《分子植物育种》网络版，2020 年，第 18 卷，第 46 篇，第 1-7 页

研究报告
Research Report

洋桔梗(*Eustoma grandiflorum*)干旱胁迫转录组初步分析

安霞[1*] 朱强[1] 楼旭平[2] 李鲁峰[2] 陈杰[3] 柳婷婷[1] 李文略[1] 骆霞虹[1] 朱关林[1] 余利隽[1]
1 浙江省萧山棉麻研究所, 浙江省园林植物与花卉研究所, 杭州, 311251; 2 杭州市萧山区农业科学技术研究所, 杭州, 311202; 3 华中农业大学, 武汉, 430070
* 通信作者, anxia@zaas.ac.cn

摘 要 洋桔梗(*Eustoma grandiflorum*)是一种原产于北美的重要观赏植物。目前为止，关于洋桔梗的基础分子研究报道数量相对较少，特别是响应干旱胁迫的分子机制相关研究。本实验利用转录组二代测序技术对干旱胁迫下的洋桔梗幼苗进行了研究。结果表明，该转录组测序数据具有良好的质量控制结果。为了在此基础上获得基因信息，利用 Trinity 和 Corset 等软件将序列进行了拼接，共得到 102 014 条非冗余基因，其中包含 2 929 条编码基因，经分析发现分属于 79 个转录因子家族。对所有基因进行注释，发现比对相似度最高的物种是中果咖啡。最后，对基因序列进行简单序列重复(Simple Sequence Repeat, SSR)分析，获得了所有非冗余基因以及转录因子编码基因所包含的 SSR 信息。本研究结果可为后续针对洋桔梗响应干旱胁迫的相关研究提供候选分子资源。

关键词 洋桔梗, 干旱胁迫, 转录组

Transcriptome Profiling of Lisianthus (*Eustoma Grandiflorum*) under Drought Stress

An Xia[1*] Zhu Qiang[1] Lou Xuping[2] Li Lufeng[2] Chen Jie[3] Liu Tingting[1] Li Wenlue[1] Luo Xiahong[1] Zhu Guanlin[1] Yu Lijun[1]
1 Zhejiang Xiaoshan Institute of Cotton and Bast Fiber Crops Research, Zhejiang Institute of Landscape Plants and Flowers, Hangzhou, 311251; 2 Hangzhou Xiaoshan District Agricultural Science and Technology Research Institute, Hangzhou, 311202; 3 Huazhong Agricultural University, Wuhan, 430070
* Corresponding author, anxia@zaas.ac.cn
DOI: 10.5376/mpb.cn.2020.18.0046

Abstract Lisianthus (*Eustoma grandiflorum*) is an important ornamental plant native to North America. So far, the number of basic molecular research reports on Lisianthus is relatively small, especially the research related to the molecular mechanism of response to drought stress. In this experiment, transcriptome second-generation sequencing technology was used to study Lisianthus seedlings under drought stress. The results showed that the transcriptome sequencing data had good quality control results. In order to obtain genetic information on this basis, the sequences were spliced using software such as Trinity and Corset, a total of 102 014 non-redundant genes were obtained, including 2 929 coding genes which were found to belong to 79 transcription factor families after analyzing. Annotating all genes, Zhongguo coffee was found to be the species with the highest comparison similarity. Finally, simple sequence repeat (SSR) analysis was performed on the gene sequences, and all SSR

收稿日期：2020 年 10 月 27 日；接受日期：2020 年 10 月 27 日；发表日期：2020 年 11 月 3 日
引用格式: 安霞, 朱强, 楼旭平, 李鲁峰, 陈杰, 柳婷婷, 李文略, 骆霞虹, 朱关林, 余利隽, 2020, 洋桔梗(*Eustoma grandiflorum*)干旱胁迫转录组初步分析, 分子植物育种(网络版), 18(46): 1-7 (doi: 10.5376/mpb.cn.2020.18.0046) (An X., Zhu Q., Lou X.P., Li L.F., Chen J., Liu T.T., Li W.L., Luo X.H., Zhu G.L., and Yu L.J., 2020, Transcriptome profiling of Lisianthus (*Eustoma grandiflorum*) under drought stress, Fengzi Zhiwu Yuzhong (Molecular Plant Breeding (online)), 18(46): 1-7 (doi: 10.5376/mpb.cn.2020.18.0046))

information contained in non-redundant genes and transcription factor coding genes had been obtained. The results of this study can provide candidate molecular resources for subsequent studies on Lisianthus 's response to drought stress in the future.

