基因组学与应用生物学，2019 年，第 38 卷，第 12 期，第 5516−5526 页

研究报告
Research Report

中国沙棘雌雄株叶片 RNA-Seq 转录组分析

周武 [1,2,4]　胡娜 [1,3]　刘晓彤 [2]　王煜伟 [2]　索有瑞 [1,2,3*]

1 中国科学院藏药研究重点实验室，中国科学院西北高原生物研究所，西宁，810008；2 省部共建三江源生态与高原农牧业国家重点实验室，青海大学，西宁，810016；3 青海省藏药研究重点实验室，西宁，810008；4 中国科学院大学，北京，100049

* 通信作者，yrsuo@nwipb.cas.cn

摘　要　应用 Illumina HiSeq 测序技术，对中国沙棘雌雄株叶片分别进行了转录组测序，实验共获得 48.31 G 有效数据，平均错误率为 0.015%。经 Trinity 软件混合拼接后共得到 320 876 条 Transcripts 和 187 362 条 Unigenes，平均长度分别为 808 bp 和 588 bp，最大长度均为 17 291 bp。将 Unigenes 与公共数据库进行比对，得到注释的 Unigenes 为 104 926 条，占总数的 56%。转录组数据经搜索共得到 33 248 个微卫星标记。中国沙棘雌株和雄株转录组基因表达量差异比较数据显示，雌雄样本间表达差异基因共 92 970 个。上述工作不仅为中国沙棘的基因克隆和基因挖掘提供了基础数据，也为初步阐明中国沙棘性别决定的分子机制打下了基础。

关键词　中国沙棘, 转录组, SSR, 差异基因

The RNA-Seq Transcriptome Analysis of Female and Male Plants Leaves in *Hippophae rhamnoides* L. Subsp. *Sinensis*

Zhou Wu [1,2,4]　Hu Na [1,3]　Liu Xiaotong [2]　Wang Yuwei [2]　Suo Yourui [1,2,3*]

1 Key Laboratory of Tibetan Medicine Research, Northwest Institute of Plateau Biology, Chinese Academy of Sciences, Xining, 810008; 2 State Key Laboratory of Plateau Ecology and Agriculture, Qinghai University, Xining, 810016; 3 Qinghai Provincial Key Laboratory of Tibetan Medicine Research, Xining, 810008; 4 University of the Chinese Academy of Sciences, Beijing, 100049

* Corresponding author, yrsuo@nwipb.cas.cn

DOI: 10.13417/j.gab.038.005516

Abstract　In this study, Illumina HiSeq high-throughput sequencing technology was applied to get the transcriptome from leaves of female and male *Hippophae rhamnoides* L. subsp. *sinensis*, and 48.31 Gb effective data was obtained with an average error rate of 0.015%. After assembly by the software Trinity, a total of 320 876 unigenes and 187 362 transcripts were generated, corresponding to maximum length of 17 291 bp both, the average length of 588 bp and 808 bp respectively. Using Blastx against the public databases, 104 926 unigene were annotated at least one databases, accounting for 56% of the total. In addition, 33 248 potential SSR loci were detected throughout the transcriptomic data. The gene expression analysis between the two transcriptomes revealed that 92 970 genes expressed differentially between male and female samples. This work provides basic data for the gene cloning and gene mining, as well as to investigate the sex-determining mechanism of dioecious *Hippophae rhamnoides* L. subsp. *sinensis*.

Keywords　*Hippophae rhamnoides* L. subsp. *sinensis*, Transcriptome, SSR, Differentially expressed genes

沙棘，是沙棘果实和整个植株的通俗称呼和统称。沙棘属植物沙棘从植物分类学上属于桃金娘目胡颓子科沙棘属，是一种落叶性灌木(廉永善和陈学林, 1996)，其适应能力很强。中国沙棘是沙棘的亚种，也是最主要

基金项目：本研究由青海省科技厅项目(2015−NK−509, 2017−SF-A8, 2017−ZJ-Y11)和中国科学院西部之光青年学者 B 类项目共同资助

引用格式：Zhou W., Hu N., Liu X.T., Wang Y.W., and Suo Y.R., 2017, The RNA-Seq transcriptome analysis of female and male plants leaves in *Hippophae rhamnoides* L. Subsp. *sinensis*, Jiyinzuxue Yu Yingyong Shengwuxue (Genomics and Applied Biology), 38 (12): 5516-5526 (周武, 胡娜, 刘晓彤, 王煜伟, 索有瑞, 2019, 中国沙棘雌雄株叶片 RNA-Seq 转录组分析, 基因组学与应用生物学, 38(12): 5516-5526)

品种，分布和种植面积最为广泛。沙棘是药食同源且具有防风固沙等多种用途的特色经济植物，富含多种矿物质和维生素，素有“维生素宝库”之称(鲁长征等, 2008)。沙棘具有祛痰利肺、消食开胃、活血散瘀、消炎止痛等功效(陈卫平和李毅敏, 2005; 包桂兰等, 2009)。中国是沙棘资源大国，其分布面积占世界沙棘总面积的 95%以上。青海是沙棘的种群发源地和原料主产区，享有“世界沙棘看中国，中国沙棘看青海”的美誉，现有沙棘资源 15.34 万公顷，占全国沙棘资源总量的 10%左右(利毛才让, 2012)。经过现代医学研究证实，沙棘具有抗缺氧、抗疲劳、防辐射等医疗保健潜力(Goel et al., 2005; Narayanan et al., 2005; Shukla et al., 2006; 逄蕾, 2009; Ni et al., 2013)。由于青海的高海拔和特殊的气候环境，造就了青海沙棘富含多类活性物质且物质活性强、有效成分含量高的特点，具备广阔的开发前景和经济价值。

RNA-Seq (转录组测序技术)借助逆转录酶，首先将特定组织或细胞中 mRNA 为模板，以 6 碱基随机引物反转录单链 cDNA，继而通过 PCR 扩增和纯化得到高质量文库，最终通过高通量测序，获得 RNA 序列信息。基于高通量转录数据，可以通过比较不同组织(细胞)的读段(Reads)数差异来鉴别出差异表达基因；借助于强大的基因注释数据，还可以识别并发现新的转录本、可变剪切和基因结构变异等(祁云霞等, 2011)。转录组测序是目前较为成熟、效费比高且应用广泛的高通量测序技术，能全面而快速地反映某一组织(细胞)在某一特定时空下的的转录组数据。如通过不同生境(贫瘠和肥沃土壤)、不同处理(正常组, 模型组和给药组)和不同发育阶段(幼虫, 成虫, 结茧和飞蛾)中转录组数据对比分析研究可为揭示其差异形成的机制指明研究的方向。转录组数据的拼接和组装可以不依赖参考基因组，这种技术优势为非模式植物功能基因的研究和深度挖掘提供了有效手段(姜福星等, 2017; 刘杰等, 2017)。通过 RNA-Seq测序不仅可以获得大量转录本序列信息，同时这些序列也为分析微卫星 SSR (Simple sequence repeats)和单核苷酸多态性 SNP (Single nucleotide polymorphisms)等分子标记提供了基础数据。中国沙棘是一种典型的雌雄异株植物(Ainsworth, 2000)，为沙棘进化和演替的原始类群，是研究植物雌雄性别进化的合适材料。本研究对中国沙棘雌雄株叶片进行转录组测序分析，组装 Unigene 数据库，通过对比分析中国沙棘雌雄株中差异表达基因，以期将来能筛选出参与调控中国沙棘性别决定的候选基因，为将来揭示中国沙棘性别决定和分化的分子机制打下工作基础。

