湖南农业大学学报(自然科学版)　2022，48(1)：27–32. DOI：10.13331/j.cnki.jhau.2022.01.005
Journal of Hunan Agricultural University(Natural Sciences)

引用格式：
刘小红. 基于 SMRT 技术的水杉全长转录组分析及基因功能注释[J]. 湖南农业大学学报(自然科学版)，2022，48(1)：27–32.
LIU X H．Characterization of the full-length transcriptome using SMRT technology and functional annotation of the genes in *Metasequoia glyptostroboides*[J]. Journal of Hunan Agricultural University(Natural Sciences)，2022，48(1)：27–32.
投稿网址：http://xb.hunau.edu.cn

基于 SMRT 技术的水杉全长转录组分析及基因功能注释

刘小红 [1,2]

(1.西华师范大学生命科学学院，四川 南充 637002；2.西南野生动植物资源保护教育部重点实验室，四川 南充 637002)

摘　要：以水杉原生种为材料，提取其幼嫩植株的根、茎和叶的总 RNA，经 mRNA 纯化、反转录、全长转录组文库构建等过程，再采用 SMRT 技术测定其全长转录本序列，运用生物信息学对获得的原始转录组数据进行分析。结果显示：提取的总 RNA 符合建库要求；全长转录组测序共获得了包含 5 914 711 个亚读段(subreads)的 14.0 Gb 的数据，质量控制处理后包括 236 130 个全长非嵌合读段(reads)和 97 626 个一致性 reads；转录本经去冗余处理后共获得 61 057 个全长一致性读段(unigenes)，其中有 54 099 个被成功注释到 7 个数据库中，注释比达 88.60%；对水杉 unigenes 作 CDS(coding sequence) 长度分布及转录因子分析，其 CDS 长度范围为 144~6477 nt，平均长度约为 679 nt；共检测到 2386 个转录因子，这些转录因子可以归类为 29 个家族。

关　键　词：水杉；单分子测序技术；全长转录组；基因；功能注释

中图分类号：Q943.2　　**文献标志码**：A　　**文章编号**：1007−1032(2022)01−0027−06

Characterization of the full-length transcriptome using SMRT technology and functional annotation of the genes in *Metasequoia glyptostroboides*

LIU Xiaohong[1,2]

(1.College of Life Science, China West Normal University, Nanchong, Sichuan 637002, China; 2.Key Laboratory of Southwest China Wildlife Resources Conservation(Ministry of Education), Nanchong, Sichuan 637002, China)

Abstract: This study selected the young roots, stems and leaves of the native species *Metasequoia glyptostroboides*(*M. glyptostroboides*) as material and the total RNA was extracted from them, respectively. The RNA products were equally mixed and then was used for mRNA purification, reverse transcription and full-length transcriptome library construction. The SMRT(single-molecule real-time) technology was employed to sequence transcriptome library, and bioinformatics methods were used to analyze transcriptome data. The results showed that the extracted total RNA could meet the requirements of library construction; a total of 14.0 Gb of data containing 5 914 711 subreads were obtained, including 236 130 full-length non-chimeric reads and 97 626 consensus reads after quality control treatment. After redundancy removal, a total of 61 057 full-length unigenes were identified, among which 54 099 were successfully annotated into seven databases, accounting for 88.60%. In addition, CDS(coding sequence) length distribution and TFs(transcription factors) of these unigenes were analyzed. The CDS length ranged from 144 to 6 477 nt, with an average length of 679 nt; a total of 2386 transcription factors were detected, which could be classified into 29 families.

