秦姣姣,杨玉琦,胡晓艳,等. 银杏转录组数据中 EST－SSR 位点的生物信息学分析[J]. 江苏农业科学,2020,48(3):90－94.
doi:10.15889/j.issn.1002－1302.2020.03.015

银杏转录组数据中 EST－SSR 位点的生物信息学分析

秦姣姣, 杨玉琦, 胡晓艳, 杜淑辉
(山西农业大学林学院,山西太谷 030800)

摘要:银杏(*Ginkgo biloba* L.)是雌雄异株植物,其植株价值因性别不同而异。银杏转录组数据中 EST－SSR 位点的生物信息学分析将为银杏遗传学研究开展提供重要的理论与方法支持。首先通过高通量测序技术获得银杏大、小孢子叶球转录组数据,然后开展数据拼接组装与 EST－SSR 位点挖掘及相应的生物信息学分析。转录组数据处理及拼接组装后共获得 108 307 条 unigenes,然后利用 MISA 软件发掘银杏转录组数据中的 SSR 位点,最终从 8 178 条 unigenes 中检索出 9 668 个 SSR 位点。其中,单核苷酸重复的数量最多,有 5 663 个;其次是二核苷酸和三核苷酸,重复数量分别为 2 471、1 438 个;四核苷酸至六核苷酸重复的数量相对较少,共有 96 个。银杏转录组 EST－SSR 位点共包含 147 种重复基元。在单核苷酸重复中,A 和 T 是优势重复基元类型,分别有 2 808、2 685 个;在二核苷酸重复基元中,AT 与 TA 数量较多,分别为 469、383 个,所占比例为 34.48%。此外,设计得到 6 809 对银杏 EST－SSR 位点特异引物。银杏转录组 EST－SSR 位点的发掘将为银杏遗传图谱构建、遗传性状分析、幼年期性别鉴定方法的建立等提供有力的理论与方法支持。

关键词:银杏;转录组;SSR 位点;重复基元
中图分类号:S664.301　　**文献标志码**:A　　**文章编号**:1002－1302(2020)03－0090－05

微卫星序列(simple sequence repeat,SSR)是指由 1～6 个核苷酸为重复单位组成的串联重复序列,如 T_n、$(AG)_n$、$(ATG)_n$、$(ATGC)_n$ 等,在真核生物基因组中随机分布。不同物种间 SSR 位点的分布差异较大,主要表现在基序类型、重复长度以及在染色体上的分布情况等,从而反映出物种间高水平的等位基因多样性。虽然不同物种间 SSR 位点的差异性较大,但是 SSR 位点两端的序列比较保守,因此可以根据 SSR 位点两端的保守序列设计特异性引物以获得其长度多态性,即 SSR 分子标记。SSR 分子标记技术除了具有操作简单、易检测、共显性、稳定性好等优点外,还具有特异性强、等位基因变异多、受选择压力小等特点[1]。开发 SSR 分子标记的传统方法所需费用高、工作量大,并且成功获得阳性克隆和多态性引物的概率偏低[2]。当前,高通

收稿日期:2019－11－23
基金项目:山西省高等学校大学生创新创业训练计划(编号:2019109);山西省高等学校科技创新项目(编号:2019L0369);山西农业大学博士科研启动项目(编号:2017YJ22);山西省优秀博士来晋工作奖励项目(编号:SXYBKY201742)。
作者简介:秦姣姣(1998—),女,重庆人,主要从事园林植物种质资源评价与保存研究。E－mail:qinjj@163.com。
通信作者:杜淑辉,博士,副教授,主要从事园林植物种质资源研究。E－mail:dshhxy@163.com。

