# 褐飞虱 EST 资源的微卫星信息分析

刘玉娣[1,2], 侯茂林[1,*]

(1. 中国农业科学院植物保护研究所, 植物病虫害生物学国家重点实验室, 北京 100193;
2. 广东省野生动物保护与利用公共实验室, 广州 510260)

**摘要**: 表达序列标签(expressed sequence tags, ESTs)是开发微卫星标记的一个重要的资源。褐飞虱 *Nilaparvata lugens* (Stål) EST 序列的公布为开发 EST-SSRs 提供了宝贵的数据资源, 本研究利用生物信息学对 NCBI 公共数据库中的 37 398 条褐飞虱 ESTs 序列进行 EST-SSRs 特征分析, 得到全长为 7 619.324 kb 的无冗余 EST 9 852 条。按照 3 个不同的查找标准在这些序列中搜索 SSR。查找结果显示: 褐飞虱 EST-SSRs 主要重复基元以 1 ~ 3 碱基为主, 占总 EST-SSR 的 95% 以上。在单碱基重复基元中, A/T 是占优势的重复基元, 在二相重复类型中, AG/CT 重复基元出现的频率最多, 而 AAG/CTT 是三相重复中占绝对优势的重复基元。在褐飞虱 EST-SSRs 中未查找到 GC 重复基元。以 100 bp 为参照, 在 3 种查找标准下含有 SSR 的 EST 序列中两端侧翼序列均≥100 bp 的序列分别为 738, 89 和 42 个。通过分析褐飞虱 EST-SSRs 标记可以为褐飞虱和近缘种的 SSR 标记的开发提供信息, 同时通过分析褐飞虱 EST-SSRs 的分布频率和分布特征可以为昆虫 EST-SSRs 的研究提供借鉴和参考。

**关键词**: 褐飞虱; 生物信息学; 表达序列标签; EST-SSRs; 查找标准; 重复基元

**中图分类号**: Q966 **文献标识码**: A **文章编号**: 0454-6296(2010)03-0239-09

## Analysis of microsatellite information in EST resource of *Nilaparvata lugens* (Homoptera: Delphacidae)

LIU Yu-Di[1,2], HOU Mao-Lin[1,*] (1. State Key Laboratory for Biology of Plant Diseases and Insect Pests, Institute of Plant Protection, Chinese Academy of Agricultural Sciences, Beijing 100193, China; 2. Guangdong Provincial Public Laboratory on Wild Animal Conservation and Management, Guangzhou 510260, China)

**Abstract**: Expressed sequence tags (ESTs) are important resources for development of new SSR markers. In this study, 37 398 ESTs of *Nilaparvata lugens* (Stål) were downloaded from NCBI and analyzed. After the pre-procession, 9 852 non-redundant ESTs with the total length about 7 619.324 kb were obtained. The EST-SSRs were detected under three search qualifications. The search results indicated that the 1 - 3 repeat motifs were the major repeats among all the SSRs, which accounted for above 95% of all EST-SSRs. A/T was the most frequent motif in the mononucleotide. AG/CT and AAG/CTT were the major motifs in the dinucleotide and trinucleotide, respectively. The GC repeat motif was not found in the EST-SSRs of *N. lugens*. When 100 bp was used as the comparison, the numbers of sequences with both flanking regions ≥100 bp under three search qualifications were 738, 89, and 42, respectively. The analysis of EST-SSRs markers can provide the information for the SSR development of *N. lugens* and related species. Furthermore, the analysis of the distribution frequency and character of *N. lugens* EST-SSRs can provide help for the EST-SSRs study of insects.

