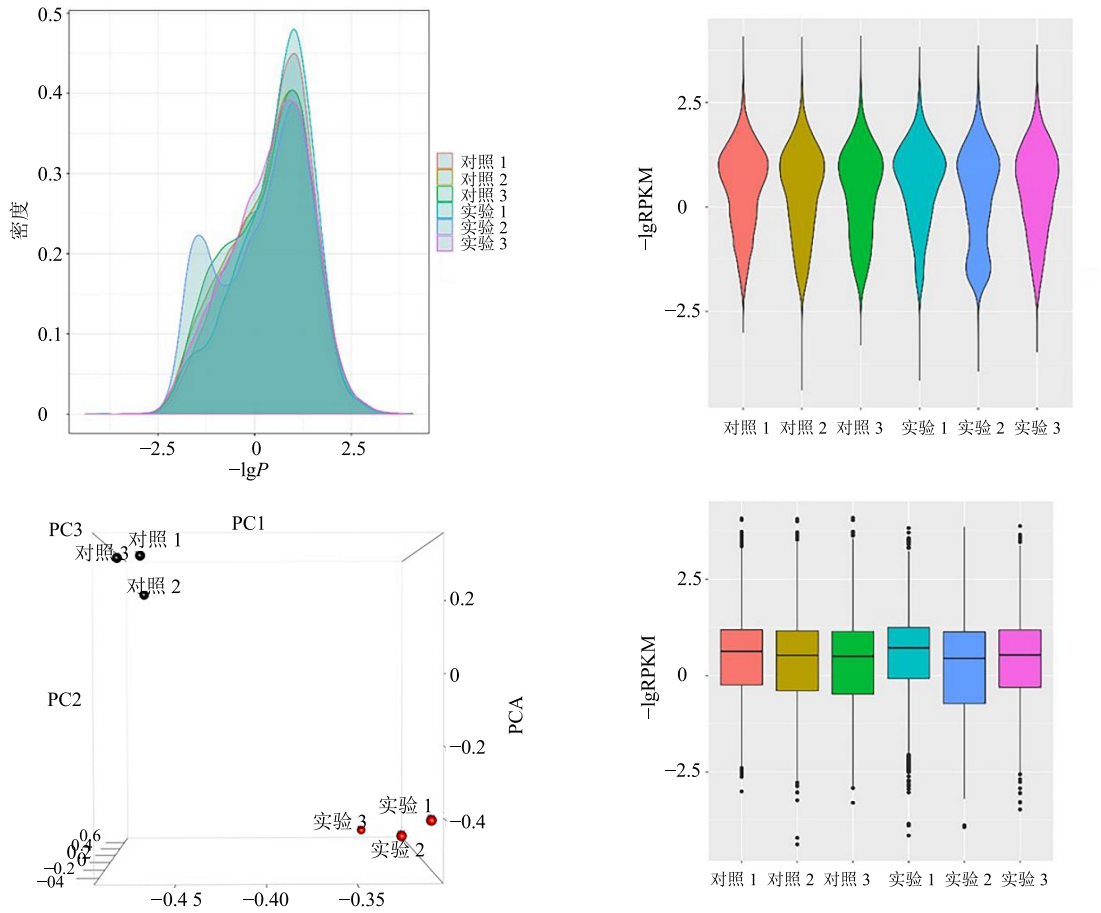


图 4 基因表达量密度分布

Fig. 4 Density distribution of gene expression

millions mapped reads）是每百万碎片（fragments）中来自某一基因每千碱基长度的碎片数目，其同时考虑了测序深度和基因长度对于碎片数的影响，是目前最为常用的基因表达量估算方法。本次采用 featureCounts 软件对各样本进行基因表达量分析，结果分别统计了不同表达水平下基因的数量以及单个基因的表达量水平。对 2 组细胞基因的 RPKM 做箱线图，进而能够在整体上检查不同条件下 RPKM 分布的情况。从箱线图中不仅可以查看单个样本基因表达水平分布的离散程度，还直观地比较不同样品的整体基因表达丰度的差异情况（图 4 右上图和右下图）。

测序得到的某个基因的 reads 数量和若干因素有关，其中最重要的有两点：（1）比对到参考序列的 reads 数量，也就是有效测序量；（2）该基因所有 Exon 的长度之和，显然一个基因的转录本越长，其获得的测序结果片段会越多。最终得到基因表达量 54 261 个；如图 4-B、D 所示 lgRPKM 标准化到（−2.5，2.5），保证数据的严谨性。