Keywords Lisianthus, Drought stress, Transcriptome

洋桔梗(*Eustoma grandiflorum*)别名草原龙胆，原产北美地区，是龙胆科(Gentianaceae)多年生草本植物。洋桔梗是重要的观赏植物，其花器官形态漂亮并且瓶插寿命长，已经成为国际越来越重要的鲜切花之一。洋桔梗对栽培环境十分敏感，国内栽培设施及技术相对不足。对洋桔梗生长发育过程中的相关分子机理进行研究，有助于为洋桔梗栽培新技术开发提供理论支持。

植物在响应干旱胁迫的过程中涉及一些功能基因和调节基因的差异表达，形成了一套复杂的信号调控网络，从而影响植物体内一系列的生理生化反应。干旱能够影响洋桔梗花茎伸长，但是目前为止关于洋桔梗如何响应干旱胁迫的相关研究鲜有报导。本研究通过对干旱胁迫处理下的洋桔梗植株进行转录组测序，共获得 6.43 Gb 高质量测序数据。对该测序结果进行拼接，获得了 102 014 条基因。对所获得的基因进行注释，发现和现有其它研究基础较好的物种之间亲缘关系都较远。因此，该项研究不仅能为后续针对洋桔梗的相关研究提供大量分子资源，还能为龙胆科其它物种高通量测序研究提供借鉴。

1 结果与分析

1.1 转录组数据获得、处理与拼接

首先对本研究中的转录组测序数据进行严格的质量控制，主要去除接头以及低质量序列信息，总共获得高质量测序数据(Clean reads) 6.43 Gb，然后采用 Trinity 软件(Grabherr et al., 2011)对本研究中获得的这些高质量测序数据进行拼接后，共得到 132 929 条转录本(Transcripts)。随后，使用软件 Corset 对获得的高质量测序数据进行分析，将所有 reads 与转录本进行比对并进行层次聚类，然后得到 102 014 个非冗余的基因(Unigenes)。这些转录本和基因在序列长度方面具有相似分布规律(图 1A; 图 1B)，且转录本与非冗余基因之间的数量差异主要体现在较短序列(<500 bp)上(图 1C)。该结果表明本次测序结果质量较高，主要表现为在较长序列(>1 000 bp)范围内，转录本与非冗余基因数目没有显著差别。

1.2 基因注释和功能分类

为了更科学系统地分析转录组测序获得的序列所涉及的基因功能信息，本研究对拼接得到的 102 014 条基因，分别采用不同的公共数据库对获得的基因进行注释。在不同数据库中获得注释基因的统计信息见表 1，其中有 79.94%的基因在至少一个数据库中有注释结果，而来源于 Nr 数据库的注释结果最多，占所有基因的 76.52%。选取基因在 Nr、Nt、pfam、GO 和 KOG 五个数据库的注释结果进行分析表明，特异在 Nr 数据库中得到注释信息的基因数有