基因于不同生长发育时期的表达[J]. 植物生理学报,2017,53(11):2031－2036.
[12]Xu Z,Jing M,Qu C,et al. Identification and expression analyses of the *alanine aminotransferase*(*AlaAT*) gene family in poplar seedlings [J]. Scientific Reports,2017,7:45933.
[13]Enosawa S,Dozen M,Tada Y,et al. Electron therapy attenuated elevated alanine aminotransferase and oxidative stress values in type 2 diabetes－induced nonalcoholic steatohepatitis of rats[J]. Cell Medicine,2013,6(1/2):63－73.
[14]Tian H,Fu J,Drijber RA,et al. Expression patterns of five genes involved in nitrogen metabolism in two winter wheat (*Triticum aestivum* L.) genotypes with high and low nitrogen utilization efficiencies[J]. Journal of Cereal Science,2015,61:48－54.
[15]董召娣,易　媛,张明伟,等. 春性和半冬性小麦花后旗叶和籽粒氮代谢关键酶活性的差异[J]. 麦类作物学报,2015,35(8):1098－1106.
[16]Mauchline T H,Fowler J E,East A K,et al. Mapping the *Sinorhizobium meliloti* 1021 solute－binding protein－dependent transportome[J]. Proceedings of the National Academy of Sciences of the United States of America,2006,103(47):17933－17938.
[17]Kan C C,Chung T Y,Juo Y A,et al. Glutamine rapidly induces the expression of key transcription factor genes involved in nitrogen and stress responses in rice roots[J]. BMC Genomics,2015,16(1):731.

量测序技术发展迅速,测序成本显著降低,为 SSR 分子标记开发提供了一种全新的方法。转录组测序(RNA - Seq)可以全面快速地获得某一特定组织或器官在特定状态下几乎所有的转录组信息,也可以根据测序结果开发特异 EST - SSR 分子标记[3]。目前,RNA - Seq 技术已在刺梨(*Rosa roxbunghii*)[4]、鱼腥草(*Houttuynia cordata*)[5]、杜仲(*Eucommia ulmoides*)[6]等多种植物上开发出 EST - SSR 分子标记,并应用于多领域遗传分析。

银杏(*Ginkgo biloba* L.)为银杏科银杏属落叶乔木,雌雄异株,有"金色活化石"之称,具有良好的观赏特性与药用价值[7]。银杏种子含有银杏酸等多种生理药理活性物质[8],但银杏种子成熟后外种皮有恶臭[9],易污染环境,故在园林绿化上宜使用雄株。银杏采果园的建设与园林绿化中的资源配置都要求将雌雄株区分开来,然而银杏实生苗在定植后需 15 ~ 20 年才开花,继而才能肉眼分辨出雌雄,这显然不能满足早期定植时对雌雄性别区分的要求。形态特征鉴别法简单易行,但仍处于定性阶段,缺乏准确的定量标准;同工酶法及染色体核型分析法均可靠,但难以应用于大规模实践;分子标记法及特异蛋白方面的研究更为准确,但需更高的科技支撑[10]。因此,开发快捷、有效和可靠的银杏雌雄株早期性别鉴定方法,对银杏的资源配置及实际应用具有重要意义。利用雌雄特异的 EST - SSR 标记位点,已经成功地开发出一种早期鉴别杜仲性别的方法[11]。寻找在银杏雌雄株中存在的与性别相关联的特异 EST - SSR 标记位点,或许可以成为快速准确地鉴别银杏性别的方法,为制定科学合理的配置应用方案提供有力的技术支持。

本研究通过 RNA - Seq 技术对银杏大、小孢子叶球进行转录组测序,对测序数据进行拼接组装后获得 unigenes,再对 unigenes 中包含的 EST - SSR 位点进行分析,明确银杏转录组 EST - SSR 位点的组成和分布特征,为后续银杏遗传分析及早期性别鉴定方法的建立等研究提供理论支持。

1　材料与方法

1.1　银杏来源与总 RNA 提取

银杏采自山东农业大学林学院银杏种质资源圃,选取来自于同一家系的银杏 25 年生雌雄实生苗各 5 株,于 2015 年 3 月取初开的银杏大孢子叶球(雌花)和小孢子叶球(雄花)各 10 个作为试验材料。每棵树采集 2 个样本,每 5 个样本作为 1 个生物学重复,共设置 2 个生物学重复。所采样品用液氮速冻后保存于 - 80 ℃冰箱备用。使用改良的 CTAB 法[12]抽取银杏总 RNA,利用 Nanodrop 2000 和琼脂糖凝胶电泳检测总 RNA 质量和完整性。当总 RNA 的浓度大于 400 ng/μL、28S/18S > 1.8 时,表明所提取的 RNA 符合转录组测序的要求。