Keywords: *Metasequoia glyptostroboides*; SMRT(single-molecule real-time) sequencing technology; full-length transcrptome;

收稿日期：2020−11−10　　**修回日期**：2021−12−22
基金项目：西华师范大学科研项目(19B033)
作者简介：刘小红(1975—)，男，重庆彭水人，博士，教授，主要从事植物分子遗传学研究，350783409@qq.com

gene; functional annotation

水杉(*Metasequoia glyptostroboides*) 属孑遗植物，为杉科(Taxodiaceae)、水杉属(*Metasequoia*) 的单一种, 在植物界有“活化石”之称[1]。据史料记载，水杉起源于中生代的白垩纪，在新生代的第三纪时遍布北半球，但在第四纪的冰川期后全世界仅有极少量个体存活，直到 20 世纪 40 年代才在中国湖北省利川市的小河境内被首次发现[2]。水杉生长历史悠久，对于研究古生物、古气候、古地质以及裸子植物的系统演化等都具有重要意义[3]。目前，水杉已被列入国家一级保护植物名录。

水杉虽然具有重要的生物学作用，但关于它的研究主要集中在栽培引种[4–5]、生理生化[6–7]、系统发育[8]等方面，而在分子遗传水平上的研究较少[9]。转录组测序可以获得大量的基因信息[10–13]，能揭示潜在的代谢途径和遗传机制，可为其进一步的分子生物学研究提供依据。

关于转录组测序，以 Illumina/Solexa、Roche/454 和 ABI/SOLiD 为代表的二代测序平台具有测序时间短、成本低、准确性高、高通量等优点，目前在差异基因表达方面被广泛应用[14–15]。但二代测序读长较短，在没有参考基因组的情况下难以获得基因的全长序列信息。而近几年发展起来的三代测序技术已经成为更好的选择，如 PacBio 测序平台的单分子测序 SMRT(single–molecule real–time)技术已经成为获得全长转录序列的首要选择。三代测序技术的主要优点是读长较长(平均读长可达 20 kb), 对于逆转录生成的全长转录本不需要对其进行片段化处理，可以直接作单分子测序获得全长序列信息[16–17]。全长转录组测序技术已被用于许多植物的全长转录组分析[18–20]，但在水杉植物中还少见相关报道。基于此，本研究拟以来自湖北省利川市的原生水杉种为材料，采用 SMRT 三代测序技术对全长转录本进行测序分析，以获得相应的遗传信息，旨在为后续基因克隆及基因功能鉴定提供依据。

1 材料与方法

1.1 材料

选取来自湖北省利川市林业科学研究所(108°96′E，30°28′N) 的原生水杉种为材料，在植株长至 10 cm 左右时，分别取其根、茎、叶，液氮速冻保存。

1.2 方法

1.2.1 总 RNA 的提取及其质量检测

水杉根、茎、叶的总 RNA 的提取都用 RNeasy Plus Mini Kit(Qiagen, Valencia, CA, USA) 完成。对提取得到的总 RNA 先用 Nanodrop 2000 和 Agilent 2100 分析其浓度，测定 $OD_{260/280}$、$OD_{260/230}$、25S/18S 和 RIN 值，再用 1.0%的琼脂糖凝胶电泳分析是否有 DNA 污染和 RNA 降解现象。

1.2.2 SMRT 测序

将检测合格的水杉根、茎和叶的总 RNA 等量混匀，经 mRNA 纯化、反转录及 SMRT 测序后得到全长转录组原始数据(raw data)。该工作由北京诺禾致源科技股份有限公司的 PacBio 测序平台完成。

1.2.3 raw data 及 subreads 统计

原始数据经质量控制处理后进行 subreads 的统计分析，包括有效插入片段 subreads 数据量大小、有效插入片段 subreads 条数、平均 subreads 长度及 N50(将得到的 subreads 按照长度从大到小排序，依次累加 subreads 的长度，直至其长度不小于总 subreads 长度的 50%)的长度。绘制 subreads 长度分布图。

1.2.4 全长转录本分析

选用 SMRT Link 5.1 对全长转录本进行统计，按默认参数设置，统计以下项目：环形一致性序列(CCS) 数量；全长非嵌合 reads 数量；全长非嵌合 reads 平均长度；聚类分析之后得到的一致性序列的 reads 数量。