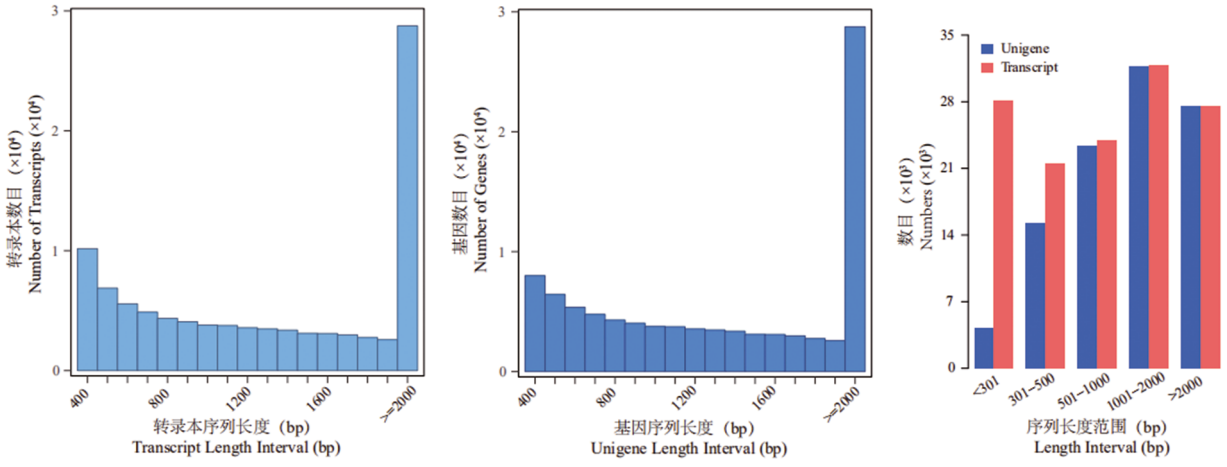

图 1 转录组序列长度分布

注: A: 转录本序列长度分布; B: 基因长度分布; C: 转录本和基因长度分布统计

Figure 1 Lengths distribution from transcriptomic data

Note: A: Lengths distribution of transcripts; B: Lengths distribution of unigenes; C: Statistical results of transcripts and unigenes

13,035 条，远远多于 Nt 数据库中特异注释的 741 条基因和 KOG 数据库中特异注释的 8 条基因，在 pfam 和 GO 数据库中没有特异注释的基因(图 2A)。因此，转录组数据在 Nr 数据库中的注释信息更全面。进一步分析 Nr 数据库注释结果，有接近一半(47.3%)的序列与目标序列具有较高相似度(超过 80%, 图2B)，且大量序列(占比 60.3%)的比对结果 e 值小于 1e-60 (图 2C)。这些序列物种注释结果中，比对到最多的物种为中果咖啡(*Coffea canephora*, 占比 39.7%)，并且有更大比例的序列(44.1%)比对到其它(Other)物种(图 2D)。在 KOG 数据库分类中，有 25 个 KOG 类别被不同数量基因所注释，共包含 28 696 条基因。其中被注释基因最多的两个类别是 O: 翻译后修饰：蛋白开关和分子伴侣(3 519 条基因)和 R: 总体功能预测(3 421 条基因) (图 3)，该结果与黄麻干旱胁迫转录组结果类似。与此同时，从代谢角度来看，这些基因被更多富集在“遗传信息处理”大类中的“翻译”、“代谢”大类中的“碳水化合物代谢”和“遗传信息处理”中的“折叠、排列和降解”等三个代谢条目中(图 4)。最后，由于转录因子往往处于基因表达通路的上游，能够调控下游一系列基因表达从而在更大程度上影响洋桔梗响应干旱胁迫的程度。与之对应，本次转录组测序结果中共有 2 929 条基因可能编码转录因子，这些转录因子属于 79 个不同的转录因子家族 (图5)。在这些转录因子家族中，包含预测基因数目最多的三个家族分别是 bHLH 家族 (256 条基因) MYB_related 家族(216 条基因)和 bZIP 家族(191 条基因)。

1.3 分子标记开发

通过对转录组测序所获得的序列进行简单重复

表 1 基因注释成功率统计

Table 1 Statistical numbers on successfully annotated genes against different databases

数据库 Databases	NR	NT	KO	SwissProt	PFAM	GO	KOG
基因数目 Number of Genes	78 070	50 844	34 057	60 288	55 628	55 628	25 389
百分比(%) Percentage (%)	76.52	49.84	33.38	59.09	54.52	54.52	24.88