**Key words**: *Nilaparvata lugens*; bioinformatics; EST; EST-SSRs; search qualification; repeat motif

微卫星(microsatellite)即简单重复序列(simple sequence repeats, SSR), 又称短串联重复序列(short tandem repeats, STR), 是指以 1 ~ 6 个碱基为基本组成单元的串联重复序列(Chambers and MacAvoy, 2000)。根据组成单元包含的碱基数目的差异, 可分为二(di-)、三(tri-)、四(tetra-)、五(penta-)、六(hexa-)个核苷酸为基本重复单位(motif)的微卫星, 简称为二相、三相、四相、五

基金项目: 广东省野生动物保护与利用公共实验室基金资助项目(2008-003); 国家自然科学基金项目(30700537)

作者简介: 刘玉娣, 1977 年生, 博士, 副研究员, 研究方向昆虫分子生态学, E-mail: ydliu@ippcaas.cn

* 通讯作者 Corresponding author, E-mail: mlhou@ippcaas.cn

收稿日期 Received: 2009-10-26; 接受日期 Accepted: 2010-01-31

相、六相重复微卫星(Schlötterer, 1998; Chambers and MacAvoy, 2000)。与其他分子标记相比，微卫星在基因组中具有较高的丰度，突变率很高($10^{-2}$～$10^{-6}$/代)的优点。另外，由于筛选的微卫星位点长度一般较短，容易进行 PCR 扩增，对模板 DNA 的质量和数量要求不高，具有较高的可重复性，不受实验时间和地点的影响(Selkoe and Toonen, 2006)。因此微卫星分子标记是目前备受遗传学家青睐的分子标记之一。

传统开发 SSR 标记的方法是通过文库的构建、应用生物素探针杂交的方法将含有 SSR 克隆进行筛选、将阳性克隆测序并分析、引物设计、PCR 引物检测等步骤，因此基因组文库的构建和筛选步骤繁琐，十分费时费力，而且要求的技术平台比较高(Zane *et al.*, 2002)。在改进 SSR 位点的分离方法时，避免文库的构建和筛选成为了一个需求。由于不同物种 EST(expressed sequence tag，表达序列标签)计划的实施，目前公共数据库中的 EST 数量迅速增加。EST 资源库的不断扩充极大地方便和加快了人们在生命科学领域的研究，也为利用这些数据来开发 EST 分子标记奠定了基础。利用生物信息学的手段从庞大的 EST 数据中直接查找含有 SSR 的 EST，利用 EST 两端的保守序列设计 PCR 引物，它操作经济简便，缩短了 SSR 的开发周期，节省了开发成本。只要有 EST 测序的物种，在此物种中进行 EST-SSRs 的查找均是可行的。由于 EST-SSRs 标记的查找利用的是公共序列，省去了 SSR 引物开发过程中的克隆和测序步骤，因而 EST-SSRs 标记的开发过程简单，成本低。与传统的 SSR 相比，EST-SSRs 标记还有很多的优点，它在近缘种之间有较高的通用性，筛选到的 EST-SSRs 扩增稳定性好(Varshney *et al.*, 2005)。

目前在植物中已做了大量的 EST-SSRs 开发研究和应用，例如小麦、水稻、玉米、大豆、茶树、白菜、大麦、柑橘和棉花等(金基强等，2006；忻雅等，2006；王长彪等，2006；李建明等，2007；陈全求等，2008；陈相艳等，2009)，而在昆虫中的应用相对较少。褐飞虱 *Nilaparvata lugens* (Stål) EST 序列在公共数据库中公布，对开展褐飞虱 EST-SSRs 的研究意义较大。截止到 2009 年 9 月 30 日，共公布了 37 398 条 EST 序列。本研究对现有的褐飞虱 EST 中的 SSR 信息进行了全面分析，比较了不同查找条件下的褐飞虱 EST-SSRs 的发生频率和分布特点，并对 ESR-SSR 序列的可用性进行了统计分析。通过分析褐飞虱 EST-SSRs 标记一方面可以为褐飞虱和近缘种的 SSR 标记的开发提供信息，另一方面通过分析褐飞虱 EST-SSRs 的分布频率和分布特征可以为昆虫 EST-SSRs 的研究提供借鉴和参考。