1.2　转录组测序及序列拼接组装

利用 NEB Next Ultra™ RNA Library Prep Kit for Illumina(NEB, USA)构建 cDNA 文库,然后利用 Illumina Hiseq 2500 测序平台对构建的 cDNA 文库进行双末端测序。对测序得到的原始序列进行去接头、去低质量读段和去重复等处理后,使用软件 Trinity[13]进行 de novo 组装,最终得到尽可能长的 unigenes。

1.3　银杏转录组 EST - SSR 位点挖掘

利用 MISA 软件[14]对银杏转录组 unigenes 序列进行 EST - SSR 位点搜索,搜索标准如表 1 所示。

表 1　银杏转录组 EST - SSR 位点的搜索标准

重复类型	最少重复次数(次)
单核苷酸重复	10
二核苷酸重复	6
三核苷酸重复	5
四核苷酸重复	5
五核苷酸重复	5
六核苷酸重复	6

1.4　银杏转录组 EST - SSR 位点引物设计

利用 Primer 3.0 进行银杏转录组 EST - SSR 位点引物设计,软件参数设置采用默认值,针对检索到的每一个 EST - SSR 位点同时设计 3 对特异引物供后期试验选择。

2　结果与分析

2.1　总 RNA 质量检测结果

经 Nanodrop 2000 和琼脂糖凝胶电泳检测后发现,总 RNA 浓度为 845.7 ng/μL,28S/18S 为 2.01,表明本研究提取的银杏总 RNA 样品质量高,能够满足后续转录组测序的要求。

2.2　原始序列组装结果

银杏转录组测序原始数据经组装拼接后共得到 108 307 条 unigenes,这些 unigenes 的总长度为 86 212 372 bp,平均长度为 796 bp。序列长度大于

1 000 bp 的 unigenes 有 23 624 条，占全部 unigenes 的 21.81%（图 1）。

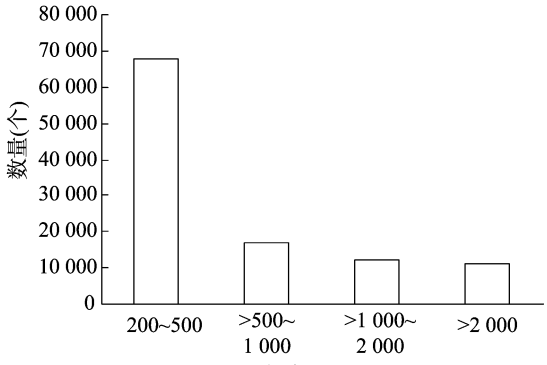

图1　unigenes的长度分布

2.3　银杏 EST－SSR 位点数量分布特征

如表 2 所示，银杏转录组 unigenes 序列经检索后，共发现 8 178 条 unigenes 含有 9 668 个 EST－SSR 位点，占总 unigenes 数量的 7.55%。从银杏转录组 unigenes 中共检索到 6 种核苷酸重复类型，出现数量最多的是单核苷酸重复，占总 EST－SSR 位点数量的 58.57%，其次是二核苷酸重复，占 25.55%，数量最少的是五核苷酸重复，仅占 0.14%。