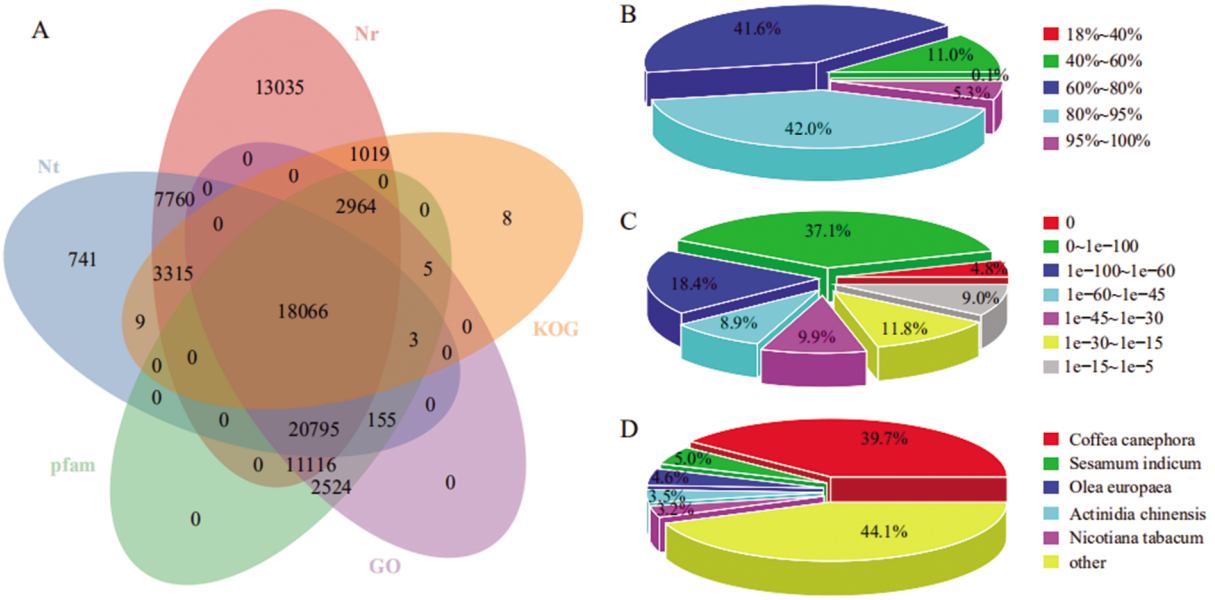

图 2 转录组注释信息

注: A: 转录组结果比对到不同数据库的基因数目; B: Nr 数据库比对结果序列相似度分布; C: Nr 数据库比对结果 e 值分布; D: Nr 数据库注释比对到最多的物种

Figure 2 Annotation of transcriptome

Note: A: Numbers of unigenes annotated by different databases; B: Distribution of sequence similarities against the Nr database; C: Distribution of e-values against the Nr database; D: The most annotated species from our transcriptome data