1 结果与分析

1.1 中国沙棘叶片总 RNA 提取

分析 1%琼脂糖凝胶电泳结果(图 1)，28 S 和 18 S 条带明亮，无明显降解，表明提取的中国沙棘叶片 RNA 完整性较好；经过 Agilent 2100 生物分析仪、Nanodrop 2000 超微量分光光度计检测和 Qubit 2.0 Fluorometer 荧光定量仪精确定量表明，RNA 样品的完整性、纯度和浓度均达到符合 HiSeq-PE150 转录组测序要求，所有样品 RIN 值均在 8.0 以上，$OD_{260/280}$ 为 2.0~2.1，28 S/18 S 为 1.7~2.4，RNA 样品浓度>300 ng/μL，RNA 总量>9 μg (表 1)。

1.2 测序数据质量

通过 Illumina HiSeq-PE150 高通量测序，将拍照捕获的 BCL 格式荧光图像文件经 CASAVA 碱基识别转化为原始测序数据(Sequenced Reads)，该数据由描述性文字、碱基序列、测序标识符和对应碱基的测序质量信息四部分组成。测序所得的原始数据，由于建库方法、测序仪器自身和测序方法的原因，序列中会含建库时加入的接头(Adapter)的和一小部分低质量的 Reads (测序碱基质量值 Qphred≤20 的碱基数占 Reads 一半以上)。为了防止后期信息分析出现错误，必须删去带接头的和低质量 Reads，还要删去未明确碱基类型比例大于千分之一的 Reads，这种去除无效数据的 Reads 称为清洁数据(Clean reads)，后续所有分析都基于清洁数据。各 3 份中国沙棘雄株和

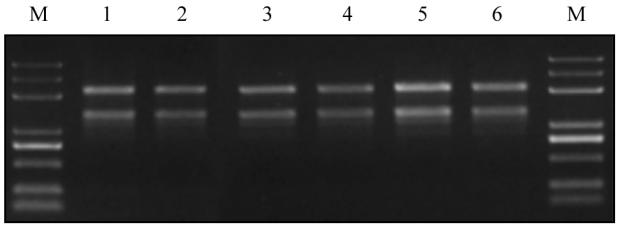

图 1 中国沙棘雌雄株叶片总 RNA 的琼脂糖凝胶电泳分析

注: M: Trans 2000 Plus Marker; 1,6: Male−1, Female−3 稀释 3 倍, 上样 1 μL; 2~4: Male−2, Male−3, Female−1 稀释 5 倍, 上样 1 μL; 5: Female−2 稀释 10 倍, 上样 1 μL

Figure 1 Agarose gel electrophoresis analysis of total RNA in female and male leaves of *Hippophae rhamnoides* L.

Note: M: Trans 2000 Plus Marker; 1 and 6: 3 times dilution of the sample Male−1 and Female−3, 1 μL volumes was uploaded; 2~4: 5 times dilution of the sample Male−2, Male−3 and Female−1, 1 μL volumes was uploaded; 5: 10 times dilution of the sample Female−2 , 1 μL volumes was uploaded

表 1 中国沙棘总 RNA 质量检测

Table 1 Quality test of total RNA in *Hippophae rhamnoides* L.

样品 Sample	浓度(ng/μL) Concentration (ng/μL)	总量(μg) Total quantity (μg)	$OD_{260/280}$	$OD_{260/230}$	28S/18S	RIN 值 RIN value
Male-1	474	15.16	2.06	2.06	1.7	8.2
Male-2	514	16.44	2.07	1.97	2.4	8.2
Male-3	624	19.96	2.06	2.16	1.8	8.3
Female-1	416	13.31	2.12	2.10	1.9	8.2
Female-2	1 242	39.74	2.10	2.29	1.9	8.3
Female-3	300	9.60	2.08	1.95	2.3	8.2

雌株转录组双端测序质量和数量统计(表 2)。各样品清洁数据碱基数均在 6.73 G 以上，共获得有效数据 48.31 G。测序错误率低于 0.02%，平均错误率为 0.015%。Q_{20} (质量分数大于等于 20 的碱基所占的比例)碱基百分比在 96%以上，Q_{30} (质量分数大于等于 30 的碱基所占的比例)碱基百分比在 92%以上，GC 含量在 41%左右。本研究测序得到的所有原始转录组数据已提交至 NCBI 的 SRA database，登陆号为 SRP140800。

1.3 转录本拼接

将测序所得原始数据进行混合，采用针对 RNA-Seq 数据专门开发的转录组拼接软件 Trinity (版本号 r20140413p1)对清洁数据进行组装拼接。从拼接转录本中挑选每条基因中最长的转录本作为该基因的代表序列，称之为 Unigene，作为后续分析和注释功能等的参考序列。经过组装，共得到 320 876 条转录本(Transcripts)和 187 362 条独立集团(Unigenes) (表 3)。Transcript 的 N_{50} 和 N_{90} (将拼接结果中归属于同一转录本的长度不一的转录本依据长度从大到小排列，并进行长度累加，当累加值不小于 Unigene 总长 50%/90%的数值就是 N_{50}/N_{90}) 分别为 1 416 和 298，Unigene 的 N 50 和 N 90 分别为 849 和 251。Unigenes 和 Transcripts 序列的长度均分布于 200 bp 至 2 000 bp 以上，其中 200 bp 至 300 bp 区间分布最多(分别为 83 315 和 107 765) (图 2)。Unigenes 和 Transcripts 序列的平均长度分别为 588 bp 和 808 bp，最大长度均为 17 291 bp。

1.4 基因功能注释

将Unigenes 序列分别与 Nr 非冗余蛋白质序列数据库，Nt 核酸序列数据库，Pfam 蛋白质结构域数据库，KOG/COG 同源蛋白簇数据库，Swiss-Prot 蛋白质序列数据库，KEGG 日本京都基因和基因组百科全书和 GO 基因本体论数据库比对分析后，给出了 All Unigene 在每个数据库中的功能注释信息。结果显示(表 4)，有 56%的 Unigene 至少被一个数据库注释，在七大数据库中一共被注释到 104 926 条。NR 数据库共注释到 79 632 条目，占 Unigenes 总数的 42.5%，是所有数据库中注释数量最多的。其次是 Swiss-prot 数据库(66 561 条, 35.52%)、GO 数据库(63 317 条, 33.79%)、Pfam (61 469 条, 32.8%)、Nt 数据库 56 019 条, 29.89%)、KEGG 数据库(35 188 条, 18.78%)、KOG 数据库注释到的数量最少，共注释到 29 512 条(15.75%)。从五大数据库注释结果绘制的 Venn 图中(图 3)，可以发现共有 14 539 条 Unigenes 能被 Nr、Nt、KOG、GO 和 Pfam 5 个数据库注释，12 272 条能同时被所有 7 个数据库注释(表 4)。

表 2 中国沙棘转录组测序质量和数量汇总

Table 2 Statistical datas of RNA-seq for *Hippophae rhamnoides* L.