表 2　银杏转录组 EST－SSR 位点的数量与分布

重复类型	SSR 位点数分布（个）															SSR 数量（个）	百分比（%）	发生频率（%）
	5	6	7	8	9	10	11	12	13	14	15	16	17	18	>18			
单核苷酸重复	—	—	—	—	—	2 730	1 260	669	372	204	132	78	49	48	121	5 663	58.57	5.23
二核苷酸重复	—	813	481	437	356	271	108	5	—	—	—	—	—	—	—	2 471	25.55	2.28
三核苷酸重复	926	343	147	20	—	1	—	—	1	—	—	—	—	—	—	1 438	14.87	1.33
四核苷酸重复	61	7	1	—	—	—	—	—	—	—	—	—	—	—	—	69	0.71	0.06
五核苷酸重复	12	—	—	1	—	—	—	—	—	—	—	—	—	—	—	13	0.14	0.01
六核苷酸重复	—	8	4	1	1	—	—	—	—	—	—	—	—	—	—	14	0.16	0.01
总计	999	1 171	633	459	357	3 002	1 368	674	373	204	132	78	49	48	121	9 668	100.00	8.92

2.4　银杏 EST－SSR 重复基元的分布特征

如图 2 所示，在 9 668 个银杏 EST－SSR 位点中，共有 147 种重复基元出现，其中单核苷酸、二核苷酸、三核苷酸、四核苷酸、五核苷酸及六核苷酸重复基元的种类分别有 4、12、60、44、13、14 种。单核苷酸重复基元中，A 和 T 是优势重复基元类型，分别有 2 808、2 685 个，占单核苷酸重复的 97.00%；二核苷酸重复基元中，出现次数最多的是 AT，有 469 个，占二核苷酸重复的 18.98%，其次是 TA，有 383 个，占 15.50%；三核苷酸重复基元中，出现频率最高的为 GAA，占三核苷酸重复的 4.79%；四核苷酸、五核苷酸和六核苷酸重复基元类型数量最少，占总 EST－SSR 位点数量的 1.01%。

2.5　银杏转录组 EST－SSR 位点引物设计

如表 3 所示，采用 Primer 3.0 软件对本研究检索到的银杏 EST－SSR 位点进行特异引物设计，共得到 6 809 对特异引物，成功率为 70.43%。在设计成功的 6 809 对引物中，扩增产物为单核苷酸重复的最多，有 4 012 个，占 58.92%；其次为二核苷酸和三核苷酸重复基元，分别有 2 413、1 095 个，分别占 35.44%、16.08%。另外，PCR 产物为复合型重复（含有 1 个以上重复基元类型）的有 921 个，占 13.53%。

2.6　银杏转录组 EST－SSR 位点的可用性评价

多态性是判定分子标记可用性的重要参考指标之一，对于 SSR 分子标记来说，长度是影响其多态性高低的一个重要因素。研究表明，当 SSR 长度大于 20 bp 时，此位点具有高度多态性，当长度在 12～20 bp 之间，此位点具有中等水平的多态性，而长度小于 12 bp 的 SSR 位点多态性较低[15]。因此本研究中对银杏转录组 EST－SSR 位点进行搜索时，筛选标准为单核苷酸重复至少 10 次，二核苷酸重复至少 6 次，而三核苷酸至六核苷酸的重复次数要大于 5 次。经统计，银杏转录组 EST－SSR 位点的长度集中分布在 12～45 bp 之间，其中长度大于 20 bp 的 EST－SSR 位点共有 734 个，占总 EST－SSR 位点的 7.59%；长度在 12～20 bp 之间的 EST－SSR 位点有 4 944 个，占总数的 51.14%。Zhang 等研究发现，高级重复基元类型 SSR 位点的多态性要低于低级重复基元类型[16]。在本研究中检索到的银杏转录组 EST－SSR 位点主要是低级重复基元类型 SSR 位点，如单核苷酸、二核苷酸、三核苷酸重复所占比例高达 84.28%，对银杏转录组 EST－SSR 位点长度进行统计分析时发现，长度大于 20 bp 的 734 个 EST－SSR 位点中，单核苷酸、二核苷酸重复基元类型有 471 个，占比达到 64.17%，