1.2.5 转录本去冗余

利用 CD–HIT[21]通过序列比对聚类法去除冗余，输出 1 个非冗余的序列文件，去冗余后的基因即为 unigenes。对转录本去冗余前后长度频数分布情况和含有相同转录本拷贝数的基因数量进行统计。

1.2.6　基因的功能注释

为了得到全面的基因功能信息，对去冗余之后的基因进行数据库(包括 NR、NT、Pfam、KOG、Swiss–Prot、KEGG 以及 GO)功能注释，分析这些基因在数据库中的注释情况，并利用 NR 数据库绘制物种分布图和基因功能的 GO 分类统计图。

1.2.7　基因的结构分析

利用 Angel[22]进行 CDS(coding sequence) 预测分析。该软件具有无错和容错 2 种预测模式。本研究中，选用默认的容错模式，对 CDS 序列的长度及其分布进行预测，绘制序列长度分布图。运用 iTAK[23]对去冗余的 unigenes 进行植物转录因子预测，并将注释到转录本数量最多的转录因子家族进行柱形图展示。

2　结果与分析

2.1　总 RNA 的分离纯化

水杉总 RNA 分离纯化后，用 Nanodrop 2000 和 Agilent 2100 检测的结果显示：在最后加入溶剂都为 30.00 μL 的条件下，叶的总 RNA 浓度最高，达 301.70 ng/μL，其后依次为茎的(268.80 ng/μL) 和根的(189.70 ng/μL)，三者的总量分别为 9.05、8.06、5.69 μg。根、茎和叶的总 RNA 的 $OD_{260/280}$ 值分别为 2.04、2.05、2.11，$OD_{260/230}$ 值分别为 1.25、1.10、1.47，25S/18S 值分别为 1.50、1.50 和 1.40，RIN 值分别为 8.50、8.70、7.70。琼脂糖凝胶电泳结果显示，每个样品都有 2 条明亮的条带，没有拖尾现象，表明提取的 RNA 质量较好，没有 DNA 污染和 RNA 降解现象。上述检测结果表明，提取的 RNA 样品能满足建库测序对质量的要求。

2.2　raw data 及 subreads 结果分析

将提取的水杉根、茎、叶 3 个部位的总 RNA 样品等量混匀，经 mRNA 纯化、反转录及 SMRT 测序，对原始数据进行质量控制处理后共得 5 914 711 个 subreads 和 14.0 Gb subread base，平均 subreads 长度为 2368 nt，N50 为 2569 nt。subreads 的长度分布如图 1 所示。在长度为 100~10 000 nt 的范围内，2300 nt 附近的 subreads 最多，而超过 7000 nt 的则非常少。

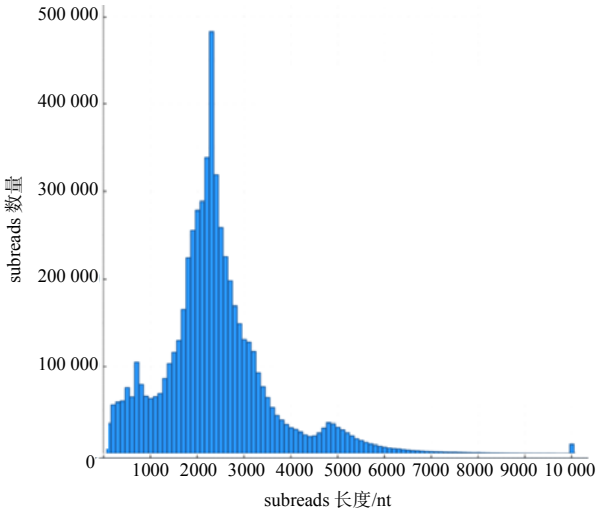


图 1　水杉全长转录组 subreads 的长度分布

Fig.1　The length distribution of the subreads from the full-length transfriptome of *Metasequoia glyptostroboides*