样品名称 Sample	原始数据 Raw reads	清洁数据 Clean reads	清洁数据碱基数(G) Clean bases (G)	错误率(%) Error (%)	Q20 (%)	Q30 (%)	GC (%)
Male-1	63 943 984	63 028 048	9.45	0.01	97.88	94.57	41.49
Male-2	63 308 850	62 113 974	9.32	0.01	97.68	94.14	41.18
Male-3	59 110 298	57 821 108	8.67	0.01	97.73	94.24	41.52
Female-1	46 863 098	45 439 564	6.82	0.02	96.88	92.42	41.65
Female-2	47 211 330	44 847 076	6.73	0.02	97.08	92.97	41.85
Female-3	51 443 054	48 772 892	7.32	0.02	97.07	92.94	41.53

表 3 拼接转录本长度分布

Table 3 Distribution of number and quality of unigenes and transcripts

	最小长度(bp) Min length (bp)	平均长度(bp) Mean length (bp)	最大长度(bp) Max length (bp)	N_{50}	N_{90}	总条数 Total number	总核苷酸数(nt) Total nucleotides (nt)
独立基因 Unigenes	201	588	17 291	849	251	187 362	110 146 870
转录本 Transcripts	201	808	17 291	1 416	298	320 876	259 142 573

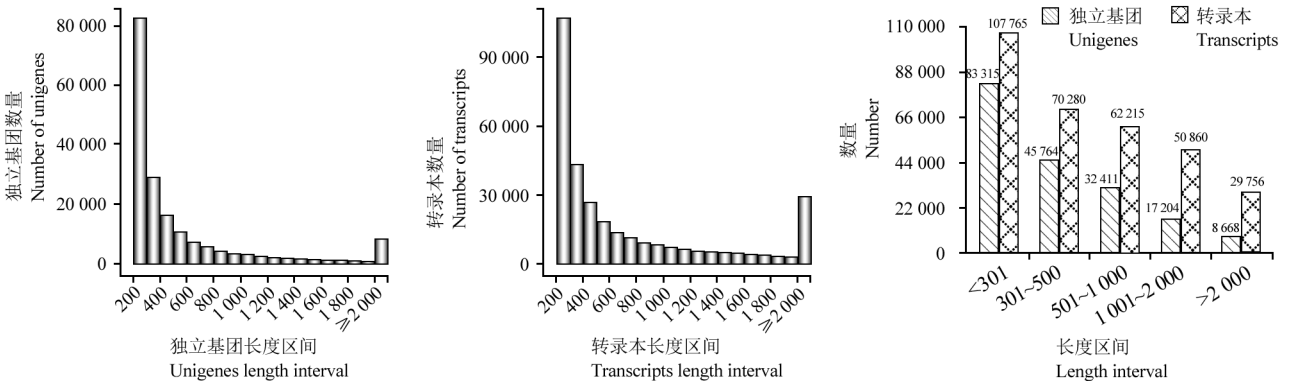

图 2 转录本和独立基因拼接长度和频数分布

Figure 2 Distribution of splice length and frequency of unigene and transcripts

表 4 中国沙棘 unigenes 注释数量统计

Table 4 The statistical number of unigenes that were functional annotated in *Hippophae rhamnoides* L.

数据库 Database	非重复序列基因数 Number of unigenes	比例(%) Percentage (%)
Nr	79 632	42.50
Nt	56 019	29.89
KEGG	35 188	18.78
SwissProt	66 561	35.52
Pfam	61 469	32.80
GO	63 317	33.79
KOG	29 512	15.75
能被所有数据库注释数目 Annotated in all databases	12 272	6.54
至少被一个数据库注释数目 Annotated in at least one database	104 926	56.00
总独立基因数目 Total unigenes	187 362	100.00

1.4.1 注释基因的 Nr 分类

通过与 Nr 数据库进行比对注释，可以获取与中国沙棘转录组序列具有相似性的近缘物种信息。注释结果绘制的物种分布(图 4)。注释结果显示，中国沙棘与葡萄、梅、桃、可可、梨具有较高的序列同源性。其中，与葡萄(*Vitis vinifera*)具有最高的序列相似性，有 10.5%的同源性；其次与梅(*Prunus mume*)有 6.8%的相似性；与桃(*Prunus persica*)有 6.2%的相似性；与可可(*Theobroma cacao*)和梨(*Pyrus x*)分别有 3.9%和 3.5%的相似性。其中 69.1%的 Unigene 属于其它序列，这可能是因为沙棘发生在距今 2 500~4 000 万年间的第 3 纪渐新世，是地球上少有的已经存在超过 3 亿年的植物“活化石”，有许多自身特有的遗传信息等待挖掘。

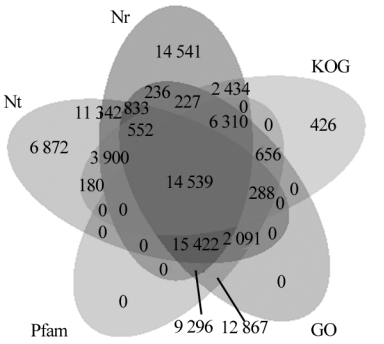

图 3 五大数据库 Unigenes 注释数量统计

Figure 3 The statistical number of unigenes that were functional annotated with the five database

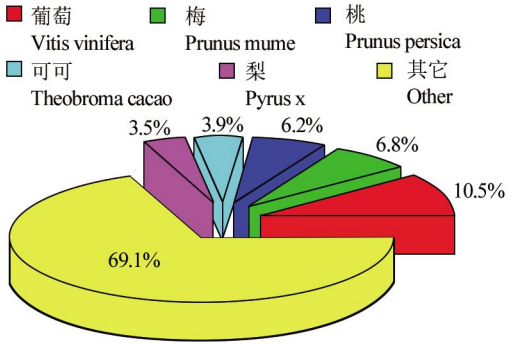

图 4 基于 Nr 数据库比对的同源基因物种来源分布

Figure 4 Species distribution of the top BLASTx hits against the Nr database for unigene in *Hippophae rhamnoides* L.