## 1 材料和方法

### 1.1 褐飞虱 EST 来源

褐飞虱 EST 来自 NCBI(美国国家生物技术信息中心)数据库(http://www.ncbi.nlm.nih.gov/)，共计 37 398 条(Noda *et al.*, 2008)。

### 1.2 EST 前处理

采用 EST-trimmer.pl 软件(http://pgrc.ipk-gatersleben.de/misa/download/est-trimmer.pl)去除长度小于 100 bp 的 EST 序列。

### 1.3 聚类去冗余

前处理后的 ESTs 通过软件 Cap3Win(Huang and Madan, 1999)进行片段重叠群分析和聚类。拼接时设定的初始装配参数为：最小匹配碱基数(minmatch)为 30，最小分值(minscore)为 30。对错误拼接序列设置比较高的装配参数再次进行拼接，判别，共进行了 3 次。将分析后的重叠群(contigs)和单一序列(singlets)合并后，采用 EST-trimmer 软件去除 5′端或 3′端 50 bp 的 polyT 或 polyA 序列。

### 1.4 EST-SSR 筛选

应用 MISA.pl 软件(http://pgrc.ipk-gatersleben.de/misa/misa.html)对聚类和去除冗余序列后的 ESTs 进行 SSR 查找。分 3 个标准进行查找：查找标准 1 (1/10，2/6，3/5，4/5，5/5，6/5)，即单核苷酸重复的次数在 10 次或 10 次以上，二核苷酸重复的次数在 6 次或 6 次以上，三至六核苷酸重复的次数在 5 次或 5 次以上；查找标准 2 (1/15，2/10，3/8，4/8，5/8，6/8)，即单核苷酸重复的次数在 15 次或 15 次以上，二核苷酸重复的次数在 10 次或 10 次以上，三至六核苷酸重复的次数在 8 次或 8 次以上；查找标准 3 (1/20，2/12，3/10，4/10，5/10，6/10)，即单核苷酸重复的次数在 20 次或 20 次以上，二核苷酸重复的次数在 12 次或 12 次以上，三至六核苷酸重复的次数在 10 次或 10 次以上。对于复合微卫星(compound microsatellite)，查找标准为间隔(interrupted)等于 10 或小于 10 碱基的 2 个

SSR 为 1 个复合微卫星。

### 1.5　EST-SSR 可用性分析

SSR 的侧翼序列只有具有足够长度才能进行引物设计，从而进一步验证其可用性和多态性。本研究以 100 bp 为参照，当侧翼序列一端 < 100 bp 时该序列即认为是缺乏足够的侧翼序列长度，从而认为该序列是不可用的；当两端的侧翼序列均 ≥100 bp时，即有相对足够长的侧翼序列时，认为该序列是可用的。本研究通过设计程序，对含有 SSR 的 EST 序列进行统计计算，分别统计侧翼序列 < 100 bp和≥100 bp 的 SSR 的个数。

## 2　结果与分析

### 2.1　褐飞虱 EST 中出现 SSR 的频率

褐飞虱 37 398 条 EST 序列经过聚类拼接处理后共得到 9 852 条无冗余的 EST 序列，包括重叠群（contigs）3 897 个和单一序列（singlets）5 955 个。经过对处理后 9 852 条无冗余的 EST 序列按不同的查找标准进行搜索。

查找标准 1：共检出含有 SSR 的序列分别为 802 条，发生频率（含有 SSR 的 EST 数目与总 EST 数目的比值）为 8.14%。其中，690 条含单个 SSR，112 条含有 2 个或 2 个以上的 SSR。共检出 995 个 SSR，占无冗余 EST 的 10.09%。在这 995 个 SSR 中，完全重复 SSR 为 948 个，复合微卫星为 47 个。从分布情况看，褐飞虱 EST 中平均每 7.98 kb 就出现 1 个 SSR，但不同重复类型间差异很大（表 1）。