2.3　全长转录本数据统计分析

采用 SMRT 技术测序后共获得 339 296 个 CCS 序列，其中用于后续试验分析的全长非嵌合 reads 共计 236 130 个，平均长度为 2598 nt，对其作一致性序列处理，最后得到一致性 reads 97 626 个。

2.4　转录本的去冗余结果分析

对 97 626 个一致性 reads 进一步作去冗余处理，去冗余前后的结果如表 1 所示。小于 500 bp 的转录本和基因数量较少，分别仅有 739、363 个；在 1~3 kb 的转录本和基因数量较多，约占总量的 96.92%。去冗余后最终得到的基因数(unigenes 的数量) 为 61 057 个，占总转录本的 62.54%。

对去冗余的 61 057 个基因作转录本拷贝数分析，结果拷贝数最少的为 1 个拷贝，最多的为 10 个拷贝，其中 1 个拷贝的基因数达 51 523 个，占总量的 84.39%；含 9 个拷贝的基因数最少，为 134 个，仅占总量的 0.22%。

表 1　水杉转录本去冗余前后的长度分布

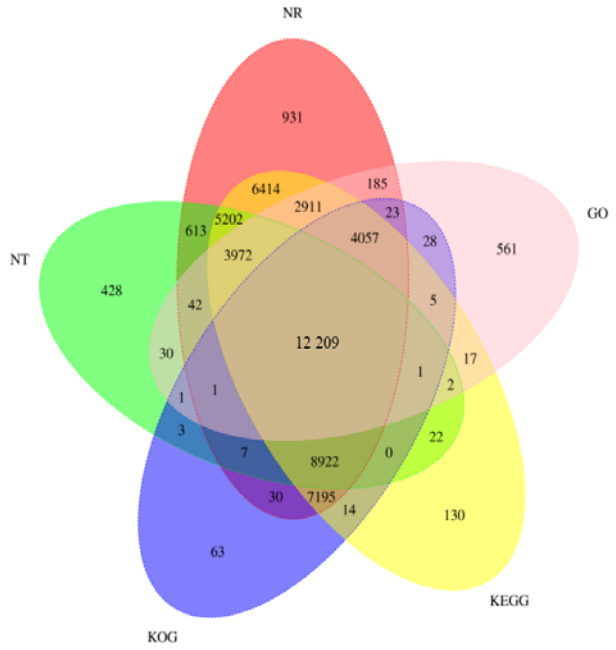
Table 1　Length distribution of the transcripts before and after redundancy removal in *Metasequoia glyptostroboides*

转录本长度范围/bp	转录本数量	基因数
<500	739	363
500 ~ 1 000	2264	1518
>1000 ~ 2 000	23 405	15 222
>2000 ~ 3000	39 089	21 543
>3000	32 129	22 411
总量	97 626	61 057

2.5 基因功能注释结果分析

2.5.1 注释基因数量

利用 NR、Swiss-Prot、KEGG、KOG、GO、NT 和 Pfam 7 个数据库对水杉全长 unigenes 进行功能注释，共有 54 099 个基因被成功注释，被注释的基因在 7 个数据库中的数量依次为 52 714、42 752、51 073、32 559、24 045、31 455、24 045 个，在 7 个数据库中均被注释的有 12 043 个。选取常用的 NR、KEGG、KOG、GO 和 NT 5 个数据库的注释情况绘制韦恩图(图 2)。从图 2 可知，在 NR 数据库中注释的基因数最多，其次为 KEGG 数据库，最少的为 GO 数据库，在 5 个数据库中均有注释的基因数为 12 209 个。



每个大圆圈中的数字之和代表该数据库注释的转录本数；圆圈交叠部分的数字表示不同数据库共同注释到的转录本数。

图 2 水杉基因功能注释韦恩图

Fig.2 The Venn diagram of gene function annotation in *Metasequoia glyptostroboides*