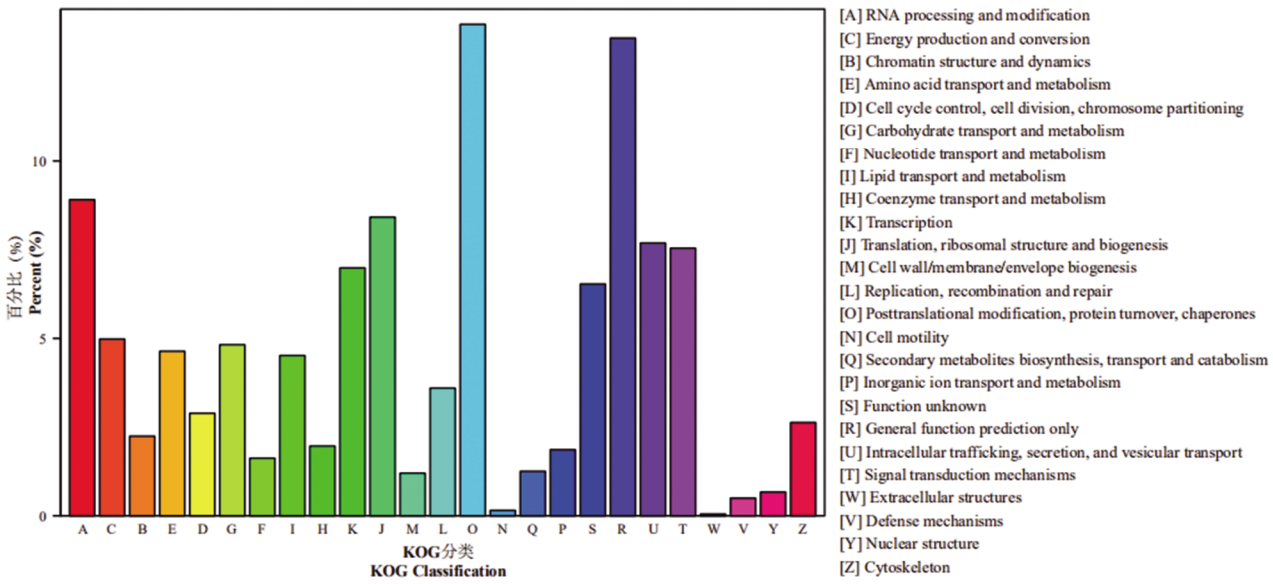

图 3 转录组结果的 KOG 分类

Figure 3 The KOG classification of our transcriptome data

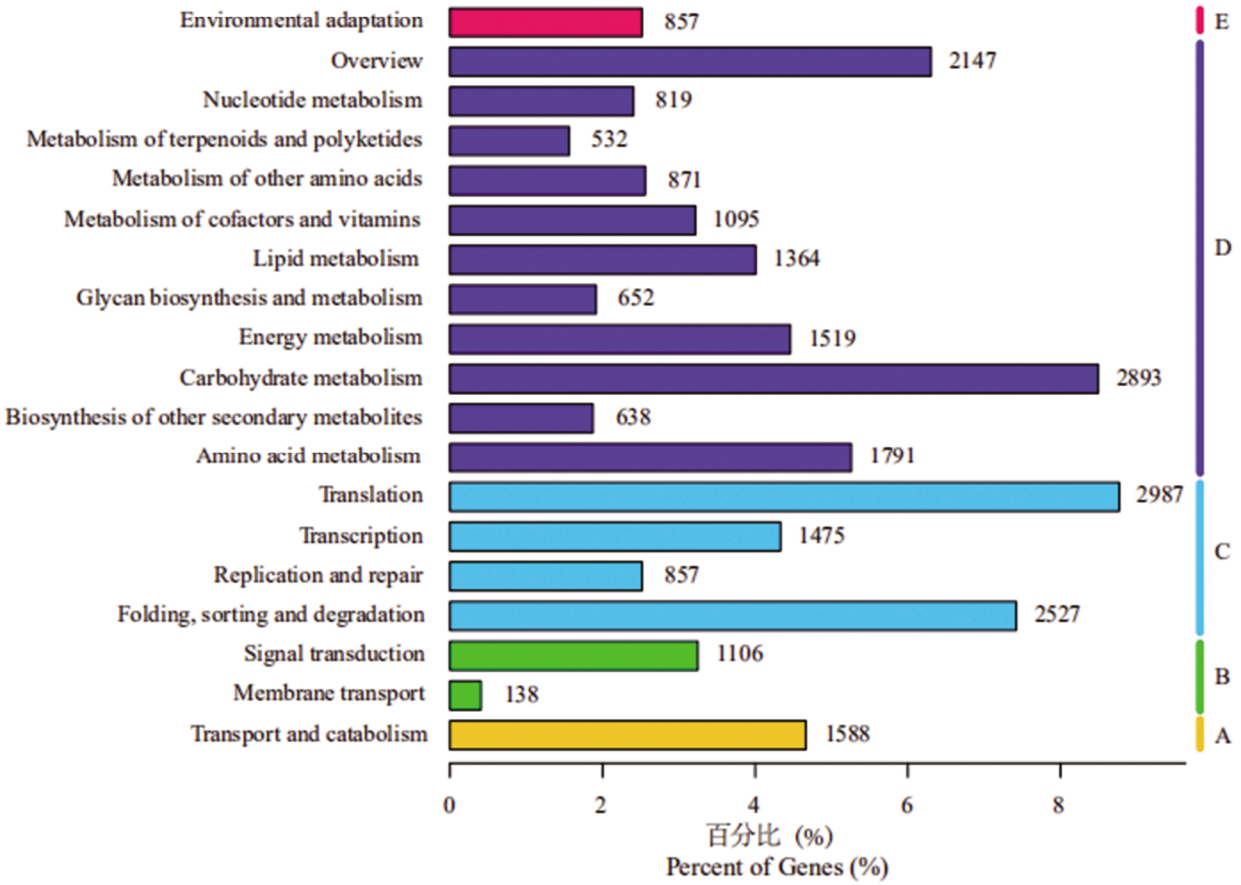

图 4 转录组结果的 KEGG 分类

注: 共分为五大类; A: 细胞进程; B: 环境信息处理; C: 遗传信息处理; D: 代谢; E: 有机系统

Figure 4 The KEGG classification of our transcriptome data

Note: A total of five categories were included: A, Cellular processes; B, Environmental information processing; C, Genetic information processing; D: Metabolism; E: Organismal systems.