1.4.2 注释基因的 GO 分类

GO 分类是依据功能及作用对基因或蛋白进行描述和分类的系统，这一分类系统既可以注释某一未知的同源基因的功能，也可用于基因资源的关联分析和深度发掘(Chen and Li, 2005; 肖国华等, 2008; 霍梦琪等, 2016)。对独立基团 Unigenes 进行 GO 注释之后，按照 GO 3 个基本大类(生物学过程, 细胞组分和分子功能)，将注释成功的全部基因按照进一步细分的功能进行聚类(图 5)。基于序列同源性，被分成 3 大主要类别共有 56 个功能群。在 GO 分类的“生物学过程”中，注释基因数量在二级分类排在前 3 位的分别是 Cellular process 细胞过程(34 336 条)、Metabolic process 代谢过程(33 821 条)和 Single-organism process 单一生物学过程(25 874 条)。在“细胞组分”分类中，Cell 细胞(19 534 条)、Cell part 细胞要素(19 520 条)和 Organelle 细胞器(13 542 条)这 3 个类别中，注释的基因数目最多。而在“分子功能”分类中，注释基因最多的类别分别是 Binding 结合(30 496 条)和 Catalytic activity 催化剂活性(27 397 条)。

1.4.3 注释基因的 KOG 分类

KOG 数据库大致上可归纳为 26 个组，将通过 KOG 成功注释的 Unigenes 按照分组进行分类(图 6)。KOG 数据库总共注释到 29 512 条 Unigenes。在这 25个分类中，“翻译后修饰、蛋白翻转、分子伴侣(Posttranslational modification, Protein turnover, Chaperones)”是聚类最多的群体，共 4 711 条，其次是 4 710 条的“翻译、核糖体结构和生物合成(Translation, Ribosomal structure and biogenesis)”。而“细胞迁移(Cell motility)”这个类别是注释最少的，只有 17 条。

1.4.4 注释基因的 KEGG 分类

对 Unigenes 进行 KO 功能注释后，依据其注释结果和参与的 KEGG 代谢通路进行分类(图 7)。总体上可依据 Unigenes 参与的代谢通路分为 5 个类别，即细胞过程、环境信息处理、遗传信息处理、代谢和有机系统。将中国沙棘雌、雄株样品的 Unigene 注释到 KEGG 数据库进行代谢分类分析，共注释到 35 188 条转录组数据，参与了 19 个 KEGG 代谢通路。其中，遗传信息处理 Genetic Information Processing 分支中的翻译 Translation 代谢途径的 Unigene 数量最多，有 5 240 条；其次，代谢 Metabolism 分支中的碳水化合物代谢 Carbohydrate metabolism 途径的 Unigene 数量有 4 166 条；而膜运输途径只有 156 条 Unigene，数量最少。

1.5 Unigene 的 SSR 分析

简单重复序列标记(simple sequence repeats, SSR)，称为短串联重复序列或微卫星 DNA 标记。对中国沙棘转录组的 187 362 条 Unigene 序列进行 SSR 分析，共鉴定出 33 248 个 SSRs，总长度为 110 146 870 bp。这些微卫星 DNA 分布于 25 686 个 Unigene 序列中，其中有 5 671 条含有一个以上简单重复序列标记，其中在化学成分形成中的 SSR 数有 3 089 个。此外，对不同重复类型的 SSR 数量分布统计表明(图 8; 图 9)，单核苷酸 SSR 数目最多，总共有 19 997 个，占 6 种核苷酸重复类型的 60.14%；其次为二核苷酸(7 518 条, 22.61%)、三核苷酸(5 091 条, 15.31%)、四核苷酸(440 条, 1.32%)、五核苷酸(104 条, 0.31%)和六核苷酸重复(98 条, 0.29%)。

1.6 中国沙棘雌雄株叶片差异基因表达分析

对通过读段(Reads)数统计得到的 Readcount 数据，采用 DESeq 进行筛选分析，筛选阈值为 $p_{adj}< 0.05$ (p-adjusted: 校正后的 p value, 同 p value 的统计学意义一样, p-adjusted 越小, 表示基因表达差异越显著)。