2.5.2 NR 数据库注释

通过与 NR 数据库进行比对注释，共有 52 714 个基因被注释到 568 个物种中。匹配基因数最多的是北美云杉(*Picea sitchensis*)，达 12 519 个；其次为无油樟(*Amborella trichopoda*)，5888 个；排名第三的为莲(*Nelumbo nucifera*)，3417 个；最少的为小油桐(*Jatropha curcas*)，仅有 442 个。

2.5.3 GO 分类

对上述 unigenes 进行 GO 注释，将注释成功的基因按照 GO 3 个大类的下一级进行分类，结果显示：与细胞组分(cellular component)有关的基因有 21 966 个，归属于 18 个下一级分类，平均每级分类为 1220 个。与生物过程(biological process)有关的基因数为 42 898 个，归属于 25 个下一级分类，平均每级为 1716 个。与分子功能(molecular function)相关的基因有 28 015 个，归属于 11 个下一级分类，平均每级为 2547 个。

2.6 基因结构分析

2.6.1 CDS 预测

CDS 预测结果如图 3 所示。共有 61 259 个 CDS 序列，其长度不一，最长的达 6 477 nt，最短的仅为 144 nt，平均长度约为 679 nt。总的趋势是转录本序列越长，则转录本数量越少。

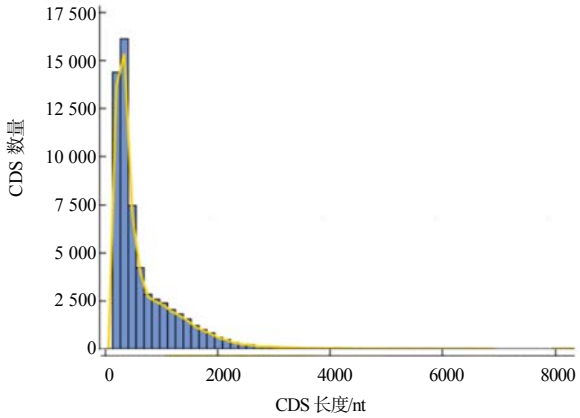


图 3 水杉基因的 CDS 长度分布情况

Fig. 3 CDS length distribution of the genes in *Metasequoia glyptostroboides*

2.6.2 转录因子分析

在 61 057 个 unigenes 中，预测到共有 2386 个表达产物为转录因子，这些转录因子可以分成 29 个家族，其含有的转录因子成员数最多的为 C3H 家族，有 234 个；其次为 GRAS 家族，有 194 个；最少的为 HSF 家族，仅有 28 个。

3 结论与讨论

全长转录组测序技术可有效获取物种转录组信息且可信度高，目前该技术被广泛应用于许多物

种[24–25]，有助于了解没有参考基因组的非模式生物的遗传信息[26]。本研究中，基于 PacBio 平台的 SMRT 测序技术被用于分析水杉幼苗全株的全长转录组测序，共获得了 5 914 711 个 subreads 和 14.0 Gb 的数据量。经过数据优化和去冗余处理，最后获得 61 057 个基因的全长 cDNA 序列信息。本研究结果与在野草莓(*Fragaria vesca*)[27]、紫苜蓿(*Medicago sativa*)[28]和高山杜鹃(*Rhododendron lapponicum*)[29]等中的研究结果类似。本研究中，转录本的平均长度达 2598 nt，远高于二代测序的平均读长[30–31]。本研究结果还表明，不同的基因在转录组测序分析中转录本拷贝数可能不一样，其中具有单拷贝转录本的最多，占全部转录本的 84.39%，总的趋势是拷贝数越多，对应的基因数量越少。