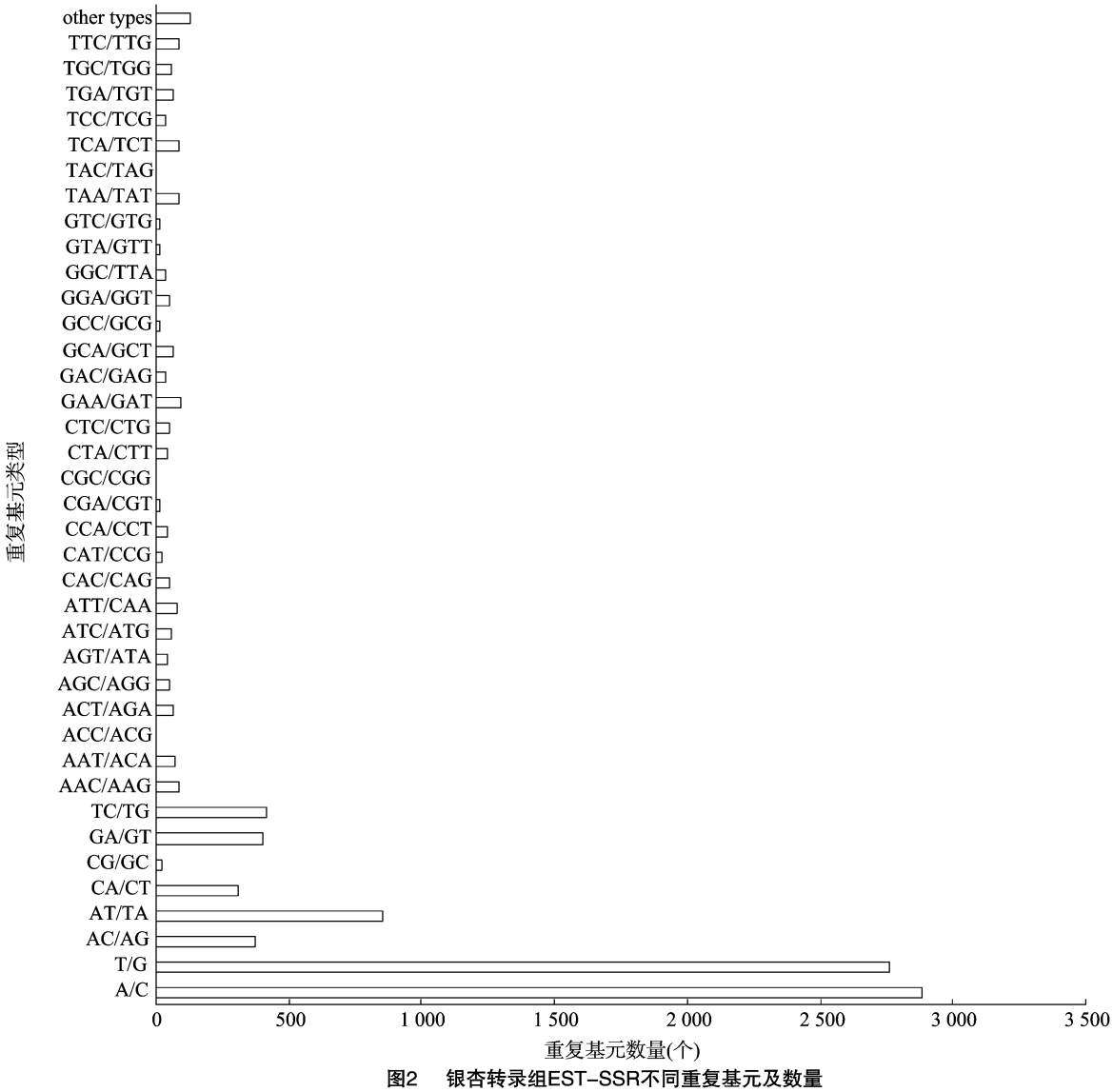

图2　银杏转录组EST-SSR不同重复基元及数量

表 3　PCR 扩增产物中包含重复基元类型

重复类型	数量（个）	比例（%）
简单重复:单核苷酸重复	4 012	58.92
二核苷酸重复	2 413	35.44
三核苷酸重复	1 095	16.08
四核苷酸重复	34	0.50
五核苷酸重复	10	0.15
六核苷酸重复	10	0.15
复合型重复	921	13.53

