

张 宁,尹美强,谭青青,等. 苦参转录组 SSR 位点及基因功能注释分析[J]. 江苏农业科学,2019,47(7):41-44.
doi:10.15889/j.issn.1002-1302.2019.07.011

苦参转录组 SSR 位点及基因功能注释分析

张 宁¹,尹美强¹,谭青青²,温银元¹,王玉国¹,王金荣¹

(1. 山西农业大学农学院,山西太谷 030801; 2. 西北大学生命科学学院,陕西西安 710069)

摘要:分析苦参转录组中的简单重复序列(SSR)位点信息,为开发分子标记奠定基础。利用 Fastqc 软件对苦参转录组测序的原始读长(reads)进行质量评估,再用 Trimmomatic 软件对 reads 质量较差的碱基进行过滤,利用 Trinity 软件对 Trimmomatic 处理后的 reads 进行序列组装,之后使用基因组完整性评估(BUSCO)软件对转录组组装的序列进行质量评估,并分析组装的 contig 序列的开放阅读框(open reading frame,简称 ORF);利用 MicroSAteLLite(MISA)软件对无冗余独立基因(unigene)进行 SSR 搜索。利用 Trinity 软件最终筛选得到 23074 条 ORF 信息;使用 MISA 软件从 unigenes 序列中发现 8 798 个 SSR 位点,分布于 7 339 条 unigene 中,总体上 unigenes 序列中 SSR 占比为 2.16%,SSR 位点平均间隔是 5.28 bp,其中占比最高的是单核苷重复基序,为 50.53%;其次是出现频率分别为 22.28%、24.73% 的二、三核苷酸。苦参转录组中 SSR 类型众多,出现频率高,在后续的苦参遗传性状分析,及次生代谢(苦参碱和黄酮等次生代谢产物)途径等相关基因定位等方面具有很好的应用潜力。

关键词:苦参;转录组;SSR;位点信息;基因功能;分子标记

中图分类号:R285 **文献标志码:**A **文章编号:**1002-1302(2019)07-0041-04

苦参(*Sophora flavescens* Ait.)是豆科槐属植物,以其干燥根入药,味苦,性寒,具有清热除燥湿、杀虫和利尿等药效。其主要药用成分是生物碱类和黄酮类化合物,已从苦参中分离出生物碱类 39 个,黄酮类 122 个成分^[1]。苦参主产于山西、陕西、河南、河北等地,在医学临床、农业、畜牧业和日用品等中有广泛的应用^[2]。气候的变化和人为过度的采挖造成野生苦参资源数量急剧减少^[3]。因此,保护和利用好野生苦参资源是当务之急,势在必行。

分子标记开发可对制定合理有效的种质资源保护策略提供科学依据,但目前还缺乏能够应用于苦参种质鉴定、遗传图谱构建、功能基因定位等研究的简便、高效、稳定且具有种属特异性的分子标记体系。简单重复序列(simple sequence repeat,简称 SSR)是由核苷酸构成的重复序列,在真核生物和原核生物基因中都有存在。SSR 位点标记具有在生物中分布

广泛、重复类型多样、出现频度高等特点^[4],主要应用于分子育种优良基因定位、生物多样性分析、遗传图谱绘制、突变体单核苷酸多态性(single nucleotide polymorphism,简称 SNP)位点分析辅助等。传统寻找基因组中 SSR 标记的方法存在位点开发成本高、步骤较多、操作繁琐等问题^[5]。转录组 SSR 位点开发具有方便快捷、效率高等特点,且成本低廉。SSR 开发引物能够直接快速地定位基因信息。随着苦参研究的深入,目前还未发现有关苦参转录组 SSR 开发的报道。本研究通过分析苦参转录组中的 SSR 位点信息,为苦参遗传性状分析、次生代谢(苦参碱和黄酮等次生代谢产物)途径、分子标记辅助育种及苦参遗传多样性研究提供依据和参考。

1 材料与方法

1.1 转录组数据来源

从 NCBI(美国国家生物技术中心)数据共享平台获得苦参转录组数据,从 SRA(Sequence Read Archive)数据库(<https://www.ncbi.nlm.nih.gov/sra/>)获得苦参叶片 RNA-Seq 原始测序数据,下载编号是 SAMD00029896,使用 Illumina HiSeq1000 对苦参组织进行建库测序,原始数据 reads 为 90 bp,采取双端(paired-end sequencing)测序,获得 1.3 GB 转录组数据,下载网址是 <ftp://ftp.ncbi.nlm.nih.gov> 中的 DRR031281^[6]。

2014,345(6199):950-953.