样本间基因表达水平相关性是检验实验可靠性和样本选择是否合理的重要指标。相关系数越接近 1，表明样本之间表达模式的相似度越高。若样本中有生物学重复，通常生物学重复间相关系数要求较高。所有样本间基因表达量相关性分析热图结果见图 5。结果显示，对照组和实验组样本间的相关性都非常好。

3.4 差异基因表达分析

实验将 $\log_2$FC＞1 或＜−1，FDR＜0.05 且表达差异在 2 倍以上的基因定义为差异表达基因。差异表达基因分布情况用火山图（图 6）可以直接反映，表明在给予月腺大戟素 A 前后转录组表达水平有明显的差异（红色上调基因、蓝色下调基因、黑色为无差异基因）；总计得到的差异基因数为 1 695 个，其中上调基因有 770 个，925 个下调基因（图 7）。为了更好地反映对照组和实验组差异基因的表达情况，用 Cluster 3.0 以 lgRPKM 值对不同表达模式的基因进行聚类，图 8 为基因的富集聚类图，直观地反映出给予月腺

图 5　样本间相关系数热图

Fig. 5　Heat map of correlation coefficient between samples

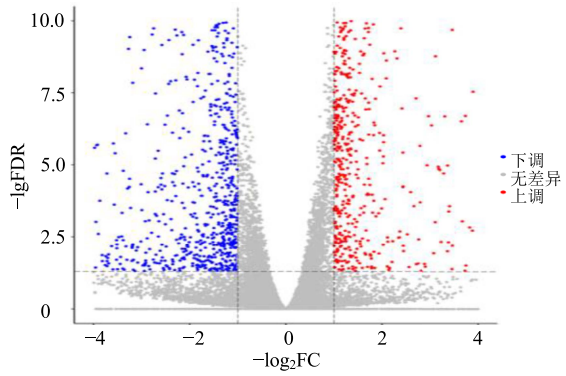


图 6　差异表达基因火山图

Fig. 6　Volcano plot of differentially expressed genes

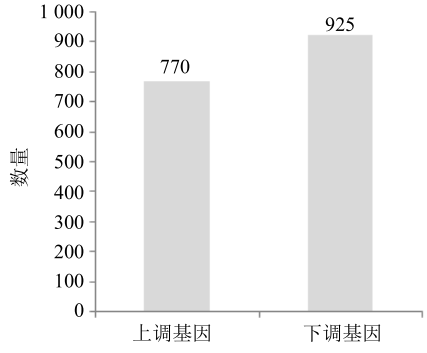


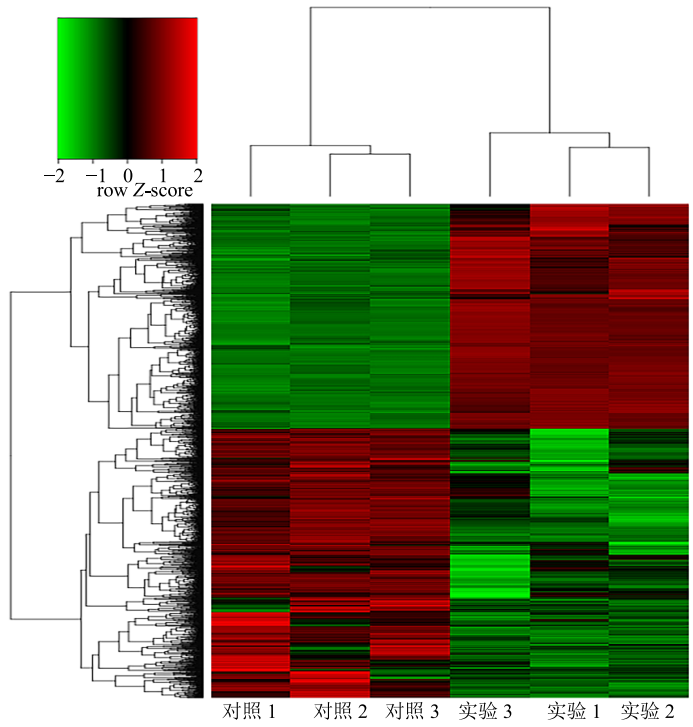
图 7　对照组与实验组差异表达基因统计

Fig. 7　Statistics of differentially expressed genes in control and experimental groups