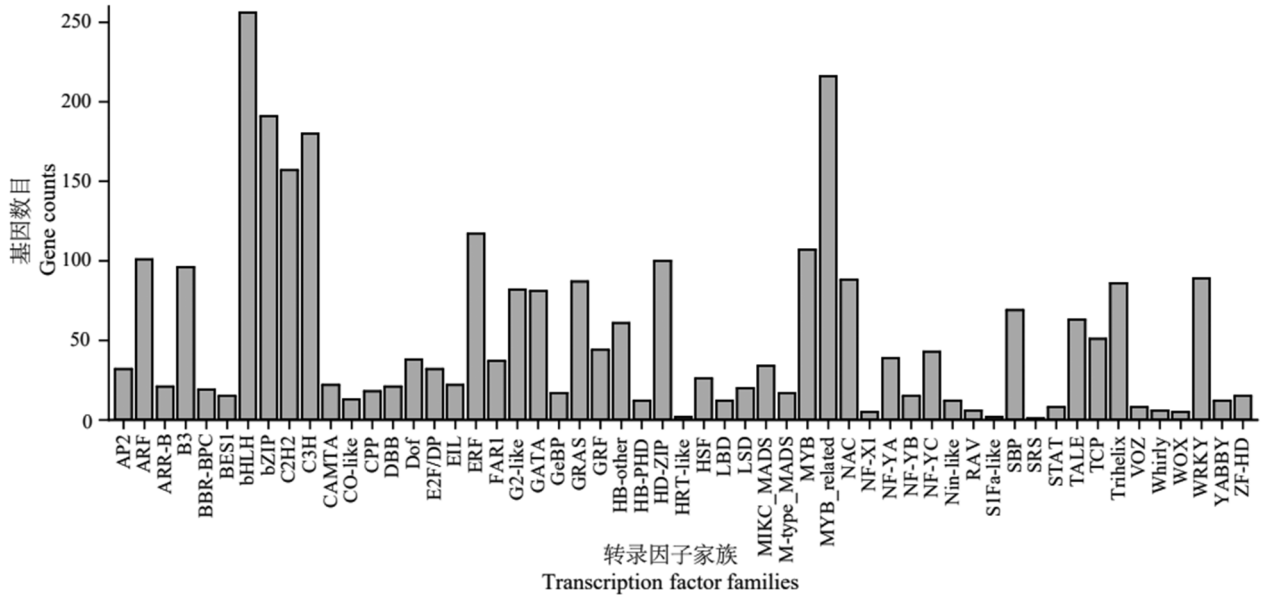

图 5 转录因子数目统计

Figure 5 Numbers of transcription factors from different families

序列(simple sequence repeat, SSR)分析，共在 21 329 条基因序列(占所有基因序列的 20.91%)中发现了 25 468 个 SSR。这些 SSR 序列主要包括单碱基至六碱基不同程度的重复，以及复杂重复序列。如图 6A 所示，除了二碱基重复以外，SSR 重复序列平均总长度随着重复单位的复杂性增加而递增，其中复杂重复单元的重复序列长度最长。在编码转录因子的 2,929 条基因中，有 833 条序列(占比 28.44%)具有不同的 SSR，这些 SSR 序列总长度也具有前述类似规律(图 6B)。最后分析 SSR 可能位于基因的不同位置，表明对于所有基因来说，超过一半(51.46%)的重复序列可能横跨相邻的两个基因结构(5’非翻译区: utr5; 编码区: cds; 3’非翻译区: utr3)，而位于编码区的 SSR 所占比例最少(图 6C)。对于转录因子编码基因来说，横跨两个基因结构的 SSR 大幅度减少(27.25%)，而位于编码区的重复序列所占比例依然最少(图 6C)。后