[13] Zeng Z B. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci [J]. Proceedings of the National Academy of Sciences of the United States of America, 1993, 90(23):10972-10976.

[14] Zeng Z B. Precision mapping of quantitative trait loci [J]. Genetics, 1994, 136(4):1457-1468.

[15] McCouch S R, Cho Y G, Yano M, et al. Report on QTL

nomenclature [J]. Rice Genet News, 1997, 14(11):11-13.

[16] Shi J, Li R, Qiu D, et al. Unraveling the complex trait of crop yield with quantitative trait loci mapping in *Brassica napus* [J]. Genetics, 2009, 182(3):851-861.

[17] 高必军. 甘蓝型油菜 *napin* 基因启动子的克隆与几个重要农艺性状的初步 QTL 定位 [D]. 雅安:四川农业大学, 2007:1-121.

[18] 吴建忠. 甘蓝型油菜结实相关性状分析及 QTL 定位 [D]. 武汉:华中农业大学, 2010:1-60.

1.2 转录组的从头组装

首先通过 Sratoolkit. 2. 8. 2 - 1 将 sra 格式转录组原始数据转换为 fastq 格式^[7]; 使用 Fastqc 软件进行转录组原始数据质量评估, 然后, 利用 Trimmomatic 软件对 fastq 格式的序列进行低质量去除, leading 头部去掉质量低于 3 的碱基, trailing 尾部过滤掉质量低于 3 的碱基, 每 4 个碱基是一个阅读框, 4 个连续碱基的平均质量低于 15 的过滤掉, reads 中最小长度小于 40 序列的过滤掉^[8]; 随后, 对高质量 reads 采用 Trinity 软件进行从头 (de novo) 组装^[9], 最短 contig 长度设置为 200 bp (参数为默认参数)。筛选每个基因最长的转录本作为 unigene, 最后组装得到苦参转录组的全部转录本 (包含可变剪切)。

1.3 苦参转录组数据组装完整性评估

选取由 Trinity 软件组装的序列, 使用 BUSCO V 2. 0. 1 软件进行苦参叶片转录组数据完整性评价^[10]。BUSCO V 2. 0. 1 软件依据 Ortho DB 数据库, 组成了几个大的进化分支单拷贝基因集, 将转录本 reads 拼接结果与该基因集数据进行比较 (基因集直接使用 HMMER3 与参考数据库比对), 依据比对上的比例、完整性评估拼接结果的准确性和完整性。

1.4 ORF 预测

使用 Trinity 软件中的 TransDecoder LongOrfs 工具对 unigene 进行开放阅读框 (open reading frame, 简称 ORF) 预测, 筛选大于 100 个氨基酸的 ORF 序列, 获得最佳的 ORF 区域, 使用 Pfam (<http://pfam.xfam.org/>) 和 UniProt (<http://www.uniprot.org>) 数据库对预测结果进行校正, 将比对结果保留到 Pfam 和 UniProt 数据库的蛋白质序列中^[11]。

1.5 SSR 位点搜索

使用 MISA 软件^[12]对苦参转录组数据 unigene 的 SSR 位点进行定位搜索, 查询定位规则是三碱基、四碱基、五碱基和六碱基重复至少 5 次, 二碱基重复不得少于 6 次, 2 个 SSR 位点之间不足 100bp 则视为复合型 SSR。

1.6 含 SSR 序列的基因功能注释及生物碱基因挖掘

通过 diamond blastx 和 diamond blastp 分别将苦参中含 SSR 的 8248 条 unigene 序列与 uniprot_sprot、Pfam 和 eggnog、Kegg、基因本体论 (gene ontology, 简称 GO) 等数据库进行比对, 比对参数 *e* 值 < 10⁻⁵, 然后利用 WEGO ([http://wego.](http://wego.genomics.org.cn/)

[genomics.org.cn/](http://wego.genomics.org.cn/)) 在线分析工具进行 GO 功能分类统计, 分析含有 SSR unigene 的功能分布特征; 通过与 GO 库进行比对后, 得到的 unigene 注释结果按照 GO 数据库的 23 个类别进行分类统计。通过对 WEGO 注释结果 (3 个大类) 23 个子类更深入分析挖掘苦参碱相关基因, 为进一步研究奠定基础。