查找标准 2：共检出含有 SSR 的序列分别为 158 条，发生频率为 1.60%。其中，68 条含单个 SSR，106 条含有 2 个或 2 个以上的 SSR。共检出 174 个 SSR，占无冗余 EST 的 1.77%。在这 174 个 SSR 中，完全重复 SSR 为 165 个，复合微卫星为 9 个（表 1）。

查找标准 3：共检出含有 SSR 的序列分别为 87 条，发生频率为 0.88%。共检出 95 个 SSR，占无冗余 EST 的 0.96%。在这 95 个 SSR 中，完全重复 SSR 为 90 个，复合微卫星为 5 个。其中，31 条含单个 SSR，64 条含有 2 个或 2 个以上的 SSR（表 1）。

随着查找标准的上升，不同重复类型的 SSR 数量随之下降（表 1）。在 3 种查找方式下三核苷酸的比例均最高，其次是单核苷酸和二核苷酸。在查找标准为 2 和 3 时没有搜索到五核苷酸重复的 SSR（表 1）。

**表 1　褐飞虱 SSR 在无冗余 EST 中出现的频率**

**Table 1　Occurrence of SSRs in non-redundant *Nilaparvata lugens* ESTs**

| 重复基元 Repeat motif | SSR 数目 Number of SSRs | | | | | | 各类型所占比例（%） Proportion in all SSRs | | | | | | 平均距离（kb） Average distance | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SQ-1 | | SQ-2 | | SQ-3 | | SQ-1 | | SQ-2 | | SQ-3 | | SQ-1 | | SQ-2 | | SQ-3 | |
| | P | C | P | C | P | C | P | C | P | C | P | C | P | C | P | C | P | C |
| 单核苷酸 Mononucleotide | 359 | 7 | 68 | 0 | 31 | 0 | 37.59 | 0.73 | 39.08 | 0 | 32.63 | 0 | 21.22 | 1 088.47 | 112.05 | – | 245.78 | – |
| 二核苷酸 Dinucleotide | 165 | 14 | 21 | 1 | 16 | 1 | 17.28 | 1.47 | 12.07 | 0.57 | 16.84 | 1.05 | 46.18 | 544.24 | 362.82 | 7 619.32 | 476.21 | 7 619.32 |
| 三核苷酸 Trinucleotide | 357 | 23 | 69 | 7 | 40 | 4 | 37.38 | 2.41 | 39.66 | 4.02 | 42.11 | 14.74 | 21.34 | 331.27 | 110.42 | 1 088.47 | 190.48 | 1 904.83 |
| 四核苷酸 Tetranucleotide | 20 | 2 | 6 | 1 | 2 | 0 | 2.09 | 0.21 | 3.45 | 0.57 | 2.11 | 0 | 380.97 | 3 809.66 | 1 269.89 | 7 619.32 | 3 809.66 | – |
| 五核苷酸 Pentanucleotide | 5 | 0 | 0 | 0 | 0 | 0 | 0.52 | 0 | 0 | 0 | 0 | 0 | 1 523.86 | – | – | – | – | – |
| 六核苷酸 Hexanucleotide | 2 | 1 | 1 | 0 | 1 | 0 | 0.21 | 0.10 | 0.57 | 0 | 1.05 | 0 | 3 809.66 | 7 619.32 | 7 619.32 | – | 7 619.32 | – |
| | 908 | 47 | 165 | 9 | 90 | 5 | 95.08 | 4.92 | 94.83 | 5.17 | 94.74 | 5.26 | 8.39 | 162.11 | 46.18 | 846.59 | 84.66 | 1 523.86 |
| 总计 Total | 955 | | 174 | | 95 | | 100 | | 100 | | 100 | | | | | | | |

SQ-1：查找标准 1（Search qualification 1）；SQ-2：查找标准 2（Search qualification 2）；SQ-3：查找标准 3（Search qualification 3）. 下同 The same below. P：完全重复 SSR（Perfect SSR）；C：复合 SSR（Compound SSR）. –：未搜索到（No SSR can be searched）.