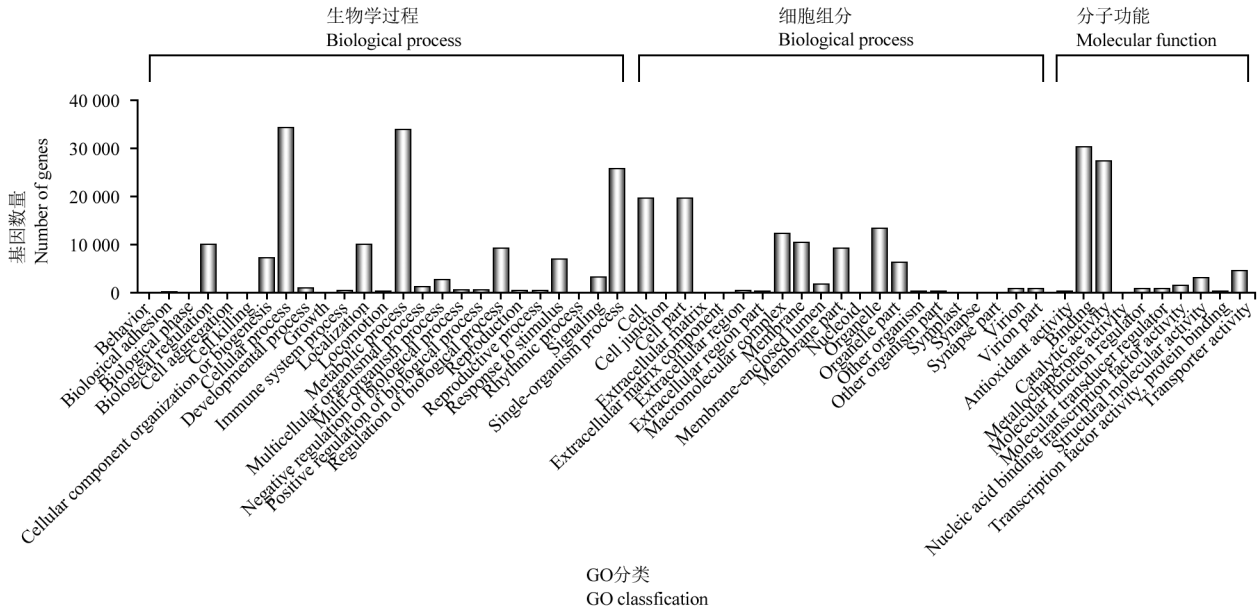

图 5 GO 注释分类

Figure 5 GO function annotion and classification of unigene

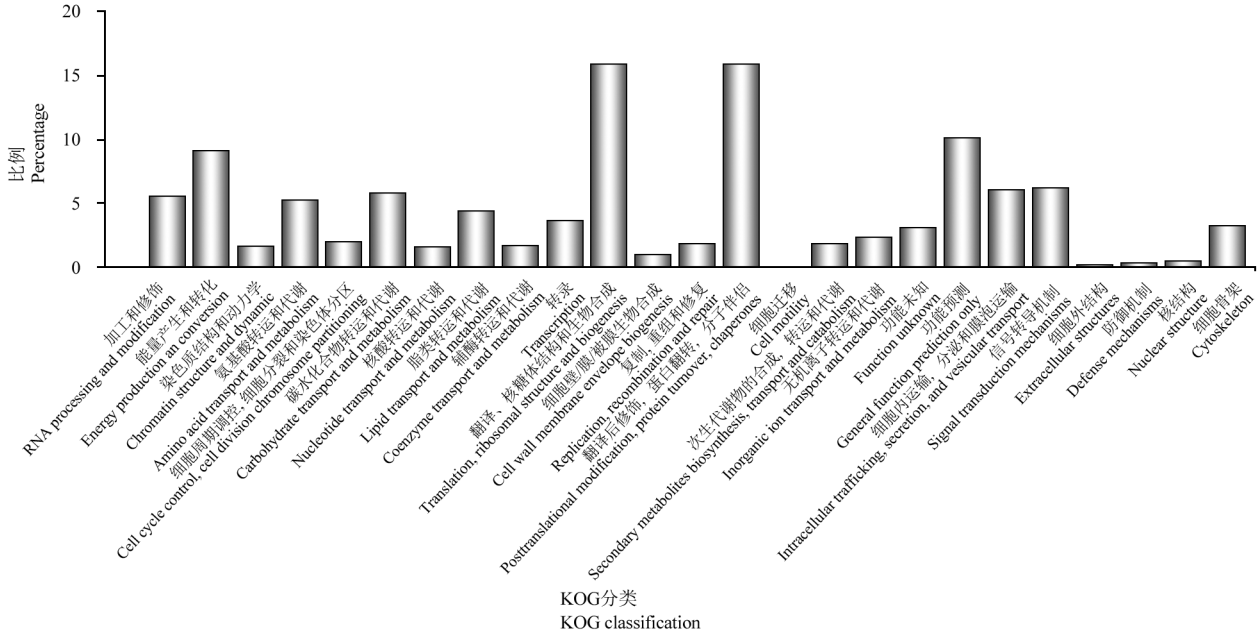

图 6 KOG 注释分类

Figure 6 KOG function classification

依据沙棘植株的性别，将 6 个样品分为雄株(Male)和雌株(Female) 2 个组，各设 3 个重复。分别将中国沙棘雌雄株叶片的测序数据与混合所有测序数据拼接得到的转录本数据进行比对，依据分析结果绘制的基因表达韦恩图(图 10) (fpkm>0.3)可知，雄株共表达了 126 739 条基因，雌株共表达了 87 183 条基因，比雄株少了 39 556 条基因。雄株和雌株共有的表达基因为 60 476 条，而雄株特有的表达基因为 66 263 条，雌株特有的表达基因为 26 707 条，雌雄差异基因总共 92 970 条。差异基因的筛选条件设置为 p_{adj}<0.05 时，所绘制的火山图能直接反映 q value 和 log2 (Fold-changge)的关系，更直观展示了上调表达和下调表达基因数量。根据火山图展示结果可知(图 11)，中国沙棘雄株比雌株显著上调基因有 71 个，下调基因有 109 个。

2 讨论

沙棘作为一种重要的生物资源，中国从上世纪

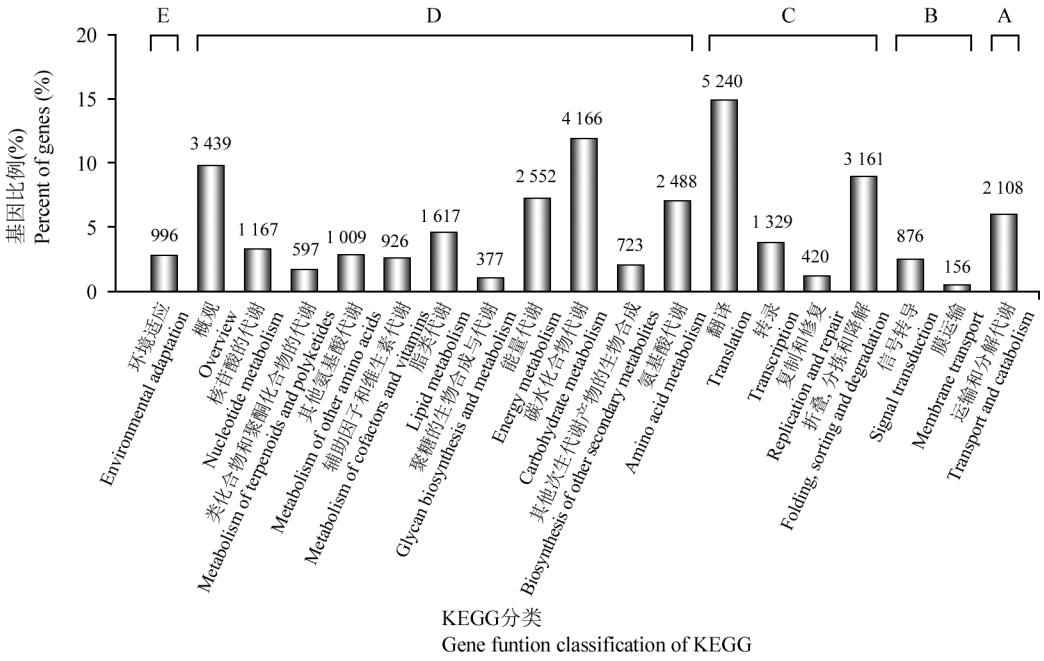

图 7 KEGG 注释分类

Figure 7 Summary of KEGG pathways

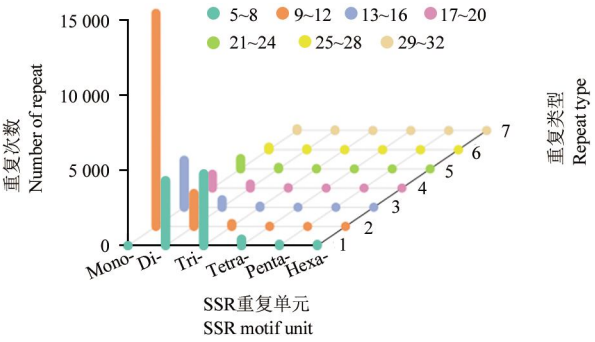

图 8 沙棘转录组 SSR 基序单元的数量分布

注: X 坐标为 SSR 类型, Y 坐标数值是坐标; 具体重复的次数应按照颜色与图例对应, Z 坐标是 SSR 数目

Figure 8 The number distribution of SSR motif unit in *Hippophae rhamnoides* L. transcriptome

Note: The X coordinate represents SSR type, and the Y coordinate represents numbers; The number of concrete repetitions corresponded to the color and legend, and the Z coordinate represents the number of SSR

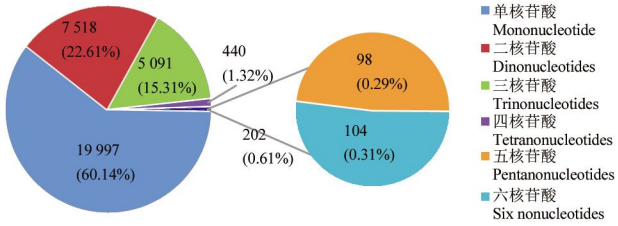

图 9 沙棘重复基元种类分布

Figure 9 Types and distribution of SSRs in *Hippophae rhamnoides* L.

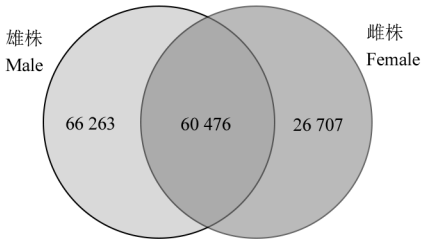

图 10 沙棘雄雌株差异基因表达

Figure 10 Differentially expressed genes of male and female *Hippophae rhamnoides* L.