水杉虽然不是模式植物，但可以基于已有参考植物的数据库信息对本研究获得的 61 057 个全长基因进行功能注释，结果共有 54 099 个基因被注释到 7 个数据库中，被注释的基因占 88.60%，其中有 12 043 个同时被 7 个数据库注释，占总量的 19.72%。通过 NR 数据库的注释，发现水杉和北美云杉之间的亲缘关系较近。此外，通过 GO 数据库对基因功能进行分类统计，并将这些基因分成了生物过程、细胞组分及分子功能 3 大类。在基因结构分析中，CDS 预测结果显示，转录本序列越长，则其在细胞中的拷贝数越少，这与前人[20, 24]在其他物种作全长转录组分析的结果一致。转录因子分析发现，获得的部分基因的表达产物其实是与细胞中基因转录调控密切相关的转录因子，与细胞中很多基因能否转录及转录强度密切相关。该结果为分析水杉基因的功能提供了分子基础。一小部分基因不能被注释，可能是因为这些 unigenes 是一些新基因，在数据库中还没找到与此相似的转录本序列，有待于对这些全长转录本对应的基因的结构和功能作进一步研究。

参考文献：

[1] MA J S. The chronology of the “Living Fossil” *Metasequoia glyptostroboides*(*Taxodiaceae*): a review (1943–2003)[J]. Harvard Papers in Botany, 2003, 8(1): 9–18.

[2] 林勇，艾训儒，姚兰，等. 水杉原生母树种群结构与动态[J]. 生态学杂志，2017，36(6)：1531–1538.

[3] JUVIK O J, NGUYEN X H T, ANDERSEN H L, et al. Growing with dinosaurs: natural products from the cretaceous relict *Metasequoia glyptostroboides* Hu & Cheng-a molecular reservoir from the ancient world with potential in modern medicine[J]. Phytochemistry Reviews, 2016, 15: 161–195.

[4] LIU M, FENG Z K, MA C H, et al. Influencing factors and growth state classification of a natural *Metasequoia* population[J]. Journal of Forestry Research, 2019, 30: 337–345.

[5] LIU H, ZHU Y F, LIU X, et al. Effect of artificially accelerated aging on the vigor of *Metasequoia glyptostroboides* seeds[J]. Journal of Forestry Research, 2020, 31: 769–779.

[6] BAJPAI V K, KANG S C. Antifungal activity of leaf essential oil and extracts of *Metasequoia glyptostroboides* Miki ex Hu[J]. Journal of the American Oil Chemists’ Society, 2010, 87: 327–336.

[7] 张垚，白金峰，徐凯莉，等. 具有抗菌活性的水杉内生真菌的分离及筛选[J]. 湖北农业科学，2018，57(11)：36–38.

[8] FAN K X, AI X R, YAO L, et al. Do climate and human disturbance determine the sizes of endangered *Metasequoia glyptostroboides* trees in their native range? [J]. Global Ecology and Conservation, 2020, 21: e00850.

[9] WANG J J, HAN S, YIN W L, et al. Comparison of reliable reference genes following different hormone treatments by various algorithms for qRT-PCR analysis of *Metasequoia*[J]. International Journal of Molecular Sciences, 2019, 20(1): 34.

[10] BRIESE M, SAAL L, APPENZELLER S, et al. Whole transcriptome profiling reveals the RNA content of motor axons[J]. Nucleic Acids Research, 2016, 44(4): e33.

[11] KERR S C, GAITI F, BEVERIDGE C A, et al. *De novo* transcriptome assembly reveals high transcriptional complexity in *Pisum sativum* axillary buds and shows rapid changes in expression of diurnally regulated genes[J]. BMC Genomics, 2017, 18: 221.

[12] LIU H B, SMITH T P L, NONNEMAN D J, et al. A high-quality annotated transcriptome of swine peripheral blood[J]. BMC Genomics, 2017, 18: 479.

[13] SZABO E X, REICHERT P, LEHNIGER M K, et al. Metabolic labeling of RNAs uncovers hidden features and dynamics of the *Arabidopsis* transcriptome[J]. The Plant Cell, 2020, 32(4): 871–887.

