周军永,陆丽娟,刘　茂,等. 基于李府贡枣转录组测序的 SSR 和 SNP 特征分析[J]. 江苏农业科学,2019,47(4):51 - 54.
doi:10.15889/j.issn.1002 - 1302.2019.04.011

基于李府贡枣转录组测序的 SSR 和 SNP 特征分析

周军永, 陆丽娟, 刘　茂, 朱淑芳, 仇鹏辉, 孙其宝
(安徽省农业科学院园艺研究所,安徽合肥 230031)

摘要:为了简单重复序列(simple sequence repeats,SSR)和单核苷酸多态性标记(single nucleotide polymorphism,SNP)开发等研究,以李府贡枣不同处理枣果实的转录组序列为基础,分析了转录组数据中 SSR 和 SNP 位点的分布。结果表明:转录组数据共获得了 226 488 条 contig 序列,其中有 42 570 条 unigene 在数据库中得到注释。利用鉴定简单重复序列的软件(MIcroSAtellite identification tool,MISA)进行 SSR 位点的搜索,共得到 18 016 个 SSR 位点,SSR 位点的出现频率为 0.43 个/kb。SSR 位点共包含 164 种重复基元,其中以 A/T 类型为主的单核苷酸重复所占的比例最高(6 942 个,38.44%),其次是 AG/CT 类型为主的二核苷酸重复(6 113 个,33.85%)和以 AAG/CTT 为主的三核苷酸重复(4 242 个,23.49%),四核苷酸重复、五核苷酸重复和六核苷酸重复基本相同。在转录组得到的 unigene 中共发现 SNP 位点 163 360 个,发生频率为 1/254 bp,6 种单核普酸变异中以 Transition 类型的 A/G 和 C/T 发生频率最高,分别为总数的 30.80% 和 30.49%;其他 4 种 Transversion 类型的 SNP 为 C/G、G/T、A/C 和 A/T,分别占到总数的 9.83%、9.78%、9.78% 和 9.32%。其中 Transition 类型显著高于 Transversion 类型,在转换类型中 A/G 和 C/T 发生频率基本一致,但以 A/G 发生频率略高。

关键词:枣;转录组;SSR;SNP;特征分析
中图分类号: S665.101　　**文献标志码:** A　　**文章编号:**1002 - 1302(2019)04 - 0051 - 04

枣(*Ziziphus jujuba* Mill.)具有重要的经济价值和生态价值,在我国栽培历史悠久,是许多省份和地区重要的经济林树种,枣产业成为当地的支柱产业之一。我国枣种质资源丰富、品种繁多,近年来国内外学者利用简单重复序列间扩增(inter - simple sequence repeat,ISSR)[1-2]、扩增片段长度多态性(amplified fragment length polymorphism,AFLP)[3]等分子标记技术在枣的品种分类、鉴别以及遗传多样性方面开展了相关研究工作。

简单重复序列是由 1 ~ 6 个碱基组成的简单串联重复序列,普遍存在于真核生物基因组[4],SSR 按来源可分为有基因组 SSR 和转录组来源的 SSR[5],与基因组 SSR 相比,转录组来源的 SSR 无须构建基因组文库等工作。SSR 标记具有影响转录、基因调节、蛋白质功能以及基因组构[6-7],被认为是遗传学研究中最理想的分子标记手段之一[8],同时转录组来源的 SSR 反映了基因组的编码区域,直接获得物种基因表达信息,因此 EST - SSR 多态性可能与基因功能直接相关[9]。与常规的 AFLP、随机扩增多态性 DNA(random amplified polymorphism DNA,RAPD)、ISSR 等分子标记相比,SSR 标记具有数量丰富、分布广泛、共显性遗传、多态性丰富等特点,由于枣 SSR 标记开发较晚,目前 SSR 标记已被应用于枣指纹图谱构建、亲缘关系、遗传多样性分析等研究领域[10-11]。SNP 标记是指基因组 DNA 序列中由于单个核普酸替换或较短片段的插入缺失所引起的多态性,以其分布广泛、稳定性强等优点已被广泛应用于的遗传分析领域,SNPs 标记在苹果、西瓜、柑橘、葡萄、柿等作物中得到了开发和利用[12-13]。目前,在枣中开发了一些基因组 SSR[14] 和转录组 SSR 标记[15],分别在基因组和转录组水平上分析了枣微卫星的特点;而枣 SNP 研究处于标记发现阶段,研究报道较少。