80 年代就已经开始了对沙棘进行系统而全面的研究和开发。目前，已经拥有了丰富的天然沙棘种质资源，带动了一大批龙头企业参与沙棘产品的开发，形成了规模化的产业链，同时也锻炼和培育了庞大的的沙棘研究和开发队伍。目前国际沙棘研究公认的方向有：沙棘生态林营造、沙棘种子园建设、广泛开展不同生态型及亚种间的远缘杂交育种、人工诱变育种、高生化成分新品种选育、沙棘主要性状遗传规律及分子生物学方面的基础研究等。其中针对沙棘育种中存在的问题，国内学者的研究重点集中于沙棘良种选育和良种快速繁殖体系的建立(硬木扦插和软木扦插)，以满足大规模人工沙棘园和生态恢复的种植需求。而目前各类沙棘研讨会普遍得出一个共识：即便选育到再优良的沙棘品种，不解决沙棘植株幼苗早期雌雄性别鉴定的技术难题，改变人工沙棘林雌雄株比例也难以大幅提高人工沙棘林结实率和

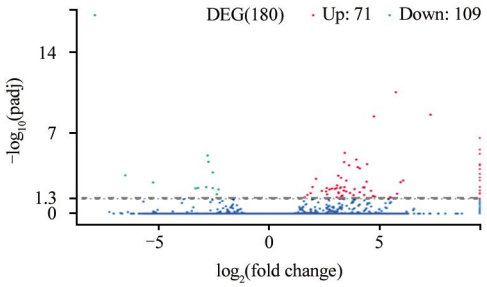

图 11 沙棘雌雄差异基因表达模式的比较分析

注: 横坐标表示基因在中国沙棘雌雄株样品间差异表达倍数的变化; 纵坐标表示基因表达量变化的统计学检验的显著程度, 校正后的 p value 越小, −log10 (校正后的 p value)越大, 即差异越显著; 图中的散点代表各个基因, 蓝色圆点表示无显著性差异的基因, 红色圆点表示有显著性差异的上调基因, 绿色圆点表示有显著性差异的下调基因(p_{adj}<0.05)

Figure 11 Comparison and analysis of expression patterns of differential unigenes identified between male and female in *Hippophae rhamnoides* L.

Note: The x-axis represents fold-change (log10) of differentially expressed genes in two groups; The y-axis represents the adjusted p value (log10), statistical difference in gene expression; Dots in the graph represent genes; Blue dots represent genes that did not change significantly; The red dots indicate significantly up-regulated genes under the condition of p_{adj}<0.05, the yellow dots indicate the significantly up-regulated genes under p_{adj}<0.05

单位面积产果量，实现沙棘浆果采集的标准化和现代化。如果能开创性解决沙棘幼苗早期性别鉴定的难题，将人工沙棘林中雌株的比例从 50%左右提高到 91%左右，即雌雄比例为 10:1 (Jadhav and Sharma, 2014; Das et al., 2016)，必将大力推进和实现沙棘资源的高效利用，提高生产效益，也有助于揭示中国沙棘性别决定的分子机制。

中国沙棘是典型的雌雄异株植物，为沙棘进化和演替的原始类群，是研究植物雌雄性别进化的合适材料。大量研究表明，在雌雄异株植物中，性别决定和性别表达多与性染色体紧密相关，有的受性染色体上性别决定基因直接驱动，有的受常染色体上性别相关基因与性染色上性别决定基因互作的影响，还有的受环境与性染色上性别决定基因相互作用的影响。而目前，尚未有筛选和克隆沙棘性别决定基因的报道。

在雌雄异株植物性别决定与分化的研究中，转录组学作为一种重要研究方法发挥了重要作用。通过对雌雄异株植物转录组进行比较分析，可筛选和鉴定出大量在雌、雄株中存在表达差异的基因，这些基因有的参与生殖过程、调控性别分化和表达，从这些候选基因中甚至有可能筛选和鉴定出性别决定基因。另外，对性别差异表达基因的功能注释和代谢途径聚类分析有助于了解性别分化和表达过程中的分子调控机制和代谢差异，这对于开展雌雄株植物的快速鉴别具有重要意义，如可根据代谢途径中一种性别特异性催化酶，开发一种化学指示剂通过滴在叶片上的显色反应直接鉴别植物性别。目前研究人员已经运用 RNA-seq 技术在研究动植物的性别分化中取得了一系列突破性研究成果(Wolf and Bryk, 2011; Petropoulos et al., 2016)。

本研究运用 RNA-Seq 高通量测序获得了中国沙棘雌雄株叶片的转录组数据，这些数据为设计 SSR 引物和筛选 SSR 分子标记提供了数据支持；通过差异基因表达分析，也筛选得到了一批在中国沙棘雌雄株叶片中存在差异表达的基因。这一结果不仅初步揭示了中国沙棘雌雄株生理生化过程和代谢调控网络的差异，同时也是下一步筛选中国沙棘性别决定和性别表达相关的候选基因，特别是这其中包含仅在中国沙棘雌性或雄性一种性别植株中几乎不表达而在另外一种性别植株中显著性表达的基因 77 个(p_{adj}<0.01)。但这些基因表达上的差异也许只是由于基因表达的时空特异性导致的，同时二代转录组测序也存在测序长度的限制，基因表达水平量化上存在一定误差，因此这些基因必须经过下一步 Real-Time PCR 检验和基因克隆验证。

3 材料与方法

3.1 实验材料

试验材料取至青海省大通县朔北藏族乡边麻沟野生沙棘林(E101°50'40.05", N36°57'45.36", 海拔 3 010 m)。选择雌株已经挂果的 9 月，取树龄 10 年以上，长势良好，无病虫害的雌雄株嫩叶各 3 份，放置于液氮中速冻。带回实验室后，置于 −80℃冰箱中备用。

3.2 RNA 提取与检测

采用 TaKaRa MiniBEST Plant RNA Extraction Kit 试剂盒提取中国沙棘雌雄株叶片总 RNA。采用 1%琼脂糖凝胶电泳分析 RNA 是否存在污染，用 Agilent 2100 生物分析仪(Agilent Technologies, Palo Alto, CA, USA)测试 RNA 的完整度(28S/18S rRNA)，用 Nanodrop 2000 (Thermo Fisher Scientific, Wilmington, DE, USA)检测 RNA 的纯度($OD_{260/280}$ 比值)，用 Qubit 2.0 Fluorometer (Thermo Scientific Inc., Waltham, MA, USA)对 RNA 浓度进行精确定量。

3.3 测序文库的构建、库检及测序

RNA 的质量和浓度检验合格后，通过表面偶联有 Oligo (dT)的磁珠富集信使 RNA，将 mRNA 打断成短的 Fragments。以 Fragments 为扩增模板，使用 6 bp 随机引物，通过 rt-PCR 扩增反应依次合成单链 cDNA 和双链 cDNA。经 AMPure XP beads 吸附纯化的双链 cDNA，在平末端修复后，先在 3' 端加上腺嘌呤 A 尾巴并连上测序 Adaptor，再进行片段大小筛选和富集，最后经过 PCR 扩增和纯化，得到测序文库。