大戟素 A 前后细胞内基因的差异变化情况。

3.5　差异表达基因功能分析

GO 功能显著性富集分析给出与基因组背景相比，在差异表达基因中显著富集的 GO 功能条目，从而给出差异表达基因与哪些生物学功能显著相关。该分析首先把所有差异表达基因向 GO 数据库（http://www.geneontology.org/）的各个 term 映射，计算每个 term 的基因数目，然后找出与整个基因组背景相比，在差异表达中显著富集的条目。图 9 是按照不同分类下最显著的 15 个 term 绘制的柱状图。差异表达基因的 GO 功能注释分类显著性大小结果，−lg*P* 表示对 *P* 值进行负对数的计算，*P* 值越小，−lg*P* 越大，表明此 term 变化的可能性越显著。对 2 组细胞中筛选出的显著差异蛋白进行 GO 注释聚类分析，主要包括 3 个层面：生物学过程（biological process，BP）、细胞组分（cellular component，CC）以及分子功能（molecular function，MF），BP、CC、MF 3 大类别 45 小类，分别包含了 4 360、590、1 231 个条目；在 BP 分类下基因涉及的生物学功能最复杂，其中显著性差异较大涉及的是细胞外基质组织、基因表达的正调节、细胞增殖的负调节、凋亡过程的积极调节、RNA 聚合酶 II 启动子转录的正调节、活性氧代谢过程的积极调节等细胞繁殖过程；CC 大类下主要差异表达基因显著性较大的是细胞膜的



行代表不同的基因，列代表不同的样品；红色表示基因表达水平上调，绿色表示下调；左上角的 color key，将 lgRPKM 标准化到（−2，2）范围内
Rows represent different genes, and columns represent different samples; Red means up-regulation of gene expression, green means down-regulation; Color key in the upper left corner, normalizing lgRPKM to (−2, 2)