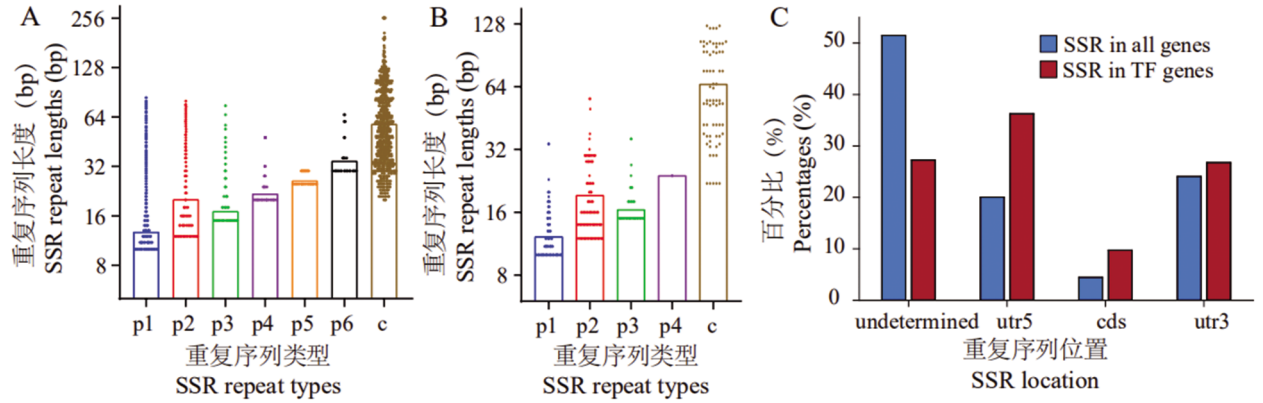

图 6 简单序列重复(SSR)信息统计

注: A: 转录组中 SSR 序列长度统计; B: 转录因子 unigene 中 SSR 序列长度统计; C: SSR 位于不同基因功能区域统计; 重复序列类型包括单碱基重复至六碱基重复(p1-p6)以及复杂重复单元(c); 这些 SSR 可能位于基因的 5’非翻译区(utr5), 编码区(cds), 3’非翻译区(utr3)或者未知位置(undetermined)。

Figure 6 Statistical on simple sequence repeats (SSR)

Note: A: Sequence lengths distribution of SSRs amongst the whole transcriptome data; B: Sequence lengths distribution of SSRs from transcription factor coding genes; C: Location of SSRs on varied districts of unigenes; The SSR units included mononucleotides to hexanucleotides (p1 to p6), and complex units (c); These SSRs may locate on the 5’ untranslated regions (utr5), 3’ untranslated regions (utr3), coding sequences (cds), or currently unknown positions (undetermined).

续可以通过开发这些 SSR 的特异引物，对特定基因，或者转录因子进行针对性更强的检测和研究。

2 讨论

洋桔梗是重要的观赏植物。然而，针对该物种的分子生物学研究基础较欠缺。目前为止，仅在早期构建过一个针对盐胁迫的差减文库 (王继刚等, 2008)，鉴定了可能的差异性表达基因。然而，差减文库的通量一般较低，和现行高通量测序相比，远远不能满足研究的需求。在分子资源开发方面，也仅有早期针对花期进行的转录组测序(Kawabata et al., 2012)。在该项转录组测序中，所得到的 63 401 条 contig 仅有 65%在 NCBI 数据库中得到了比对结果，小于本研究中的 76.52% (表 1)，说明随着测序技术的发展和拼接方法的成熟，本次测序结果的序列注释情况得到了提高。然而，可能是由于与洋桔梗的近源物种分子研究基础均较薄弱，本次转录组注释结果，比对到最高比例的物种(图 2D)是茜草科(Rubiaceae)的中果咖啡(Coffea canephora)。茜草科和龙胆科同属龙胆目，因此该中果咖啡可能是和洋桔梗亲缘关系最近的有一定分子研究资源的物种。除此之外，洋桔梗转录组序列的比对结果零散地分布在其他物种上(图 2D)。

曾经有研究对洋桔梗 MADS 家族基因进行鉴定(Ishimori and Kawabata, 2014)和功能研究(Li et al., 2015)。然而，在没有转录组或者基因组等高通量测序结果的支持下，进行基因功能研究，或者基因家族鉴定往往显得更加困难(Nakano et al., 2011)。本研究对洋桔梗在干旱胁迫下进行了转录组测序和并对测序数据进行初步分析，相关研究结果为后期研究洋桔梗响应干旱胁迫处理的分子机理提供相应数据支撑。与此同时，也有研究者完成了洋桔梗质体组测序(Yan et al., 2019)，相关分子资源为通过质体遗传信息调节花形态等农艺性状(Jin and Daniell, 2015)提供了相应基础。该质体组测序结果(Yan et al., 2019) 与龙胆科其它物种对应测序结果序列比对显示，洋桔梗与这些物种亲缘关系都更远。该结果从侧面印证了本次转录组测序注释结果中，比对到最多的物种是龙胆目茜草科下的中果咖啡(占 39.7%)，而其余大部分序列信息均零散地比对到其它物种(图 2D)。最后，本次转录组测序结果都将为后续研究，如鉴定挖掘干旱响应相关基因或者干旱胁迫相关的转录因子提供了分子数据。

3 材料与方法

3.1 植物材料

洋桔梗(*Eustoma grandiflorum*)品种“雪莱”在市场上购买。植株长至 8 cm 左右，对其进行干旱胁迫处理。处理 36 h 后，使用液氮取全株植物样品，用于总 RNA 提取。