[14] EGEKWU N, SONENSHINE D E, BISSINGER B W, et al. Transcriptome of the female synganglion of the black-legged tick *Ixodes scapularis*(Acari: Ixodidae) with comparison between Illumina and 454 systems[J]. PLoS

One，2014，9(7)：e102667.

[15] ALVES R N，GOMES A S，STUEBER K，et al. The transcriptome of metamorphosing flatfish[J]. BMC Genomics，2016，17：413.

[16] YAN B，BOITANO M，CLARK T A，et al. SMRT-Cappable-seq reveals complex operon variants in bacteria[J]. Nature Communications，2018，9：3676.

[17] TENG K，TENG W J，WEN H F，et al. PacBio single-molecule long-read sequencing shed new light on the complexity of the *Carex breviculmis* transcriptome[J]. BMC Genomics，2019，20：789.

[18] DONG L L，LIU H F，ZHANG J C，et al. Single-molecule real-time transcript sequencing facilitates common wheat genome annotation and grain transcriptome research[J]. BMC Genomics，2015，16：e1039.

[19] CHEN X Z，LI J R，WANG X B，et al. Full-length transcriptome sequencing and methyl jasmonate-induced expression profile analysis of genes related to patchoulol biosynthesis and regulation in *Pogostemon cablin*[J]. BMC Plant Biology，2019，19(1)：266.

[20] ROACH N P，SADOWSKI N，ALESSI A F，et al. The full-length transcriptome of *C. elegans* using direct RNA sequencing[J]. Genome Research，2020，30：299–312.

[21] FU L M，NIU B F，ZHU Z W，et al. CD-HIT：accelerated for clustering the next-generation sequencing data[J]. Bioinformatics，2012，28(23)：3150–3152.

[22] SHIMIZU K，ADACHI J，MURAOKA Y. Angle：a sequencing errors resistant program for predicting protein coding regions in unfinished cDNA[J]. Journal of Bioinformatics and Computational Biology，2006，4(3)：649–664.

[23] ZHENG Y，JIAO C，SUN H H，et al. iTAK：a program for genome-wide prediction and classification of plant transcription factors，transcriptional regulators，and protein kinases[J]. Molecular Plant，2016，9(12)：1667–1670.

[24] YI S K，ZHOU X Y，LI J，et al. Full-length transcriptome of *Misgurnus anguillicaudatus* provides insights into evolution of genus *Misgurnus*[J]. Scientific Reports，2018，8：11699.

[25] SONESON C，YAO Y，BRATUS-NEUENSCHWANDER A，et al. A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes[J]. Nature Communications，2019，10：3359.

[26] LI Q Y，XIANG C L，XU L，et al. SMRT sequencing of a full-length transcriptome reveals transcript variants involved in C18 unsaturated fatty acid biosynthesis and metabolism pathways at chilling temperature in *Pennisetum giganteum*[J]. BMC Genomics，2020，21：52.

[27] LI Y P，DAI C，HU C G，et al. Global identification of alternative splicing via comparative analysis of SMRT- and Illumina-based RNA-seq in strawberry[J]. The Plant Journal，2017，90(1)：164–176.

[28] CHAO Y H，YUAN J B，GUO T，et al. Analysis of transcripts and splice isoforms in *Medicago sativa* L. by single-molecule long-read sequencing[J]. Plant Molecular Biology，2019，99：219–235.

[29] JIA X P，TANG L，MEI X Y，et al. Single-molecule long-read sequencing of the full-length transcriptome of *Rhododendron lapponicum* L[J]. Scientific Reports，2020，10：6755.

[30] WANG Z Y，FANG B P，CHEN J Y，et al. *De novo* assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweet potato(*Ipomoea batatas*)[J]. BMC Genomics，2010，11：726.

[31] WEI W L，QI X Q，WANG L H，et al. Characterization of the sesame(*Sesamum indicum* L.) global transcriptome using Illumina paired-end sequencing and development of EST-SSR markers[J]. BMC Genomics，2011，12：451.

责任编辑：毛友纯
英文编辑：柳 正