图 8　差异表达基因等级聚类图

Fig. 8　Clustering chart of differentially expressed gene level

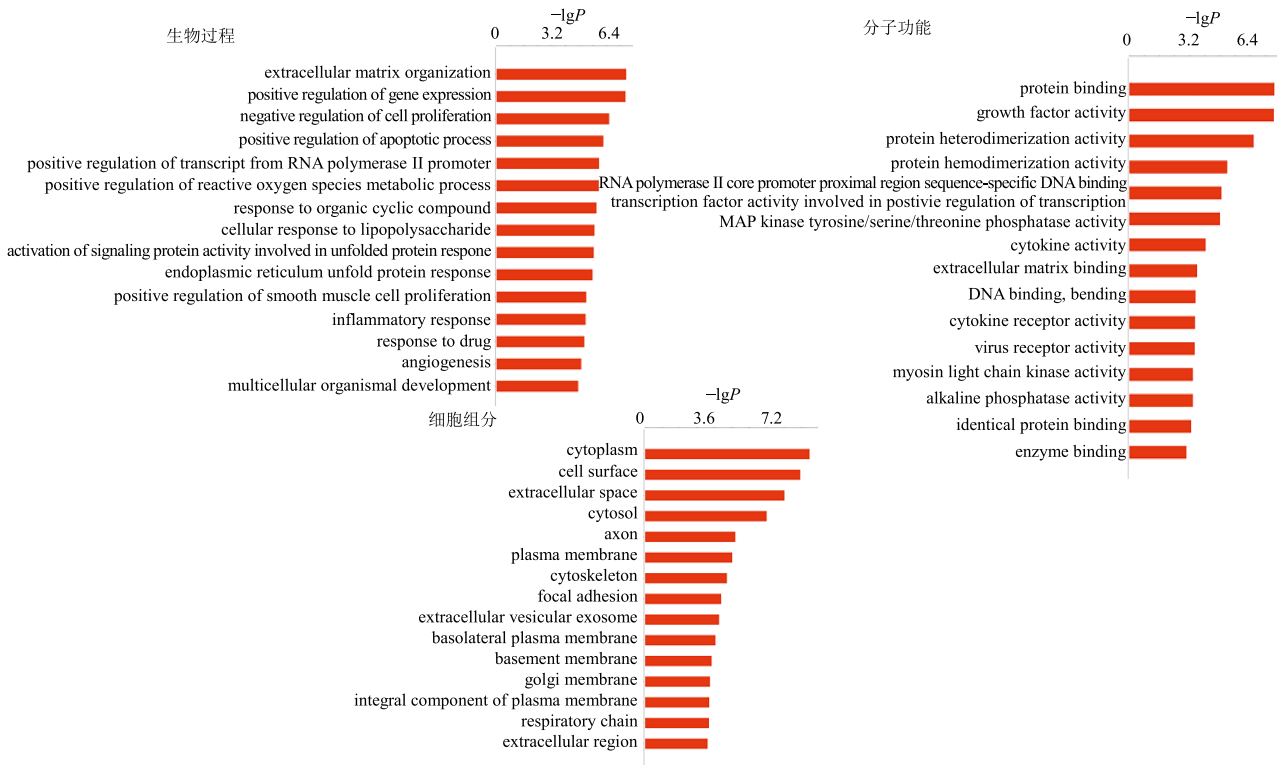


图 9　GO 功能注释分类图

Fig. 9　Classification map of GO functional annotation

组成以及膜的相关功能，主要为细胞质、细胞表面、细胞外空间、细胞液、轴突、质膜、骨架等胞膜成分；在 MF 分类中，显著性差异的为蛋白质结合、生长因子活动、蛋白质异二聚化活性、蛋白质同源二聚化活性、RNA 聚合酶 II 核心启动子近端区域序列特异性 DNA 结合转录因子活性参与转录的正调节、MAP 激酶酪氨酸/丝氨酸/苏氨酸磷酸酶活性等。

3.6　功能调控网络构建

GO 层次树形关系图能直观展示差异基因富集的 GO term 及其层级关系。树形关系图为差异基因 GO 富集分析的结果图形化展示方式，分支代表包含关系，从上至下所定义的功能范围越来越具体。根据以上显著性的 GO，GO 的层次结构，相互调控及关系差异基因做 GO-analysis 中的显著性 GO-term（P＜0.01）为研究对象进行功能调控分析，构建功能调控网络。图 10 为显著性功能间的层次树形关系图。本实验对 GO 3 大类中的 MF gene term 中显著性通路进行功能富集分析，下图通过包含关系将相关联的 GO-term 一起展示，用方框表示，而颜色的不同代表通路处于上升或下调状态以及每个通路与上下级通路之间的关系，主要包括：包含关系、反馈调节、负反馈调节等。位于网络调节图中间位置的通路可以作为本实验的重要研究对象。