对文库纯度、完整度和有效浓度检验合格后，根据测定的有效浓度将文库按照预测的沙棘基因组大小和转录组分析要求下机数据量 Pooling 后，开始进行 HiSeq-PE150 (Illumina, San Diego, CA, USA)测序工作。

3.4 转录本组装与拼接

由于目前暂无报道的中国沙棘全基因组信息，因此需要进行无参转录组拼接。在转录组 Clean reads 下机后，采用 Trinity 对 Clean reads 进行拼接(Grabherr et al., 2011)。Trinity 版本号为 r20140413p1，其中最小的 kmer 覆盖度 min_kmer_cov 设置为 2，其余参数均为默认设置。

3.5 基因功能注释

为获得全面的基因功能信息，将拼接得到的中国沙棘总的 Unigene 与 7 种数据库 Nr、Nt、Pfam、KOG/COG、Swiss-prot、KEGG、GO 进行比对(Conesa et al., 2005; Mao et al., 2005; Kanehisa et al., 2007; Young et al., 2010; Finn et al., 2016)。使用 NCBI blast 2.2.28+软件与 Nr、Nt、Swiss-prot、KOG/COG 数据库比对时，E-value 值分别设置为 1e−5、1e−5、1e−5、1e−3。使用 KAAS 软件与 KEGG 数据库比对时，E-value 值分别设置为 1e−10。使用 Hmmscan (HMMER 3.0)与 Pfam 数据库比对时，E-value 值分别设置为 0.01。使用 Blast2GO 与 GO 数据库比对时，E-value 值分别设置为 1e−6。

3.6 SSR 标记筛选

采用 MISA (1.0 版, 默认参数)对 Unigene 进行 SSR 检测(王希等, 2016, 中国农学通报, 32(10): 150-156)，并且对不同 SSR 类型在基因转录本的密度分布进行统计。采用 Primer3 (2.3.5 版, 默认参数)进行 SSR 引物设计。

3.7 中国沙棘雌雄株叶片差异基因表达分析

以拼接的 Unigene 作为 RefSeq (参考序列)，将测序得到 Clean reads 与 Unigene 作比对。采用 RSEM 软件(bowtie 2, mismatch=0)对结果进行分析(Li and Dewey, 2011)，得到样品的 Readcount 数目。对 Readcount 进行 FPKM (Fragments Per Kilobase Million)转换，得到基因的表达水平。在 RNA-Seq 技术中，FPKM 适用于双端测序文库，兼顾了测序深度和基因长度对测序结果的影响(Trapnell et al., 2010)。将 FPKM 转换得到的数据进行 DESeq 分析，筛选阈值为 $p_{adj}< 0.05$ (p_{adj} 为校正后的 p-value)，|log2 (foldchange)|>1 (Anders and Huber, 2010; Wang et al., 2010)。

作者贡献

周武是本研究的试验设计和研究执行人，负责数据分析与论文写作；刘晓彤和王煜伟负责试验操作与数据处理；索有瑞和胡娜主要负责试验指导、方案构思、论文修改与审阅。全体作者都阅读并同意最终的文本。

致谢

本研究由青海省科技厅项目(2015−NK−509, 2017−SF-A8, 2017−ZJ-Y11)和中国科学院西部之光青年学者 B 类项目共同资助。

参考文献

Ainsworth C., 2000, Boys and girls come out to play: The molecular biology of dioecious plants, Annals of Botany, 86(2): 211-221

Anders S., and Huber W., 2010, Differential expression analysis for sequence count data, Genome Biol., 11(10): R106

Bao G.L., Du Z.M., Zhao Z.H., Bai X.H., Wang Y.Y., Liu M.J., and Yu L.J., 2009, Experimental studies on antitussive, expectorant and antiasthmatic actions of Wuwei Shaji pulvis, Xiandai Zhongxiyi Jiehe Zazhi (Modern Journal of Integrated Traditional Chinese and Western Medicine), 18(3): 243-244 (包桂兰, 杜智敏, 赵中华, 白旭华, 王玉莹, 刘明洁, 于丽君, 2009, 五味沙棘散止咳祛痰平喘作用的实验研究, 现代中西医结合杂志, 18(3): 243-244)

Chen W.P., and Li Y.M., 2005, Effects of flavonoids from hippophae rhamnoides leaves and fruits on myocardial ischemia and hypoxia, in: Zhongguo Zhongxiyi Jiehe Xuehui Huoxue Huayu Zhuanye Weiyuanhui (eds.), Diliuci Quanguo Zhongxiyi Jiehe Xueyuzheng Ji Huoxue Huayu Yanjiu Xueshu Dahui Lunwen Huibian (Compilation of the Sixth National Conference on Integrated Traditional Chinese and Western Medicine for Blood Stasis and Promoting Blood Circulation

and Removing Blood Stasis), Zhongguo Zhongxiyi Jiehe Xuehui Huoxue Huayu Zhuanye Weiyuanhui, Jilin, China, pp.156-157 (陈卫平, 李毅敏, 2005, 沙棘叶和沙棘果黄酮的抗心肌缺血, 缺氧作用研究, 见: 中国中西医结合学会活血化淤专业委员会, 第六次全国中西医结合血瘀症及活血化瘀研究学术大会论文汇编, 中国中西医结合学会活血化淤专业委员会, 中国, 吉林, pp.156-157)

Chen Z., Xue C., Sheng Z., Zhou F., Ling X., Liu G., and Chen L., 2005, GoPipe: Streamlined gene ontology annotation for batch anonymous sequences with statistics, Progress in Biochemistry and Biophysics, 32(2): 187-191

Conesa A., Götz S., Garcíagómez J.M., Terol J., Talón M., and Robles M., 2005, Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research, Bioinforma., 21(18): 3674-3676

Das K., Ganie S.H., Mangla Y., Chaudhary M., Thakur R.K., Tandon R., Chaudhary M., Thakur R.K., Tandon R., Raina S.N., and Goel S., 2016, ISSR markers for gender identification and genetic diagnosis of *Hippophae rhamnoides* ssp. *turkestanica* growing at high altitudes in Ladakh region (Jammu and Kashmir), Protoplasma, 254(2): 1-15

Finn R.D., Coggill P., Eberhardt R.Y., Eddy S.R., Mistry J., Mitchell A.L., Potter S.C., Punta M., Qureshi M., Sangrador-Vegas A., Salazar G.A., Tate J., and Bateman A., 2016, The Pfam protein families database: Towards a more sustainable future, Nucleic Acids Res., 44(D): 279-285

Goel H.C., Gupta D., Gupta S., Garg A.P., and Bala M., 2005, Protection of mitochondrial system by *Hippophae rhamnoides* L. against radiation-induced oxidative damage in mice, J. Pharm. Pharmacol., 57(1): 135-143