本研究利用转录组测序技术对李府贡枣不同处理果实进行转录组测序和数据组装,通过分析其特征为 SSR 和 SNP 标记的开发和利用提供生物信息学基础,同时为枣遗传结构和遗传分化以及构建遗传图谱奠定基础,也将为其功能基因的开发利用、比较基因组学的研究等提供依据。

1　材料与方法

1.1　研究材料

材料取自安徽省农业科学院园艺研究所枣种质资源圃,2016 年选取树龄 5 年,处于盛果期的李府贡枣为试材,当枣果实进入白熟期后进行灌水处理,分别设置 ZJ(未灌水)、ZJ1(灌水后 8 h)、ZJ2(灌水后 30 h)、ZJ3(开裂)等 4 个处理,处理后分别采集果实表皮各 2 份,液氮冷冻后在 -80 ℃ 保存。

1.2　总 RNA 的提取

采用 TRIzol 试剂提取枣果实总核糖核苷酸(RNA),提取后用琼脂糖凝胶电泳检测,然后利用安捷伦 2100 芯片生物分析仪(Agilent 2100 Bioanalyzer)检测提取的 RNA 是否达到转录组测序(RNA - Seq)的试验标准。

收稿日期:2017 - 12 - 03
基金项目:安徽省农业科学院院长青年创新基金(编号:16B0306);科技部科技基础性工作专项(编号:2012FY110100);安徽省一般性转移支付科技项目“黄营灵枣新品种选育及配套技术研究与应用”“玉铃铛枣节本优质绿色生产技术研究与应用”。
作者简介:周军永(1983—),男,山东淄博人,硕士,助理研究员,主要从事果树育种与栽培等研究。E - mail:coplmn@163.com。
通信作者:孙其宝,硕士,研究员,主要从事果树育种与栽培等研究。E - mail:ansqb@163.com。

1.3　转录组测序及数据组装

提取的枣果皮总 RNA 经脱氧核糖核酸酶Ⅰ(DNase Ⅰ)处理后,用带有多聚胸腺嘧啶[Oligo(dT)]的磁珠富集真核生物信使核糖核酸(messenger RNA,mRNA)。然后加入打断试剂将 mRNA 打断成短片段,并以打断后的 mRNA 为模板用六碱基随机引物(random hexamers)合成 1 链互补脱氧核糖核酸(complementary deoxyribonucleic acid,cDNA),加入缓冲液、三磷酸碱基脱氧核苷酸(deoxyribonucleoside triphosphates,dNTPs)和 DNA 聚合酶Ⅰ(DNA polymerase Ⅰ)合成 cDNA 第 2 链,经试剂盒纯化回收、黏性末端修复、3′末端加上碱基“A”和连接测序接头,再将得到的片段进行大小选择后 PCR 扩增富集。构建好的文库经 Agilent 2100 Bioanalyzer 和美国应用生物系统公司的实时荧光定量 PCR 仪(ABI StepOnePlus Real – Time PCR System)质检合格后使用 Illumina 测序平台进行测序。转录组测序工作由深圳市恒创基因科技有限公司完成。对 4 份枣果皮样品测序得到的原始数据过滤掉里面含有带接头的、低质量的测序序列(read)得到干净序列(clean reads)。利用转录组 Trinity 组装软件对所有样品的干净序列进行混合拼接成转录本序列,取每条基因中最长的转录本为基因组数据库,得到的基因组数据库数据库用于后续分析。

1.4　SSR 和 SNP 分析方法

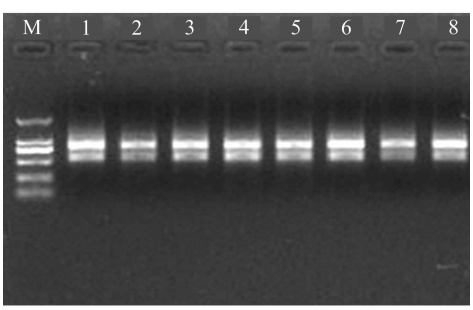
SSR 位点搜索主要是利用 MISA 软件(http://pgrc. ipk – gatersleben. de/misa/)搜索得到基因组数据库,其参数设置:单碱基、二碱基、三碱基、四碱基、五碱基、六碱基的最短重复次数分别为 12、6、5、5、4、4。