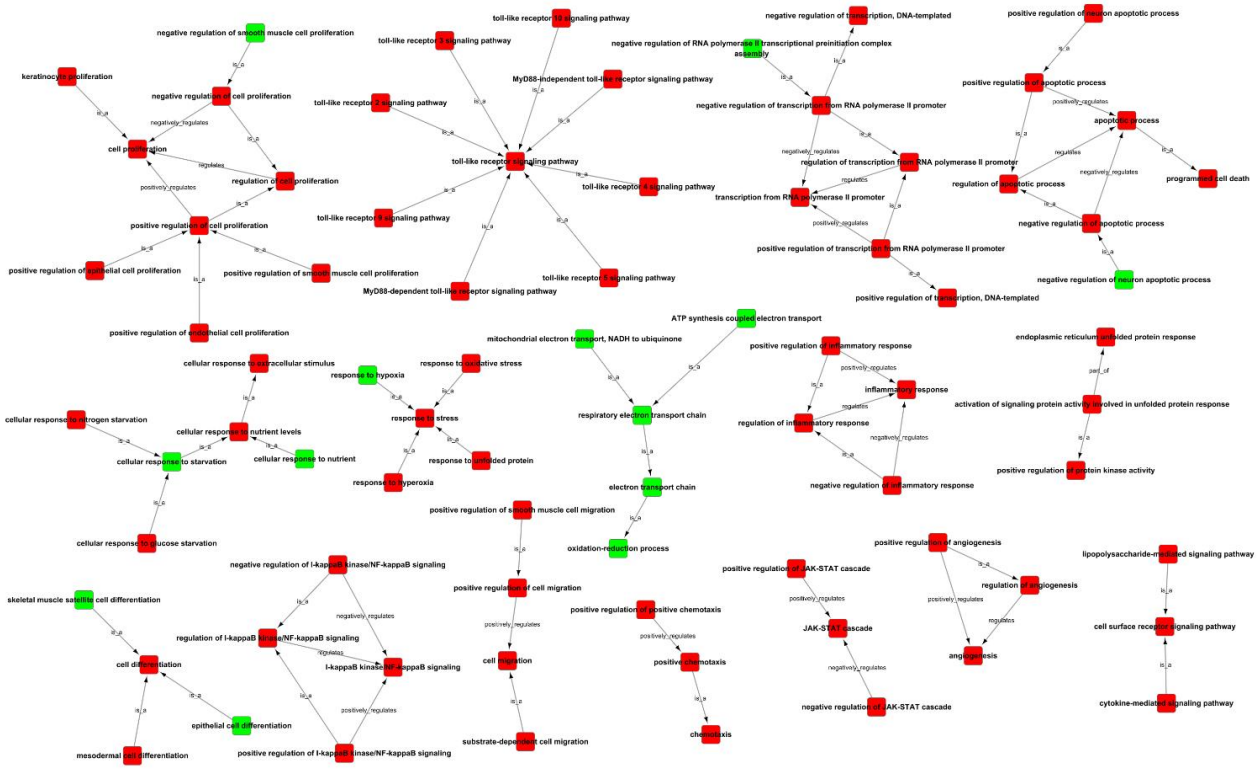
3.7　信号通路相互作用关系分析

选用显著性的 pathway term，进行信号通路相互关系网络构建，构建方法如同功能调控网络构建。结果见图 11。

3.8　信号通路分析

在生物体内，不同基因相互协调行使其生物学功能，通过通路显著性富集能确定差异表达基因参与的最主要生化代谢途径和信号转导途径。KEGG 是有关 pathway 的主要公共数据。Pathway 显著性富集分析以 KEGG pathway 为单位，应用超几何检验，找出差异基因相对于所有注释到的基因显著富集的 pathway。通过差异基因表达的 KEGG 功能注释的结果，得到差异表达基因共涉及 263 个信号通路。

差异基因 KEGG 富集散点图是 KEGG 富集分析结果的图形化展示方式。在图 12 中，KEGG 富集程度通过富集基因（rich factor）、FDR 和富集到此通路上的基因个数来衡量。其中 rich factor 指差

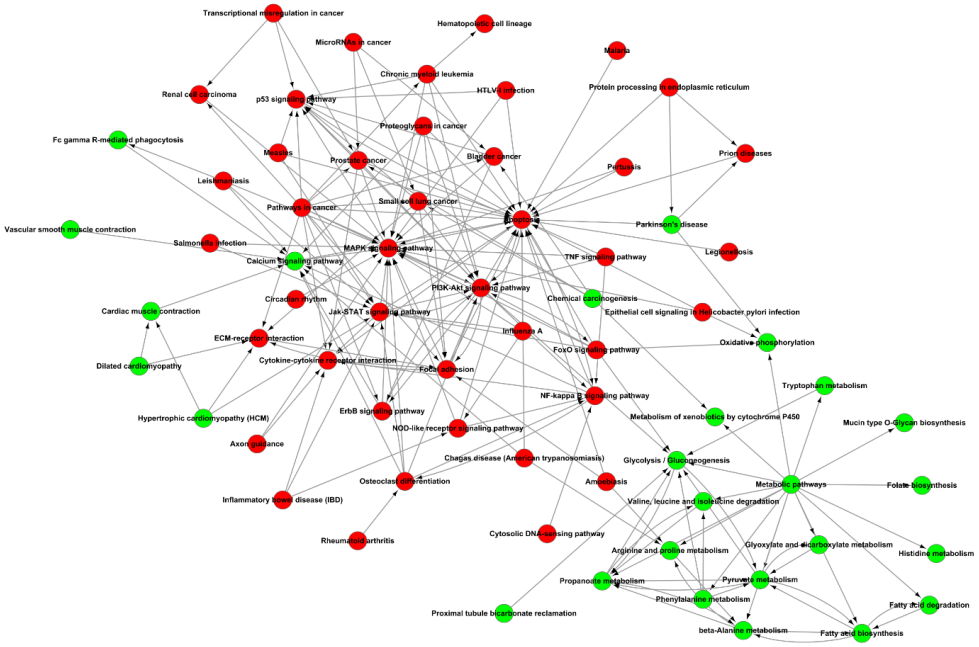


线上的标签代表靶向调控形式，红色表示该基因功能显著上调，绿色表示该基因功能显著下调（P＜0.01）

The label on the line represents the form of targeted regulation, red indicates that the function of the gene is significantly up-regulated, green indicates a significant down regulation of the gene's function (P < 0.01)

图 10　显著性功能间的层次树形关系

Fig. 10　Hierarchical tree diagram of significant function



红色代表该信号通路显著上调，绿色代表下调，箭头代表靶向调控的方向

Red represents the significant up-regulation of the signaling pathway, green represents down-regulation, and arrow represents the direction of targeted regulation