表明这部分银杏转录组 EST－SSR 位点具有高度多态性潜能,有很好的利用潜质。

3　讨论与结论

SSR 位点广泛分布于真核生物基因组中,据统计,真核生物基因组中每隔 10～50 kb 就存在 1 个 SSR 位点,在植物基因组中,平均每 23.3 kb 就有 1 个 SSR 位点[17]。目前,转录组学研究涉及的物种越来越广泛,尤其是基因组序列还未公布的物种,产生了大量的转录组测序数据,对于这些数据的深度挖掘成为目前研究的热点。本研究通过 RNA－Seq 技术对银杏大、小孢子叶球进行了转录组测序,经过拼接组装后得到 108 307 条 unigenes,检索后得到符合条件的 EST－SSR 位点 9 668 个,出现频率为 8.92%,其出现频率明显高于鱼腥草[5]、云南松[18]等物种,低于刺梨[4]等物种。造成不同物种间 EST－SSR 位点出现频率差异的原因可能是物种间 SSR 位点组成及分布的差异性。银杏转录组中 EST－SSR 位点种类与数量均比较丰富,可为银杏 SSR 分子标记的开发提供重要的参考。

不同物种之间 EST－SSR 位点主要重复类型同

样有所差异。很多植物的 EST－SSR 位点主要以二核苷酸、三核苷酸重复基元类型为主，比如云南松[18]。本研究发现，银杏转录组 EST－SSR 位点重复基元类型主要以单核苷酸重复为主，占全部 SSR 位点的 58.57%，其次是二核苷酸重复，这与红松[19]、白皮松[20]等相似，但与云南松[18]、鱼腥草[5]、刺梨[4]等物种有差异，这些物种 EST－SSR 位点的主要重复基元是三核苷酸重复。SSR 位点基序类型中普遍存在 A/T 优势，而 G/C 重复基序类型出现频率较低，在多数植物中很难发现。导致上述现象的可能原因是打破 A/T 碱基对之间氢键所需的能量要低于 G/C 碱基对，基因组中 A/T 的波动较 G/C 容易[21]。但也有观点认为，基因组甲基化使 C 转化为 T，同时 3′末端 polyA 序列插入形成富含 A 的原始 SSR 位点，导致重复基序中 A/T 优势的出现[22]。本研究发现，银杏转录组 EST－SSR 位点重复基序类型中单核苷酸重复基元类型出现最多的是 A 与 T，两者构成的 SSR 位点占总 SSR 位点数量的 56.82%。其次，二核苷酸重复基元类型中，AT 和 TA 重复基序类型出现次数同样很高，所占 SSR 位点比例分别为 4.85%、3.96%，表现出较明显的 A/T 优势。而 G/C 在所有重复基元类型中的出现频率较低，由 C 和 G 组成的单核苷酸重复基元共 19 个，占总 SSR 位点的 0.16%，二核苷酸重复中 GC 和 CG 所占的比例仅分别为 0.01%、0.003 7%。银杏转录组 EST－SSR 位点不仅出现频率高、平均分布频率广，且类型丰富，具有较高的多态性潜能和可用性。本研究积累了大量银杏 EST－SSR 位点并明确了其基本特征，可为开发银杏 SSR 分子标记奠定重要的理论基础。本研究的开展对于加快银杏功能基因资源的开发利用，建立银杏种质资源评价和改良机制、快速准确的苗期性别鉴定方法等具有重要的意义。

参考文献：

[1]赵　罕，朱高浦，刘梦培，等. 微卫星分子标记及其在林业中的应用[J]. 世界林业研究，2013，26(6)：21－26.

[2]程小毛，黄晓霞. SSR 标记开发及其在植物中的应用[J]. 中国农学通报，2011，27(5)：304－307.