### 2.2 褐飞虱 EST-SSR 的特性

二相重复的微卫星(图 1)：3 种查找方式下 AG/CT 是出现频率最多的重复基元。在查找标准 1 条件下，AT/AT 和 AC/GT 二相重复基元出现的频率相近，而在查找标准 2 和 3 条件下，AC/GT 重复基元出现的频率高于 AT/AT 重复基元。

三相重复的微卫星(图 2)：3 种查找方式下 AAG/CTT 是出现频率最高的重复基元。重复基元 AAT/ATT 在 3 种查找方式下出现的频率相近。CCG/CGG 为查找方式 1 中出现频率最低的基元，仅为 1.85%。在查找标准 2 和 3 条件下，均缺失重复基元 ACC/GGT 和 CCG/CGG)。

褐飞虱的 EST-SSRs 种类十分丰富，一至六核苷酸重复类型都能看到，但各种类型出现的频率相差很大，主要集中在一至三核苷酸重复上(表 2)。在查找标准 1 条件下(表 2)：共观察到 30 种重复基元，单核苷酸有 2 种，二核苷酸重复基元有 3 种，三、四、五、六核苷酸重复基元分别有 10，8，

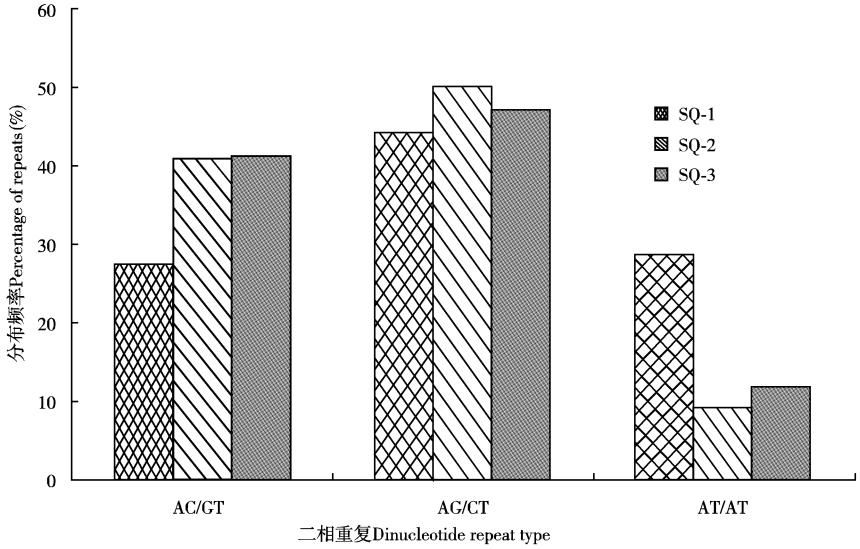


图 1 不同查找标准下褐飞虱 UniGene 中二核苷酸 SSR 分布频率

Fig. 1 Distribution frequency of the dinucleotide repeat type under different search qualifications in UniGene sequences of *Nilaparvata lugens*