图 11　显著性信号通路关系网络

Fig. 11　Network diagram of significant signaling pathway relationship

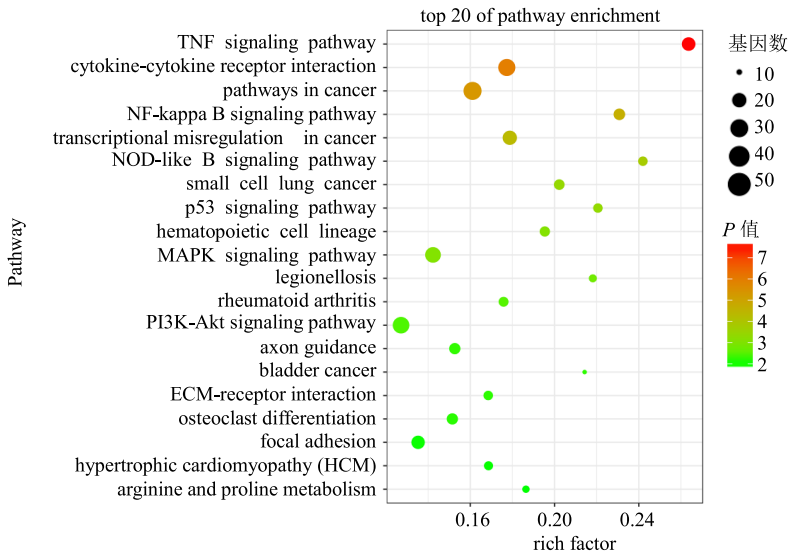


图 12　信号通路显著性富集分析

Fig. 12　Significant enrichment analysis of signaling pathway

异表达的基因中位于该 pathway 条目的基因数目与所有注释到的基因中位于该 pathway 条目的基因总数的比值。FDR 是做过多重假设检验校正之后的 *P* 值，FDR 的取值范围为 [0，1]，越接近于 0，表示富集越显著。挑选了富集最显著的 20 条 pathway 在图中进行展示。图 12 主要有 TNF 信号通路、细胞因子-细胞因子受体相互作用、癌症的途径、NF-κB 信号通路、癌症中的转录失调、NOD 样受体信号通路、小细胞肺癌通路、p53 信号通路、造血细胞谱系、MAPK 信号通路、军团病通路、类风湿关节炎通路、PI3K-Akt 信号通路、轴突导向通路，膀胱癌通路，ECM-受体相互作用、破骨细胞分化、黏着力、肥厚

型心肌病（HCM）通路、精氨酸和脯氨酸代谢通路。

4 讨论

探索乳腺癌细胞基因与药物之间的相互关系以及作用机制，成为目前治愈癌症、寻找药物靶点以及新药研发等的热点和难点。在本研究中，就采取了基于二代测序技术的转录组学方法，最终确定了在月腺大戟素 A 作用下乳腺癌细胞内较为全面的、动态的基因含量变化过程。通过对转录组学数据进行深度挖掘分析，最终筛选出 1 695 种具有显著性差异的蛋白基因以及最有可能发生紊乱的细胞生命活动途径。采取生物分析软件对筛选出的具有显著性差异的基因进行信号通路显著性富集，分析后发现，图中得分较高、且位于信号通路相互关系网络中间关键部位的通路主要有 MAPK 信号通路、钙信号通路、PI3K-Akt 信号通路、NF-κB 信号通路、TNF 信号通路、p53 信号通路、细胞凋亡，而以上这些通路全部与乳腺癌细胞凋亡相关。

丝裂原活化蛋白激酶（mitogen-activated protein kinases，MAPK）信号通路，是一组进化保守的丝/苏氨酸蛋白激酶，它们会被一系列细胞外的刺激信号激活并介导信号从细胞膜向细胞核传导，它们调控着许多生理活动，如炎症、凋亡、癌化、肿瘤细胞的侵袭和转移等[25]。实验结果可以发现，MAPK 信号通路明显上调，表明在实验组中，乳腺癌细胞凋亡、癌化受到明显影响；研究发现，钙离子参与真核细胞跨膜信号转导途径，胞浆内钙离子作为第二信使，对细胞的生存与凋亡起着重要的调控作用，而钙信号通路上调，会导致细胞内钙离子内流增加，钙稳态失调，ROS 增加促进细胞凋亡[26-27]。磷脂酰肌醇 3-激酶（PI3Ks）蛋白家族参与细胞增殖、分化、凋亡和葡萄糖转运等多种细胞功能的调节，PI3K 活性的增加常与多种癌症相关[28-30]。活化的 Akt 通过磷酸化多种酶、激酶和转录因子等下游因子，进而调节细胞的功能，Akt 通过下游多种途径对靶蛋白进行磷酸化而发挥抗凋亡作用。Akt 通过激活 IκB 激酶（IKKα），导致 NF-κB 的抑制剂 IκB 的降解，从而使 NF-κB 从细胞质中释放出来进行核转位，激活其靶基因而促进细胞的存活[31-33]。Akt 能经磷酸化 p53 结合蛋白 MDM2 影响 p53 的活性，磷酸化的 MDM2 转位到细胞核与 p53 结合，增加 p53 蛋白的降解而影响细胞存活[34-38]；而 p53 基因是一种与肿瘤关系最密切的抑癌基因，参与细胞周期的调控和损伤 DNA 修复[39-40]。从本实验结果可以发现，p53 信号通路明显上调，表明在实验组中，乳腺癌细胞内 p53 基因参与的信号通路被上调，同时通过死亡信号受体蛋白途径刺激线粒体释放高毒性的氧自由基诱导细胞凋亡，调控 TNF 受体等。