SNP 位点的搜索通过 Samtool 和 Picard – tools 等工具对比对结果进行染色体坐标排序、去掉重复的序列等处理,最后通过变异检测软件 UATK3 进行单核苷酸多态性标记调用(SNP calling),并对原始结果进行过滤。

2　结果与分析

2.1　RNA 质量检测

提取的总 RNA 样品先进行电泳检测,结果如图 1 所示,28S 和 18S 条带明亮,无杂质。



M—2 000 bp marker; 1~4—4 个样品处理,
5~8—4 个样品处理的重复。

图1　李府贡枣总 RNA 琼脂糖电泳

Aligent2100 检测总 RNA 样品质量,RNA 完整值(RNA integrity number,RIN)都在 7.0 ~ 8.0 之间,总 RNA 的浓度和总量等指标均已达到测序要求,可用于后续转录组测序等试验(表 1)。

表 1　样品 RNA 检测指标

处理	浓度(mg/L)	总量(μg)	$D_{260\ nm}/D_{230\ nm}$	$D_{260\ nm}/D_{280\ nm}$	28S/18S	RIN
ZJ	405	18.2	0.6	1.8	1.0	7.0
ZJ1	510	24.5	0.3	1.8	1.1	8.0
ZJ2	430	24.1	0.6	1.9	1.2	7.6
ZJ3	380	19.0	0.2	1.8	1.0	7.8

2.2　转录组数据组装结果及统计

枣果实转录组测序共获得 41 471 760 条干净序列,对干净序列进行组装拼接获得 226 488 条拼接序列。拼接序列长度范围主要分布在 200 ~ 2 000 bp 之间,其中以 200 ~ 300 bp 序列数量居多,约占总拼接序列的 61.70%,大于 2 000 bp 的序列约占总拼接序列的 4.64%(表 2)。

表 2　李府贡枣转录组测序组装结果

长度范围(bp)	拼接序列数量(条)	所占比例(%)
200 ~ <300	139 741	61.70
300 ~ <500	26 681	11.78
500 ~ <1 000	24 555	10.84
1 000 ~2 000	25 004	11.04
>2 000	10 507	4.64
总数(个)	226 488	
总长度(bp)	116 726 218	
N50 长度(bp)	1 177	
平均长度(bp)	515	

组装拼接获得 42 570 条基因组数据库,序列长度主要分布在 300 ~ 3 000 bp 范围内,平均长度为 974 bp。300 ~ 2 000 bp 序列数量最多,占全部基因组数据库序列的 87.39%;2 000 ~ 3 000 bp 的基因组数据库序列有 3 651 条,占全部基因组数据库序列的 8.58%;≥3 000 bp 的基因组数据库序列有 1 719 条,占 4.04%(表 3)。

表 3　李府贡枣转录组基因组数据库长度统计情况

长度范围(bp)	基因组数据库数量(条)	所占比例(%)
300 ~ <400	16 340	38.38
400 ~ <600	4 919	11.56
600 ~ <800	3 308	7.77
800 ~ <1 000	2 715	6.38
1 000 ~ <1 200	2 402	5.64
1 200 ~ <1 400	2 267	5.33
1 400 ~ <1 600	2 081	4.89
1 600 ~ <1 800	1 742	4.09
1 800 ~ <2 000	1 426	3.35
2 000 ~ <2 200	1 154	2.71
2 200 ~ <2 400	888	2.09
2 400 ~ <2 600	686	1.61
2 600 ~ <2 800	508	1.19
2 800 ~ <3 000	415	0.97
≥3 000	1 719	4.04
总数(条)	42 570	
总长度(bp)	41 447 083	
N50 长度(bp)	1 631	
平均长度(bp)	974	

2.3　微卫星特征分析

2.3.1　微卫星数量及分布特点　在转录组的 42 570 条基因组数据库序列中发现 18 016 个 SSR 位点，其中包含 1 442 个混合型 SSR 和 13 033 个完整型 SSR 位点，完整型 SSR 占总 SSR 位点的 72.3%，包含 2 个及以上 SSR 位点的基因组数据库共有 3 762 条。SSR 位点的出现频率为 0.43 个/kb，即每 2.3 kb 就出现 1 个 SSR 位点。