2 结果与分析

2.1 苦参转录组 de novo 组装

从 NCBI 数据库下载得到的苦参转录组测序 (RNA - Seq) 数据中共包含 14 636 096 个双端测序 reads, 通过 Trimmomatic 软件过滤得到 14 578 802 个高质量 reads。转录组 de novo 组装获得 53 179 个长度大于 200 bp 的 contigs, 拼接获得的长序列 (contigs) 平均长度为 813 bp, 最长的 contig 为 22 546 bp, N50 为 1 464 bp; 筛选每个基因中最长的转录本, 共得到 54 221 条 unigenes, 平均长度为 715. 87 bp, 最长的 unigene 为 12 122 bp, N50 为 1 464 bp (表 1)。采用 TransDecoder 软件中 LongOrfs 功能进行 ORF 预测, 筛选获得大于 100 个氨基酸的 ORF 有 29 226 个 contigs; 通过 UniProt 蛋白质数据库比对获得 15 242 条蛋白质序列, Pfam 数据库比对获得 126 429 条蛋白质序列; 使用 TransDecoder 最终筛选得到 23 074 条 ORF 信息。

表 1 转录组拼接结果

类型	数量
N20 长度	2 590 条
N30 长度	2 124 条
N50 长度	1 464 条
中间长度	515 bp
平均长度	880.94 bp
总计核苷酸长度	65 599 017 bp

注: N20、N30、N50 分别表示总的转录本长度大于 20、30、50 kb 的条数。

contigs 和 unigenes 的鸟嘌呤 (G) 和胞嘧啶 (C) 占比都是 44. 8%。从序列长度分布看, 序列长度分布在 1 000 ~ 2 899 bp 的序列大约有 19. 3%, ≥ 2 900 bp 的序列只有 0. 2%, 600 ~ 999 bp 的序列大约有 13. 6%, 700bp 以下占 71. 4% (图 1)。

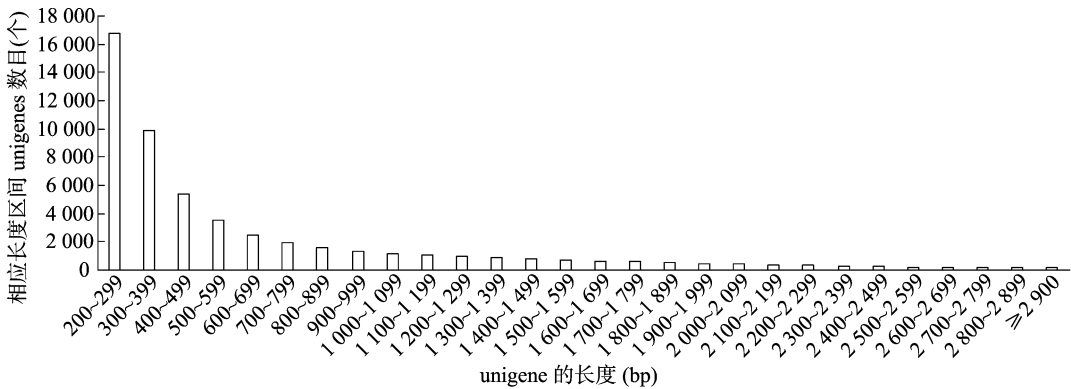


图1 序列长度分布

2.2 转录组数据完整性评估

对转录组数据进行评估, 测序、组装得到的转录序列覆盖所有可能的转录本。评估转录组数据的大小和完整性。依据植物直系同源基因数据集对苦参的转录组数据完整性进行评

估, 由表 2 可知, 在由苦参转录组序列与植物基因组匹配获得的 1440 个植物单拷贝直系同源基因中, 完全匹配到的直系同源基因 (complete) 有 1000 个, 占总 BUSCO 的 69. 4%, 部分片段匹配对应的单拷贝直系同源基因 (fragment) 有 171 个,

占总 BUSCO 的 11.9% ;没有匹配对应到的植物单拷贝直系同源基因(missing)有 269 个,占总 BUSCO 的 18.7% ,完全匹配到的单拷贝直系同源基因 (complete) 有 973 个, 占总 BUSCO 的 67.6% ,完全匹配到的多拷贝直系同源基因 (complete) 有 27 个, 占总 BUSCO 的 1.9% 。