从本实验转录组学数据和生信分析通路结果中可以得出，在月腺大戟素 A 的作用下乳腺癌细胞的病理、生理变化过程与上述一种或多种途径的上调和障碍密切相关，这也为接下来乳腺癌治疗的研究方向和靶点确立了范围，并打下了牢固的基础。

参考文献

[1] Feng R M, Zong Y N, Cao S M, *et al*. Current cancer situation in China: Good or bad news from the 2018 Global Cancer Statistics? [J]. *Null*, 2019, 39(1): 22.

[2] Rivera-Franco M M, Leon-Rodriguez E. Delays in breast cancer detection and treatment in developing countries [J]. *Breast Cancer Basic Clin Res*, 2018, doi: 10.1177/1178223417752677.

[3] 黄赛君, 葛 菲, 陈文林. MMP-1 在乳腺癌脑转移中的作用研究进展 [J]. 昆明医科大学学报, 2018, 39(12): 125-129.

[4] 余修中, 欧 华, 刘怀莉, 等. 5 种乳腺癌相关基因临床应用价值的探讨 [J]. 国际检验医学杂志, 2018, 39(20): 2514-2517.

[5] Nazir S U, Kumar R, Singh A, *et al*. Breast cancer invasion and progression by MMP-9 through Ets-1 transcription factor [J]. *Gene*, 2019, 716: 144013.

[6] Bray F, Ferlay J, Soerjomataram I, *et al*. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries [J]. *CA*: *Cancer J Clin*, 2018, 68(6): 394-424.

[7] 郑 莹. 乳腺癌在中国的流行状况和疾病特征 [J]. 中国癌症杂志, 2013, 23(8): 561-569.

[8] Bubnovskaya L, Kovelskaya A, Gumenyuk L, *et al*. Disseminated tumor cells in bone marrow of gastric cancer patients: Correlation with tumor hypoxia and clinical relevance [J]. *J Oncol*, 2014, 2014: 582140.

[9] Nounou M I, ElAmrawy F, Ahmed N, *et al*. Breast cancer: Conventional diagnosis and treatment modalities and recent patents and technologies [J]. *Breast Cancer*: *Basic Clin Res*, 2015, 9(Suppl 2): 17-34.

[10] 李 伟, 潘 燕, 李学军. HER2 阳性乳腺癌治疗药物曲妥珠单抗耐药机制及新一代靶向药物研究进展 [J]. 中国临床药理学杂志, 2014, 30(1): 48-51.

[11] 南京中医药大学. 中药大辞典 [M]. 上海: 上海科学技术出版社, 1977.

[12] 夏 青, 徐柯心, 张文婷, 等. 中药狼毒化学成分与药理作用概述 [J]. 环球中医药, 2017, 10(8): 1027-1032.

[13] 熊 爽. 月腺大戟化学成分的研究 [D]. 长春: 吉林大

学, 2009.

[14] 孙晓飞, 王淑萍, 郑泽荣. 月腺大戟化学成分研究 [J]. 中国中药杂志, 1999, 24(4): 34-35.

[15] 浮光苗, 余伯阳, 朱丹妮. 月腺大戟化学成分的研究 [J]. 中国药科大学学报, 2003, 34(4): 87-89.

[16] 颜秉强. 山东产月腺大戟生药学研究 [D]. 济南: 山东中医药大学, 2009.

[17] 李盛建, 王 莹, 王强利, 等. 月腺大戟素 A 抗乳腺癌活性 [J]. 第二军医大学学报, 2018, 39(7): 765-769.

[18] Tirosh I, Izar B, Prakadan S M, *et al*. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq [J]. *Science*, 2016, 352(6282): 189-196.

[19] Puram S V, Tirosh I, Parikh A S, *et al*. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer [J]. *Cell*, 2017, 171(7): 1611-1624.

[20] Lambrechts D, Wauters E, Boeckx B, *et al*. Phenotype molding of stromal cells in the lung tumor microenvironment [J]. *Nat Med*, 2018, 24(8): 1277-1289.