SSR 位点共包含 164 种重复基元，单核苷酸至六核苷酸分别有 2、4、10、19、32、97 种。其中 SSR 重复基元的重复次数均在 4～35 次，重复 4～10 次的 SSR 位点共有 10 606 个，占总 SSR 的 58.87%，主要为二核苷酸和三核苷酸；重复 11～16 次的 SSR 位点有 3 780 个，占 20.98%，主要为单核苷酸和二核苷酸；重复 17～20、21～35 次的 SSR 位点基本为单核苷酸（表 4）。

表 4　不同 SSR 重复基元的频率

重复次数	SSR 位点数（个）					
	单核苷酸	二核苷酸	三核苷酸	四核苷酸	五核苷酸	六核苷酸
4～10	0	5 655	4 233	219	242	257
11～16	3 316	458	5	0	0	1
17～20	3 040	0	3	0	0	0
21～35	586	0	1	0	0	0

在微卫星中，单核苷酸重复（6 942 个，38.53%）最多，其次是二核苷酸重复（6 113 个，33.93%）和三核苷酸重复（4 242 个，23.55%），四核苷酸重复、五核苷酸重复和六核苷酸重复基本相同（219、242、258 个）（图 2）。

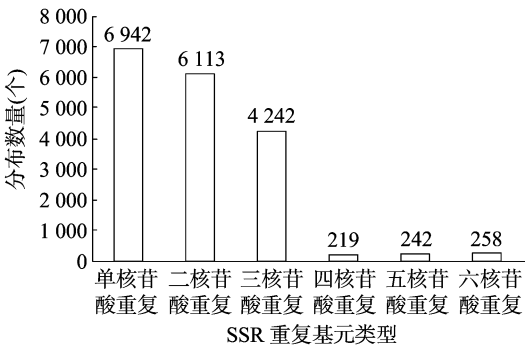


图2　李府贡枣转录组 SSR 重复基元分布

2.3.2　微卫星不同优势重复单元碱基的特征分析　SSR 位点共包含 164 种重复基元，单核苷酸至六核苷酸分别有 2、4、10、19、32、97 种。通过对枣不同类型 SSR 重复单元数量的变化的统计得出频率最高的 4 类基序，依次为 A/T（6 871 个，38.14%）、AG/CT（3 713 个，20.61%）、AT/AT（1 998 个，11.09%）和 AAG/CTT（1 462 个，8.12%）。

在 2 种单核苷酸重复微卫星中，以 A/T 为最主要的重复单元，共有 6 871 个，占 98.98%，而 C/G 只占 1.02%。

二核苷酸重复类型有 4 种（AC/GT、AG/CT、AT/AT 和 CG/CG），其中 AG/CT 重复的数量最多，共有 3 713 个，占二核苷酸重复微卫星总数的 60.74%；其次是 AT/AT（1 998 个），占 32.68%；再次是 AC/GT（396 个），占 6.48%；而 CG/CG 只有 6 个，占 0.10%（图 3）。

三核苷酸重复类型有 10 种，AAG/CTT 重复的数量最多，共有 1 462 个，占 4.46%；其次是 AAT/ATT（645 个）、ACC/GGT（521 个）、ATC/ATG（477 个）；再次是 AAC/GTT（360 个）、AGG/CCT（298 个）、AGC/CTG（283 个），其他重复类型则相对较少。

在 19 种四核苷酸重复类型中，以 AAAT/ATTT 重复数量最多，共 113 个，占四核苷酸 SSR 总数的 51.60%；其次为 AAAG/CTTT，有 27 个，占 12.33%。五核苷酸重复类型有 32 种，AAAAT/ATTTT 重复数量最多，有 104 个，占 42.98%。六核苷酸重复类型有 97 种，共 258 个，但每种重复类型数量都较少。

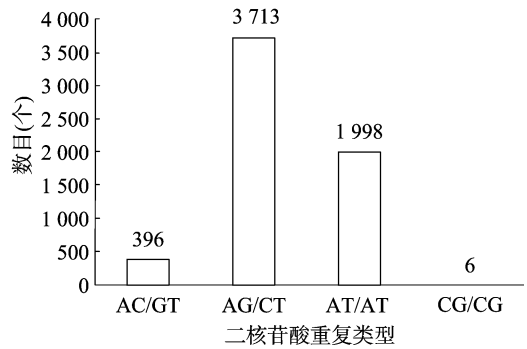


图3　二核苷酸重复类型的分布情况

通过对枣果实转录组微卫星数量分析可知，单核苷酸重复次数主要集中在 12～20 次，且随着重复次数增加呈递减趋势，未发现重复 24 次以上的单核苷酸微卫星序列。二核苷酸微卫星重复次数集中在 6～11 次；三核苷酸微卫星重复次数集中在 5～8 次；四核苷酸微卫星重复次数集中在 5～6 次；而五核苷酸微卫星和六核苷酸微卫星重复次数最少，为 4～5 次。