[3]Sharma R，Maloo S R，Choudhary S，et al. Microsatellite markers：an important DNA fingerprinting tool for characterization of crop plants [J]. The Journal of Plant Science Research，2015，31(1)：83.

[4]鄢秀芹，鲁　敏，安华明. 刺梨转录组 SSR 信息分析及其分子标记开发[J]. 园艺学报，2015，42(2)：341－349.

[5]黎晓英，刘胜贵，王　丹，等. 鱼腥草转录组 SSR 位点信息分析及其多态性研究[J]. 中草药，2016，47(10)：1762－1767.

[6]黄海燕，杜红岩，乌云塔娜，等. 基于杜仲转录组序列的 SSR 分子标记的开发[J]. 林业科学，2013，49(5)：176－181.

[7]朱丽峰. 银杏的景观价值及其在园林中的应用[J]. 林业调查规划，2012，37(1)：112－114.

[8]耿敏章. 银杏中营养成分和功能因子的研究进展[J]. 氨基酸和生物资源，2011，33(1)：63－66，83.

[9]曹福亮，沈国航. 中国银杏志[M]. 北京：中国林业出版社，2007.

[10]黄　茜，刘霁瑶，曹　敏，等. 银杏性别特征表现与鉴别研究进展[J]. 果树学报，2013，30(6)：1065－1071.

[11]林开勤，赵德刚，李　岩，等. 杜仲性别相关 EST－SSR 标记的开发[J]. 林业科学，2016，52(10)：146－152.

[12]许端祥，杜文丽，陈中钐，等. 基于瓠瓜转录组测序的 EST－SSR 标记的开发及其应用[J/OL]. 热带作物学报：1－16[2019－12－19]. http://kns.cnki.net/kcms/detail/46.1019.S.20190912.1409.008.html.

[13]Grabherr M G，Haas B J，Yassour M，et al. Trinity：reconstructing a full－length transcriptome without a genome from RNA－Seq data [J]. Nature Biotechnology，2013，29(7)：644－652.

[14]Faircloth B C. MSATCOMMANDER：detection of microsatellite repeat arrays and automated，locus－specific primer design[J]. Molecular Ecology Resourecs，2008，8(1)：92－94.

[15]Temnykh S，Park W D，Ayres N，et al. Mapping and genome organization of microsatellite sequences in rice(*Oryza sativa* L.) [J]. Theoretical and Applied Genetics，2000，100(5)：697－712.

[16]Zhang P Z，Dreisigacker S，Melchinger A E，et al. Quantifying novel sequence variation and selective advantage in synthetic hexaploid wheats and their backcross－derived lines using SSR markers[J]. Molecular Breeding，2005，15(1)：1－10.

[17]Marathi B，Guleria S，Singh N K，et al. Molecular diversity and segregation distortion measured by SSR markers in a new plant type based recombinant inbred line population of rice[J]. Indian Journal of Genetics and Plant Breeding，2011，71(4)：297－303.

[18]蔡年辉，许玉兰，徐　杨，等. 云南松转录组 SSR 的分布及其序列特征[J]. 云南大学学报(自然科学版)，2015，37(5)：770－778.

[19]张　振，张含国，莫　迟，等. 红松转录组 SSR 分析及 EST－SSR 标记开发[J]. 林业科学 2015，51(8)：114－120.

[20]李昕蔓，金卓颖，苏安然，等. 白皮松 EST－SSR 序列分布特征及引物开发[J]. 林业与生态科学，2019，34(3)：266－272.

[21]Biswas M K，Chai L J，Mayer C，et al. Exploiting BAC－end sequences for the mining，characterization and utility of new short sequences repeat (SSR) markers in *Citrus*[J]. Molecular Biology Reports，2012，39(5)：5373－5386.

[22]Li D J，Deng Z，Qin B，et al. De novo assembly and characterization of bark transcriptome using Illumina sequencing and development of EST－SSR markers in rubber tree (*Hevea brasiliensis* Muell. Arg.) [J]. BMC Genomics，2012，13(1)：192.