表 2 转录组和基因组组装完整性评估结果统计

类型	BUSCO 匹配结果 (条)	BUSCO 匹配占比 (%)
完全匹配的 BUSCOs (C)	1 000	69.4
完全匹配的单拷贝 BUSCOs (S)	973	67.6
完全匹配且有副本的 BUSCOs (D)	27	1.9
BUSCOs 部分片段 (F)	171	11.9
缺失的 BUSCOs (M)	269	18.7

2.4 转录组中 SSR 位点的分布特点

使用 Trinity 软件组装得到 54 221 条 unigenes,碱基数为 38 815 308 bp,平均每条 unigene 长度为 715.87 bp;使用 MISA 软件搜索得到 8 798 个 SSR 位点,存在于 7 339 条

unigenes 转录组序列中,包括多个 SSR 位点的 unigenes 序列有 1 173 条(包含复合 SSR 为 551 个)占 SSR unigenes 序列总数的 13.33%。总体上 unigenes 序列中 SSR 占比为 2.16% ,SSR 位点平均间隔距离是 4 411 bp。其中占比最高的是单核苷重复基序,占总 SSR 的 50.53% ;其次是出现频率分别为 22.28%、24.73% 的二、三核苷酸。SSR 最短平均分布距离是 0.99 bp 的单核苷酸重复类型,平均分布距离最长的是 1.29 bp 的五核苷酸重复类型。

苦参转录组不同重复类型的 SSR 位点都有多种基元,在考虑碱基互补且包含复合重复基元的情况下,重复类型合计 93 种,其中六核苷酸 38 种,五核苷酸 22 种,四核苷酸类型 17 种,在筛选的 SSR 中单核苷酸重复优势基元为 A/T,占比最高,为总基元类型的 98.18% ,其次是二核苷酸重复类型优势基元 AG/CT,为 65.72%。三核苷酸重复类型的优势基元是 AAG/CTT,占比 27.70% ;四、五、六核苷酸重复类型的优势基元分别是 AAAG/CTTT、AACAC/GTGTT、AGAGGG/CCCTCT,所占的比例分别是 24.17%、11.90%、7.94% (表 3)。

表 3 苦参转录组中 SSR 不同重复类型的分布特征

SSR 类型	SSR 数目 (个)	占总 SSR 比例 (%)	分布频率 (%)	平均长度 (bp)	SSR 平均分布距离 (bp)	基元种类 (种)	优势基元(占基元类型 总数的百分比)
单核苷酸	4446	50.53	5.10	11.10	0.99	2	A/T(98.18%)
二核苷酸	1960	22.28	1.51	13.03	1.08	4	AG/CT(65.72%)
三核苷酸	2176	24.73	0.72	15.23	1.14	10	AAG/CTT(27.70%)
四核苷酸	111	1.26	0.03	18.20	1.12	17	AAAG/CTTT(24.17%)
五核苷酸	42	0.48	0.01	21.79	1.29	22	AACAC/GTGTT(11.90%)
六核苷酸	63	0.72	0.01	31.14	1.25	38	AGAGGG/CCCTCT(7.94%)

注:平均分布距离由总碱基长度/SSR 数目计算得到;占基元类型总数的百分比由对应优势基元数量/SSR 重复类型总数×100% 得到。

2.5 转录组 SSR 基序重复类型和频率特征

不同重复类型苦参转录组 SSR 位点分布存在差异(表 4)。单核苷酸重复类型设置重复数≥15 次作为 SSR 位点的识别条件,因此在表中未分析单核苷酸类型。除单核苷酸外,各重复类型重复数在 5~11 次之间,随重复次数的逐渐增加,频率逐步降低。除单核苷酸外,5~7 次是主要集中次数,占 SSR 类型总数的大多数。

表 4 苦参转录组 SSR 各重复类型在不同重复次数的数量分布

重复次数 (次)	重复类型数量(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
5	0	0	1 236	77	39	55
6	0	756	512	31	3	8
7	0	431	230	8	0	0
8	0	255	178	3	0	0
9	0	173	27	1	0	0
10	1 777	82	14	0	0	0
11	924	64	8	0	0	0
≥12	1 745	202	12	0	0	0
总计	4 446	1 963	2 217	120	42	63