2.3.3　微卫星长度分布　微卫星长度也存在极显著变异，长度变化范围为 12～248 bp，平均长度为 21 bp。以重复长度为 10～20 bp 的短序列最多，占 80.12%；其次为长度在 21～29 bp 的序列，占总数的 12.18%；长度大于 50 bp 的长序列占微卫星总数的 4.36%（图 4）。

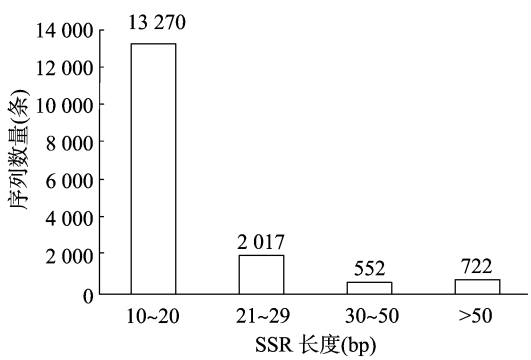


图4　李府贡枣转录组 SSR 序列长度分布

2.4 SNP 位点的特征分析

在转录组得到的基因组数据库中共发现 SNP 位点 163 360 个,发生频率为 1/254 bp,即每 254 bp 就会有 1 个 SNP 位点出现,其中转换 100 122 个,颠换 63 238 个。6 种单核苷酸变异中以转换类型的 A/G 和 C/T 发生频率最高,分别为总数的 30.80% 和 30.49%;其他 4 种颠换类型的 SNP 为 C/G、G/T、A/C 和 A/T,分别占到总数的 9.83%、9.78%、9.78% 和 9.32%。其中转换类型显著高于颠换类型,在转换类型中 A/G 和 C/T 发生频率基本一致,但以 A/G 发生频率略高。

3 结论与讨论

在李府贡枣转录组的 42 570 条基因组数据库序列中发现 18 016 个 SSR,其中包含 1 442 个混合型 SSR 和 13 033 个完整型 SSR 位点,SSR 位点的出现频率为 0.43 个/kb,比桃(0.31)、枣(0.36)出现频率[15-16]低,与柿 SSR 位点出现频率[13]相同,表明本研究中李府贡枣 SSR 标记的数量极其丰富,有望在 SSR 引物开发、遗传多样性等领域得到广泛应用。

本研究通过转录组获得的微卫星中单核苷酸重复最多,占 38.44%;其次是二核苷酸重复(33.85%)和三核苷酸重复(23.49%),四核苷酸重复、五核苷酸重复和六核苷酸重复基本相同,与前人关于枣转录组微卫星特征基本相同,但本研究获得 258 个六核苷酸重复类型。基因组序列的微卫星特征与转录组微卫星序列相比,六碱基重复微卫星出现的频率明显高于其他类型,枣转录组比基因组低级基元频率高,而高级基元比基因组的低,与前人研究[14-15]基本一致。

SSR 位点共包含 164 种重复基元,单核苷酸至六核苷酸分别有 2、4、10、19、32、97 种。其中 SSR 重复基元的重复次数均在 4~35 次,重复 4~10 次的 SSR 位点共有 10 606 个,占总 SSR 的 58.87%,主要为二核苷酸和三核苷酸;重复 11~16 次的 SSR 位点有 3 780 个,占 20.98%,主要为单核苷酸和二核苷酸;重复 17~20 次和 21~35 次的 SSR 位点基本为单核苷酸。SSR 长度变化范围为 10~248 bp,平均长度为 21 bp,以重复长度为 10~20 bp 的短序列最多,占 80.07%。

通过对本研究结果分析可知,单核苷酸重复微卫星为枣最优势微卫星,所占比例最多,而且单核苷酸微卫星重复单元次数的变化明显高于其他重复类型,其次是二核苷酸微卫星,说明单核苷酸在整个枣转录组中变异最为活跃。此外,SSR 序列以重复长度为 10~20 bp 的短序列最多,此类 SSR 位点拥有高度多态性。SSR 的长度和重复次数是影响分子标记多态性的重要因素[17],说明转录组获得的 SSR 位点可为枣遗传多样性和亲缘关系等研究有重要的价值。