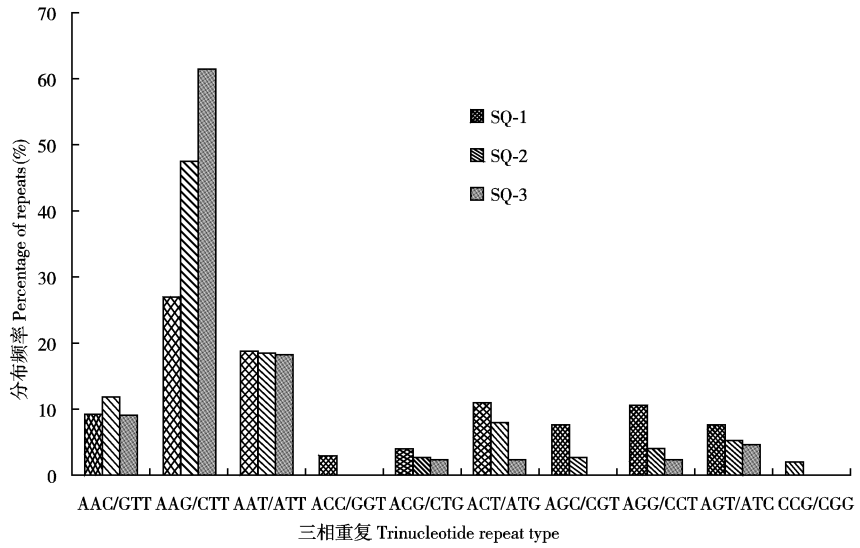


图 2 不同查找标准下褐飞虱 UniGene 中 3 核苷酸 SSR 分布频率

Fig. 2 Distribution frequency of the trinucleotide repeat type under different search qualifications in UniGene sequences of *Nilaparvata lugens*

5 和 2(表 2)。一至三核苷酸重复占总 EST-SSR 的 96.86%，其中，三核苷酸重复最为常见，占总 SSR 的 39.79%，二核苷酸重复占总 SSR 的 18.74%，而四至六核苷酸重复所占比例较小，仅占总 SSR 的 3.14%；在查找标准 2 和 3 条件下，各重复基元的数量都急剧下降，其中五相重复基元消失。在查找条件 2 下(表 2)：共观察到 18 种重复基元，单核苷酸有 1 种，二核苷酸重复基元有 3 种，三、四、五、六核苷酸重复基元分别有 8，5，0 和 1。一至三核苷酸重复占总 EST-SSR 的 95.40%，其中，3 三核苷酸重复最为常见，占总 SSR 的 43.68%；在查找标准 3 条件下(表 2)：共观察到 14 种重复基元，单核苷酸有 1 种，二核苷酸重复基元有 3 种，三、四、五、六核苷酸重复基元分别有 7，2，0 和 1。一至三核苷酸重复占总 EST-SSR 的 96.84%，三核苷酸重复也是最为常见，占总 SSR 的 46.32%。

**表 2　9 852 条褐飞虱无冗余序列中不同微卫星的出现情况**

**Table 2　Occurrence of different SSRs in 9 852 non-reduntant ESTs of *Nilaparvata lugens***

| 重复基元 Repeat motif | 重复次数 Number of repeats | | | | | | | | | | | | 合计 Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | ≥16 | |
| A/T | – | – | – | – | – | 148 | 54 | 34 | 17 | 8 | 9 | 59 | 329 |
| | | | | – | – | – | – | – | – | – | 9 | 59 | 68 |
| | | | | | | – | – | – | – | – | – | **31** | **31** |
| C/G | – | – | – | – | – | 15 | 8 | 8 | 5 | 1 | 0 | 0 | 37 |
| AC/GT | – | 27 | 5 | 5 | 3 | 2 | 0 | 0 | 2 | 0 | 0 | 5 | 49 |
| | | | | – | – | *2* | *0* | *0* | *2* | *0* | *0* | *5* | *9* |
| | | | | | | – | – | **0** | **2** | **0** | **0** | **5** | **7** |
| AG/CT | – | 36 | 20 | 4 | 8 | 2 | 1 | 2 | 0 | 3 | 1 | 2 | 79 |
| | | | | – | – | *2* | *1* | *2* | *0* | *3* | *1* | *2* | *11* |
| | | | | | | **–** | **–** | **2** | **0** | **3** | **1** | **2** | **8** |
| AT/AT | – | 25 | 9 | 13 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 51 |
| | | | | – | – | *0* | *0* | *0* | *1* | *0* | *0* | *1* | *2* |
| | | | | | | – | – | **0** | **1** | **0** | **0** | **1** | **2** |
| AAC/GTT | 11 | 9 | 6 | 3 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 2 | 35 |
| | | | | *3* | *2* | *0* | *2* | *0* | *0* | *0* | *0* | *2* | *9* |
| | | | | | | **0** | **2** | **0** | **0** | **0** | **0** | **2** | **4** |
| AAG/CTT | 39 | 17 | 10 | 6 | 3 | 4 | 2 | 1 | 2 | 0 | 5 | 13 | 102 |
| | | | | *6* | *3* | *4* | *2* | *1* | *2* | *0* | *5* | *13* | *36* |
| | | | | | | **4** | **2** | **1** | **2** | **0** | **5** | **13** | **27** |
| AAT/ATT | 42 | 9 | 6 | 6 | 0 | 0 | 2 | 2 | 0 | 1 | 2 | 1 | 71 |
| | | | | *6* | *0* | *0* | *2* | *2* | *0* | *1* | *2* | *1* | *14* |
| | | | | | | **0** | **2** | **2** | **0** | **1** | **2** | **1** | **8** |
| ACC/GGT | 7 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 |
| ACG/CTG | 3 | 6 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 15 |
| | | | | *1* | *0* | *0* | *0* | *0* | *0* | *1* | *0* | *0* | *2* |
| | | | | | | **0** | **0** | **0** | **0** | **1** | **0** | **0** | **1** |
| ACT/ATG | 24 | 6 | 6 | 4 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 42 |
| | | | | *4* | *1* | *0* | *0* | *0* | *1* | *0* | *0* | *0* | *6* |
| | | | | | | **0** | **0** | **0** | **1** | **0** | **0** | **0** | **1** |