3.2 总 RNA 提取及文库构建

样品在用于 RNA 提取之前一直保存于 -80℃超低温冰箱内。将样品取出并在液氮环境下充分研磨成粉末，使用天根公司的 RNA 提取试剂盒完成总 RNA 提取并用于构建转录组文库。

3.3 测序数据处理及转录本拼接

样品上机测序得到的直接数据为原始读数(Raw reads)，需要进行质量控制，去除不确定碱基测序结果和测序信息由于包含大于 10%的接头从而导致质量不佳等，余下读数即为高质量测序信息。对高质量测序信息的处理方式主要为序列拼接(Grabherr et al., 2011)和层次聚类，所得到的基因信息即为非冗余基因。

3.4 基因功能注释和分类

对经由上述步骤得到的非冗余基因，与四个序列数据库进行比对以进行注释。

3.5 转录因子预测

在对非冗余基因序列进行注释的同时，使用在线工具 (http://planttfdb.gao-lab.org/prediction.php)可以对这些基因信息可能编码产物进行预测，若编码转录因子则对其进行分类。

3.6 SSR 分析

在基因的碱基序列层面，存在一些规律明确的序列特征。这些序列往往以简单的序列单元(单个到多个碱基, 甚至复杂碱基单元) 为基础，重复出现多次，成为简单序列重复(SSR)。对于这些 SSR 一般使用在线工具(http://pgrc.ipk-gatersleben.de/misa/misa.html)进行预测。对于单碱基单元重复十次及以上，以二碱基为单元重复六次及以上，以及三碱基至六碱基为重复单位重复五次及以上的 SSR 均包含在统计范围内，而复杂重复序列中如果包含以上所列不同的重复单元，则每个重复单元分别满足上述要求。

作者贡献

安霞是本研究的实验设计和实验研究的执行

人，完成数据分析，论文初稿的写作；朱强、楼旭平、李鲁峰、陈杰、柳婷婷、李文略、骆霞虹、朱关林、余利隽是实验设计参与者；安霞是项目的构思者及负责人，指导实验设计，数据分析，论文写作与修改。全体作者都阅读并同意最终的文本。

致谢

本研究由省科技特派员项目(梯田景区农家乐景观提升示范与休闲产品创意)资助。

参考文献

Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D. A., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen Nir., Gnirke A., Rhind N., Palma F.D., Birren B.W., Chad Nusbaum., Lindblad-Toh K., Friedman N., and Regev A., 2013, Trinity: reconstructing a full-length transcriptome without a genome from rna-seq data, Nature Biotechnology, 29(7): 644-652

Ishimori M., and Kawabata S., 2014, Conservation and Diversification of Floral Homeotic MADS-box Genes in Eustoma grandiflorum, J. Japan, Soc. Hort. Sci., 83(2): 172-180

Jin S., and Daniell H., 2015, The engineered chloroplast genome just got smarter, Trends in Plant Science, 20(10): 622-640

Kawabata S., Li Y., and Miyamoto K., 2012, EST sequencing and microarray analysis of the floral transcriptome of Eustoma grandiflorum, Scientia Horticulturae, 144: 230-235

Li K.H., Chuang T.H., Hou C.J., and Yang C.H., 2015, Functional analysis of the FT Homolog from Eustoma grandiflorum reveals its role in Regulating A and C Functional MADS Box genes to control floral Transition and flower formation, Plant Mol. Biol. Rep., 33(4): 770-782

Nakano Y., Kawashima H., Kinoshit T., Yoshikawa H., and Hisamatsu T., 2011, Characterization of FLC, SOC1 and FT homologs in Eustoma grandiflorum: effects of vernalization and post-vernalization conditions on flowering and gene expression, Physiologia Plantarum, 141(4): 383-393

Wang J.G., Zhang K., Xu Q.J., and Li Y.H., 2008, Construction and Analysis of Eustoma grandiflorum Subtracted cDNA Library, Yuanyi Xuebao (Acta Horticulturae Sinica), 35(7): 1075 -1080. (王继刚, 张坤, 徐启江, 李玉花, 2008, 草原龙胆盐胁迫差减文库的构建及分析, 园艺学报, 35(7): 1075-1080)

Yan J., Cao Q., Wu Z., Chen S., Wang J., Zhou D., and Xie J., 2019, Complete plastome sequence of Eustoma grandiflorum (Gentianaceae), a popular cut flower, Mitochondrial DNA Part B, 4(2): 3163-3164