单核苷酸多态性在植物基因组中广泛存在[18-19]。本研究中共发现 SNP 位点 163 360 个,发生频率为 1/254 bp,与柿发生频率[13]基本一致,但与水稻和玉米等作物相比发生频率低。所获得的 SNP 位点中 Transition 类型显著高于 Transversion 类型。6 种单核普酸变异中以 Transition 类型的 A/G 和 C/T 发生频率最高。转录组来源的 SSR、SNP 多位于基因组的编码区域,可直接获得物种基因表达信息,可能与基因功能直接相关,转录组测序结果为 SSR 和 SNP 标记的开发和利用提供生物信息学基础,同时为枣遗传结构和遗传分化以及构建遗传图谱奠定基础,也将为其功能基因的开发利用、比较基因组学、分子辅助育种等研究提供依据。

参考文献:

[1]孙 俊,孙雯雯,周军永,等. 安徽及周边地区枣种质资源遗传多样性研究[J]. 园艺学报,2015,42(8):1569-1575.

[2]原勤勤,文亚峰,刘 儒,等. 枣优良品种亲缘关系的 ISSR 分析[J]. 经济林研究,2012,30(1):56-61.

[3]王永康,田建保,王永勤,等. 枣树品种品系的 AFLP 分析[J]. 果树学报,2007,24(2):146-150.

[4]Mrazek J, Guo X, Shah A. Simple sequence repeats inprokaryotic genomes[J]. PNAS,2007,10(4):8472-8477.

[5]王 东,曹玲亚,高建平. 党参转录组中 SSR 位点信息分析[J]. 中草药,2014,46(8):2390-2394.

[6]Kashi Y, King D G. Simple sequence repeat as advantageous mutators in evolution[J]. Trents in Gentic,2006,22(5):253-259.

[7]Lawson M J, Zhang L. Patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes[J]. Genome Biology,2006,7(2):R14.

[8]Liu T, Zhu S, Fu L, et al. Development and characterization of 1827 expressed sequence tag-derived simple sequence repeat markers for ramie (*Boehmeria nivea* L. Gaud)[J]. PLoS One, 2013, 8(4):e60346.

[9]Eujayl I, Sorrells M, Banm M, et al. Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat[J]. Theoretical and Applied Genetics,2002,104(2):399-407.

[10]麻丽颖,孔德仓,刘华波,等. 36 份枣品种 SSR 指纹图谱的构[J]. 园艺学报,2012,39(4):647-654.

[11]刘秀云,李 慧,刘志国,等. 基于 SSR 标记的 255 个枣品种亲缘关系和群体遗传结构分析[J]. 中国农业科学,2016,49(14):2772-2791.

[12]姚丹青,楼坚锋,顾芹芹. SNP 在农作物遗传分析中的应用[J]. 上海农业科技,2015,6:26-27.

[13]杜改改,孙 鹏,索玉静,等. 基于柿雌雄花芽转录组测序的 SSR 和 SNP 多态性分析[J]. 中国农业大学学报,2017,22(10):45-55.

[14]马秋月,戴晓港,陈赢男,等. 枣基因组的微卫星特征[J]. 林业科学,2013,49(12):81-87.

[15]魏琦琦,林 青,贾宝光,等. 枣转录组序列的微卫星特征分析[J]. 中南林业科技大学学报,2015,35(6):93-97.

[16]Wang L, Zhao S, Gu C. Deep RNA-Seq uncovers the peach transcriptome landscape[J]. Plant Molecular Biology,2013,83(4/5):365-377.

[17]赵雅楠,王 颖,张东杰,等. 小豆 SSR-PCR 反应体系优化及引物筛选[J]. 江苏农业科学,2017,45(11):33-37.

[18]雷 雨,张雪芳,罗鑫磊,等. 不同成熟期桃品种 *NAC* 基因遗传多样性研究[J]. 江苏农业科学,2017,45(22):46-49.

[19]李贝贝,刘崇怀,姜建福,等. 葡萄品种分子鉴定研究进展及展望[J]. 江苏农业科学,2017,45(15):15-20.