**续表 2 Table 2 continued**

| 重复基元 Repeat motif | 重复次数 Number of repeats | | | | | | | | | | | | 合计 Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | ≥16 | |
| AGC/CGT | 13 | 6 | 8 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 29 |
| | | | | *2* | *0* | *0* | *0* | *0* | *0* | *0* | *0* | *0* | *2* |
| AGG/CCT | 24 | 9 | 4 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 40 |
| | | | | *2* | *0* | *0* | *1* | *0* | *0* | *0* | *0* | *0* | *3* |
| | | | | | | **0** | **1** | **0** | **0** | **0** | **0** | **0** | **1** |
| AGT/ATC | 12 | 9 | 3 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 28 |
| | | | | *2* | *0* | *0* | *1* | *0* | *0* | *0* | *0* | *1* | *4* |
| | | | | | | **0** | **1** | **0** | **0** | **0** | **0** | **1** | **2** |
| CCG/CGG | 6 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 |
| AAAG/CTTT | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| AAAT/ATTT | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| | | | | *1* | *0* | *0* | *0* | *0* | *0* | *0* | *0* | *0* | *1* |
| AACT/ATTG | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| AAGC/CGTT | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| | | | | *0* | *0* | *1* | *0* | *0* | *0* | *0* | *0* | *0* | *1* |
| | | | | | | **1** | **0** | **0** | **0** | **0** | **0** | **0** | **1** |
| AATC/AGTT | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| | | | | *0* | *0* | *1* | *0* | *0* | *0* | *0* | *0* | *0* | *1* |
| | | | | | | **1** | **0** | **0** | **0** | **0** | **0** | **0** | **1** |
| AATG/ACTT | 1 | 1 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |
| | | | | *2* | *1* | *0* | *0* | *0* | *0* | *0* | *0* | *0* | *3* |
| ACTC/AGTG | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| AGAT/ATCT | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| | | | | *1* | *0* | *0* | *0* | *0* | *0* | *0* | *0* | *0* | *1* |
| AAAGT/ATTTC | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| AAATC/AGTTT | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| AATAG/ATCTT | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| AATGG/ACCTT | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| ACGTC/AGTGC | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| AAAAAG/CTTTTT | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 |
| | | | | *0* | *0* | *0* | *0* | *1* | *0* | *0* | *0* | *0* | *1* |
| | | | | | | **0** | **0** | **1** | **0** | **0** | **0** | **0** | **1** |
| AAGGTC/AGTTCC | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

查找标准 1 对应的是不同重复基元后同一行的数字；查找标准 2 对应斜体数字；查找标准 3 对应粗体数字。The numbers after the different repeat motifs on the same line indicate the search results using SQ-1, the italic numbers indicate the search results using SQ-2, while the bold numbers indicate the search results using SQ-3.