2.6 含 SSR 序列的基因功能注释及生物碱基因挖掘

为了解含有 SSR 序列苦参转录组序列的基因功能,本研究通过与公共蛋白数据库进行比对,得到含有 SSR 序列的分类信息和功能注释。结果发现,uniprot_sprot、Pfam、eggno

Kegg、GO 分别注释到 3 094、3 162、3 061、3 138、3 467 个基因。

GO 注释将基因功能分为生物进程(biological process)、细胞组分(cellular component)、功能组分(molecular function) 大类,其下又分了很多子类,从不同角度对基因的功能进行分类注释,各类间互相关联。GO 注释可以全面描述苦参中 SSR 基因和基因产物的属性。将搜索到含有 SSR 的 unigene 序列使用 blastx 比对到蛋白数据库,取比对分值最高的为序列注释信息。细胞组分注释 10312 条,生物进程注释 11 200 条,功能组分注释 4 376 条。将含有 SSR 序列的 3 467 条 unigene 编号后与其对应的 GO 分类号一起导入到 GO 分类图形显示在线分析工具 WEGO 软件中,得到其基因功能分布(图 2)。结果表明,在 3 467 条 unigene 序列中注释信息获得 23 483 个功能注释,平均 1 条 unigene 有 6.77 个 GO 注释。

苦参主要药用成分是苦参碱和黄酮类物质,通过对含有 SSR 位点的序列进行 GO 注释数据挖掘,获得 7 个生物碱代谢途径相关基因,2 个黄酮类生物合成过程相关基因。

3 讨论

苦参转录组 de novo 组装获得 51 606 个长度大于 200 bp 的 contigs,使用 uniprot 和 Pfam 蛋白质数据库进行 ORF 比对校正,uniprot 比对上 15 242 条蛋白质序列,Pfam 数据库校比对上 126 429 条蛋白质序列,TransDecoder 最终筛选得到 23 074 条 ORF 信息,unigenes 序列长度在 700 bp 以下的序列

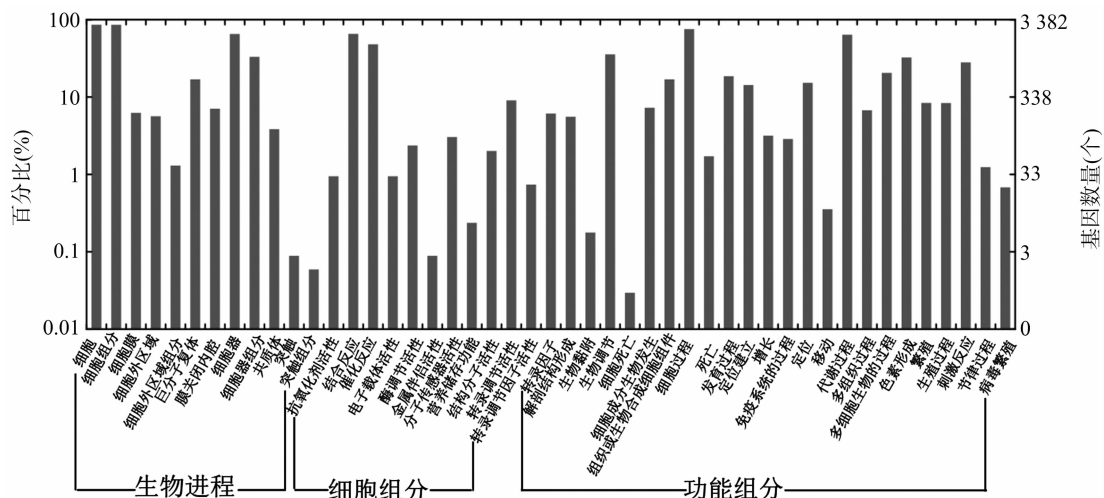


图2 GO 注释分布

数大约占总序列数的 70%。BUSCO 对转录组组装结果: C 占比为 69.5%, S 占比为 67.6%, D 占比为 1.9%, F 占比为 11.9%, M 占比为 18.6%, 总 BUSCOs 数目为 1 440 条。