Grabherr M.G., Haas B.J., Yassour M., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., Palma F.D., Birren B.W., Nusbaum C., Lindblad-Toh K., Friedman N., and Regev A., 2011, Full-length transcriptome assembly from RNA-Seq data without a reference genome, Nat. Biotechnol., 29(7): 644

Huo M.Q., Zhang Y.L., Zheng S.C., and Qiao Y.J., 2016, Mechanism of tetramethylpyrazine in treatment of coronary heart disease based on the coexpression-protein interaction network, Beijing Zhongyiyao Daxue Xuebao (Journal of Beijing University of Traditional Chinese Medicine), 39(12): 989-997 (霍梦琪, 张燕玲, 郑世超, 乔延江, 2016, 基于共表达蛋白相互作用网络探讨川芎嗪治疗冠心病的机制, 北京中医药大学学报, 39(12): 989-997)

Jadhav M.S., and Sharma T.R., 2014, Identification of gender specific DNA markers in sea buckthorn (*Hippophae rhamnoides* L.), Ind. Res. J. Genet Biotech, 6(3): 464-469

Jiang F.X., Wei P.W., Wu S., Jiang X.Y., Shi J.T., and Chen Q.B., 2017, Transcriptome analysis insights into bulblet on leaf surface in *Ornithogalum thyrsoides*, Fenzi Zhiwu Yuzhong (Molecular Plant Breeding), 15(2): 519-531 (姜福星, 魏丕伟, 吴生, 江欣燕, 施敬恬, 陈其兵, 2017, 白花虎眼万年青叶上珠芽的转录组分析, 分子植物育种, 15(2): 519-531)

Kanehisa M., Araki M., Goto S., Hattori M., Hirakawa M., Itoh M., Katayama T., Kawashima S., Okuda S., Tokimatsu T., and Yamanishi Y., 2007, KEGG for linking genomes to life and the environment, Nucleic Acids Research, 36(D): 480-484

Li B., and Dewey C.N., 2011, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC Bioinformatics, 12(1): 323

Lian Y.S., and Chen X.L., 1996, Systematic classification of *Hippophae*, Shaji (Hippophae), 9(1): 15-24 (廉永善, 陈学林, 1996, 沙棘属植物的系统分类, 沙棘, 9(1): 15-24)

Limao C.R., 2012, Study on preparation, characterization and function of important active ingredient from *Hippophae rhamnoides* L. in Qinghai, Chinese Academy of Scienes, Supervisor: Suo Y.R., and Chen Z., pp.1-18 (利毛才让, 2012, 青海沙棘重要活性成分的制备, 表征及功能作用研究, 博士学位论文, 中国科学院研究生院, 导师: 索有瑞, 陈志, pp.1-18)

Liu J., Luo C., Sun W., and Yi Y., 2017, Transcriptomics analysis of hard seeds of *Sophora viciifolia*, Fenzi Zhiwu Yuzhong (Molecular Plant Breeding), 15(3): 867-874 (刘杰, 罗充, 孙威, 乙引, 2017, 白刺花种子转录组分析, 分子植物育种, 15(3): 867-874)

Lv C.Z., Shan Y.K., Liu H.Z., Yang L.H., and Ma G., 2008, Natural vitamin king-Seabuckthorn in the application of food ingredients, Zhongguo Shipin Tianjiaji (China Food Additives), (s1): 229-235 (鲁长征, 山永凯, 刘洪智, 杨犁华, 马光, 2008, 天然维生素之王－沙棘在食品配料中的应用, 中国食品添加剂, (s1): 229-235)

Mao X., Cai T., Olyarchuk J.G., and Wei L., 2005, Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary, Bioinformatics, 21(19): 3787-3793

Narayanan S., Ruma D., Gitika B., Sharma S.K., Pauline T., Ram M.S., Ilavazhagan G., Sawhney R.C., Kumar D., and Banerjee P.K., 2005, Antioxidant activities of seabuckthorn (*Hippophae rhamnoides*) during hypoxia induced oxidative stress in glial cells, Mol. Cell Biochem., 278(1-2): 9-14

Ni W., Gao T., Wang H., Wang H., Du Y., Li J., Li C., Wei L., and Bi H., 2013, Anti-fatigue activity of polysaccharides from the fruits of four Tibetan plateau indigenous medicinal plants, J. Ethnopharmacol., 150(2): 529-535

Pang L., 2009, The mechanism of protective effect of seabuckthorn seed oil fat emulsion in mice against radiotherapy and chemotherapy injury, China Medical University, Supervisor:

Jin W.B., pp.13-40 (逄蕾, 2009, 沙棘籽油脂肪乳对小鼠放疗和化疗损伤的保护作用与机制探讨, 硕士学位论文, 中国医科大学, 导师: 金万宝, pp.13-40)

Petropoulos S., Edsgã¤Rd D., Reinius B., Deng Q., Panula S.P., Codeluppi S., Reyes A.P., Linnarsson S., Sandberg R., and Lanner F., 2016, Single-cell RNA-Seq reveals lineage and X chromosome dynamics in human preimplantation embryos, Cell, 167(1): 1012-1026

Qi Y.X., Liu Y.B., and Rong W.H., 2011, RNA-Seq and its applications: a new technology for transcriptomics, Yichuan (Hereditas), 33(11): 1191-1202 (祁云霞, 刘永斌, 荣威恒, 2011, 转录组研究新技术: RNA-Seq 及其应用, 遗传, 33(11): 1191-1202)

Shukla S.K., Chaudhary P., Kumar I.P., Samanta N., Afrin F., Gupta M.L., Sharma U.K., Sinha A.K., Sharma Y.K., and Sharma R.K., 2006, Protection from radiation-induced mitochondrial and genomic DNA damage by an extract of *Hippophae rhamnoides*, Environ. Mol. Mutagen., 47(9): 647-656

Trapnell C., Williams B.A., Pertea G., Mortazavi A., Kwan G., Van B.M.J., Salzberg S.L., Wold B.J., and Pachter L., 2010, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, Nat. Biotechnol., 28(5): 511-515

Wang L., Feng Z., Wang X., Wang X., and Zhang X., 2010, DEGseq: An R package for identifying differentially expressed genes from RNA-seq data, Bioinformatics, 26(1): 136-138

Wolf J.B., and Bryk J., 2011, General lack of global dosage compensation in ZZ/ZW systems? Broadening the perspective with RNA-seq, BMC Genomics, 12(1): 91

Xiao G.H., Li Y.J., Guo Z., Peng C.F., Wang D., Zhu J., Yang D., Yao C., and Wang J., 2008, Using gene ontology-based clustering method to study the genetic heterogeneity of leukemia, Shengwu Xinxixue (China Journal of Bioinformatics), 6(1): 9-11 (肖国华, 李永进, 郭政, 彭春方, 王栋, 朱晶, 杨达, 姚晨, 王靖, 2008, 采用基于 Gene Ontology 的聚类方法研究白血病的遗传异质性, 生物信息学, 6(1): 9-11)

Young M.D., Wakefield M.J., Smyth G.K., and Oshlack A., 2010, Gene ontology analysis for RNA-seq: Accounting for selection bias, Genome Biol., 11(2): R14