[21] Zepp, J A, Zacharias W J, Frank D B, *et al*. Distinct mesenchymal lineages and niches promote epithelial self-renewal and myofibrogenesis in the lung [J]. *Cell*, 2017, 170(6): 1134-1148.

[22] Boyle E I, Weng S, Gollub J, *et al*. GO: TermFinder-open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes [J]. *Bioinformatics,* 2004, 20(18): 3710-3715.

[23] Du J, Yuan Z, Ma Z, *et al*. KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model [J]. *Mol Biosystems*, 2014, 10(9): 2441-2447.

[24] Basu P, Cai T T, Das K, *et al*. Weighted false discovery rate control in large-scale multiple testing [J]. *J Am Stat Assoc*, 2018, 113(523): 1172-1183.

[25] El Zaoui I, Bucher M, Rimoldi D, *et al*. Conjunctival melanoma targeted therapy: MAPK and PI3K/mTOR pathways inhibition [J]. *Invest Ophthalmol Visual Sci*, 2019, 60(7): 2764-2772.

[26] Lecourieux D, Ranjeva R, Pugin A. Calcium in plant defence-signalling pathways [J]. *New Phytolt*, 2006, 171(2): 249-269.

[27] Li Z, Shi J, Hu D, *et al*. A polysaccharide found in Dendrobium nobile Lindl stimulates calcium signaling pathway and enhances tobacco defense against TMV [J]. *Int J Biol Macromol*, 2019, 137: 1286-1297.

[28] Aksoy E, Saveanu L. The isoform selective roles of PI3Ks in dendritic cell biology and function [J]. *Front Immunol*, 2018, doi: 10.3389/fimmu.2018.02574.

[29] Narayanankutty A. PI3K/Akt/mTOR pathway as a therapeutic target for colorectal cancer: A review of preclinical and clinical evidence [J]. *Curr Drug Targets*, 2019, doi: 10.2174/156800909789271521.

[30] Noorolyai S, Shajari N, Baghbani E, *et al*. The relation between PI3K/AKT signalling pathway and cancer [J]. *Gene*, 2019, doi: 10.1016/j.gene.2019.02.076.

[31] Zeng C X, Fu S B, Feng W S, *et al*. TCF19 enhances cell proliferation in hepatocellular carcinoma by activating the ATK/FOXO1 signaling pathway [J]. *Neoplasma*, 2019, 66(1): 46-53.

[32] Liu H Y, Zhang Y Y, Zhu B L, *et al*. MiR-203a-3p regulates the biological behaviors of ovarian cancer cells through mediating the Akt/GSK-3β/Snail signaling pathway by targeting ATM [J]. *J Ovar Res*, 2019, doi: 10.1186/s13048-019-0532-2.

[33] Hao Y, Liu J, Wang Z, *et al*. Piceatannol protects human retinal pigment epithelial cells against hydrogen peroxide induced oxidative stress and apoptosis through modulating PI3K/Akt signaling pathway [J]. *Nutrients*, 2019, doi: 10.3390/nu11071515.

[34] Hay J, Gilroy K, Huser C, *et al*. Collaboration of MYC and RUNX2 in lymphoma simulates T-cell receptor signaling and attenuates p53 pathway activity [J]. *J Cell Biochem*, 2019, 120: 18332-18345.

[35] Ma Z, Guo D, Wang Q, *et al*. Lgr5-mediated p53 Repression through PDCD5 leads to doxorubicin resistance in hepatocellular carcinoma [J]. *Theranostics*, 2019, 9(10): 2967-2983.

[36] Gao Y, Yin H, Zhang Y, *et al*. Dexmedetomidine protects hippocampal neurons against hypoxia/reoxygenation-induced apoptosis through activation HIF-1α/p53 signaling [J]. *Life Sci*, 2019, 232: 116611.

[37] Yu J, Wang S, Zhang Y, *et al*. TRIM67 activates p53 to suppress colorectal cancer initiation and progression [J]. *Cancer Res*, 2019, 79(16): 4086–4098.

[38] Wang J, Wang C, Yu H, *et al*. Bacterial quorum sensing signal IQS induces host cell apoptosis by targeting POT1-p53 signalling pathway [J]. *Cell Microbiol*, 2019, 21(10): e13076.

[39] Zhao W, Li H, Tang Y, *et al*. Fluorometric determination of the p53 cancer gene using strand displacement amplification on gold nanoparticles [J]. *Mikrochim Acta*, 2019, 186(8): 517.

[40] Di Agostino S, Fontemaggi G, Strano S, *et al*. Targeting mutant p53 in cancer: The latest insights [J]. *J Exper Clin Cancer Res: CR*, 2019, 38(1): 290.