苦参转录组序列通过 MISA 搜索到 8 798 个 SSR 位点, SSR 位点的 unigenes 序列在苦参转录组序列中 SSR 位点占比为 2.16%, 平均分布距离 4 411 bp 出现 1 个 SSR。与其他药用植物比较, 高于党参的 0.022%^[13], 低于丹参的 0.047%^[14], 高于西洋参的 0.013 3%^[15] 和人参的 0.017 2%^[16]; 与豆科模式植物大豆相比, 高于大豆的 0.013 5%^[17]。表明苦参的 SSR 位点数量较为丰富。通过对含有 SSR 位点序列的注释进一步分析获得苦参生物碱相关代谢基因, 为后续相关研究提供参考。

本研究结果为苦参转录组数据中的 SSR 位点分析提供依据。本研究对转录组序列进行了 ORF 预测, 反映了基因组中基因的编码区域, 可进一步确定基因位置, 省去了 SSR 引物设计开发过程中的克隆和测序步骤, 充分利用了生物信息数据库现有测序数据, 降低了开发成本。同时也明确了苦参 SSR 位点的基本特点, 为进一步开发设计新的苦参功能基因 SSR 标记奠定了基础。苦参中 SSR 对于苦参基因功能资源的开发利用、遗传资源评估、丰富的分子标记、种质资源改良和比较基因组学研究都具有重要的价值。

参考文献:

- [1] 国家药典委员会. 中华人民共和国药典[M]. 北京: 化学工业出版社, 2015.
- [2] 张贵君. 精编中草药彩色图谱[M]. 北京: 中国医药科技出版社, 2016.
- [3] 张 翊. 苦参茎叶中化学成分的研究[D]. 天津: 天津中医药大学, 2013.
- [4] 段永红, 渠云芳, 王长彪, 等. 药用植物苦参 SSR-PCR 体系的优化与验证[J]. 中国农业大学学报, 2014, 19(5): 95-100.
- [5] He J Y, Zhu S, Komatsu K, et al. Genetic polymorphism of medicinally-used *Codonopsis* species in an internal transcribed spacer sequence of nuclear ribosomal DNA and its application to authenticate *Codonopsis Radix* [J]. Journal of Natural Medicines,

2014, 68(1): 112-124.

- [6] Han R, Takahashi H, Nakamura M, et al. Transcriptome analysis of nine tissues to discover genes involved in the biosynthesis of active ingredients in *Sophora flavescens* [J]. Biological and Pharmaceutical Bulletin, 2015, 38(6): 876-883.
- [7] Staff S. Using the SRA Toolkit to convert .sra files into other formats [EB/OL]. (2015-08-22) [2017-12-06]. <http://www.ncbi.nlm.nih.gov/books/NBK158900/>.
- [8] Bonnal R J P, Ranzani V, Arrigoni A, et al. De novo transcriptome profiling of highly purified human lymphocytes primary cells [J]. Scientific Data, 2015, 2: 150051.
- [9] 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, 2011, 29(7): 644-652.
- [10] 王 林. 白色链霉菌和白背飞虱的基因组学研究[D]. 合肥: 中国科学技术大学, 2017.
- [11] 舒江平, 刘 莉, 沈 慧, 等. 基于系统基因组学分析揭示早期陆生植物的复杂网状进化关系[J]. 生物多样性, 2017, 25(6): 675-682.
- [12] 王 希, 陈 丽, 赵春雷. 利用 MISA 工具对不同类型序列进行 SSR 标记位点挖掘的探讨[J]. 中国农学通报, 2016, 32(10): 150-156.
- [13] 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, 45(16): 2390-2394.
- [14] 邓科君, 张 勇, 熊丙全, 等. 药用植物丹参 EST-SSR 标记的鉴定[J]. 药学报, 2009, 44(10): 1165-1172.
- [15] 杨维泽, 金 航, 赵振玲, 等. 西洋参 EST 资源的 SSR 信息分析[J]. 西南农业学报, 2011, 24(1): 275-278.
- [16] Li C F, Zhu Y J, Guo X, et al. Transcriptome analysis reveals ginsenosides biosynthetic genes, microRNAs and simple sequence repeats in *Panax ginseng* C. A. Meyer [J]. BMC Genomics, 2013, 14: 245.
- [17] Dreisigacker S, Zhang P, Warburton M L, et al. SSR and pedigree analyses of genetic diversity among CIMMYT wheat lines targeted to different megaenvironments [J]. Crop Science, 2004, 44(2): 381-388.