### 2.3　EST-SSR 可用性分析

本研究以 100 bp 为参照，对含有 SSR 的 EST 序列进行统计计算。分别统计单侧翼序列 < 100 bp，两侧翼序列均 < 100 bp，两侧翼序列均≥100 bp 前提下在 3 种查找标准条件下的 EST-SSRs 序列个数。统计结果显示：单侧序列侧翼列 < 100 bp 的序列在查找标准 1 下达到了 355 个，查找标准 2 和 3 下分别为 86 和 53 个。随着查找标准的提高，单侧序列侧翼列 < 100 bp 的序列所占的比例上升（表 3）。两侧翼序列均 < 100 bp 的序列在 3 种查找标准下分别为 43，7 和 4。两端侧翼序列均≥100 bp 的序列统计结果显示在 3 种查找标准下的分别为 738，89 和 42 个（表 3）；随着查找标准的提高所占的比例下降，在查找标准 1 条件下，所占得比例为 77.28%，在查找标准 2 和 3 条件下所占的比例分别为 51.15% 和 44.21%（表 3）。

**表 3　不同查找标准下不同侧翼序列长度的褐飞虱 EST-SSR 序列统计**

**Table 3　Statistics of EST-SSR sequences with different lengths of flanking regions under different search qualifications**

| 微卫星重复基元 Microsatellite repeat motif | 单侧翼序列 < 100 bp Single flanking region < 100 bp | | | 两侧翼序列均 < 100 bp Both flanking regions < 100 bp | | | 两侧翼序列均≥100 bp Both flanking regions≥100 bp | | |
|---|---|---|---|---|---|---|---|---|---|
| | SQ-1 | SQ-2 | SQ-3 | SQ-1 | SQ-2 | SQ-3 | SQ-1 | SQ-2 | SQ-3 |
| 完全重复单相微卫星 Mononucleotide microsatellite | 146 | 36 | 21 | 15 | 2 | 2 | 259 | 35 | 12 |
| 完全重复二相微卫星 Dinucleotide microsatellite | 66 | 8 | 8 | 7 | 1 | 1 | 137 | 13 | 8 |
| 完全重复三相微卫星 Trinucleotide microsatellite | 120 | 34 | 18 | 20 | 3 | 0 | 291 | 33 | 19 |
| 完全重复四相微卫星 Tetranucleotide microsatellite | 5 | 4 | 3 | 1 | 1 | 1 | 16 | 2 | 0 |
| 完全重复五相微卫星 Pentanucleotide microsatellite | 1 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 |
| 完全重复六相微卫星 Hexanucleotide microsatellite | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 |
| 复合微卫星 Compound microsatellite | 17 | 3 | 2 | 0 | 0 | 0 | 30 | 6 | 3 |
| 总计 Total | 355 | 86 | 53 | 43 | 7 | 4 | 738 | 89 | 42 |
| 各类型所占比例 Proportion in all SSRs(%) | 37.17 | 49.43 | 55.79 | 4.50 | 4.02 | 4.21 | 77.28 | 51.15 | 44.21 |

## 3　讨论

本研究中通过对褐飞虱 37 398 条 EST 序列经过聚类拼接处理后共得到 9 852 条无冗余的 EST 序列，对 7 619.324 kb 的序列按不同的查找标准进行了 SSR 的查找。不同的查找条件下，褐飞虱 EST-SSRs 中三相重复的微卫星出现的频率最高，这与许多的禾本科植物中的 EST-SSRs（Varshney，2002）以及曼氏血吸虫的 EST-SSRs 分布相似（唐远菊等，2007），而与赤拟谷盗（张琳琳等，2008）和蜜蜂（李斌等，2004）的 EST-SSRs 分布不同。张琳琳等（2008）的研究发现赤拟谷盗 EST 中单碱基重复序列占主导地位，其次是六碱基重复序列。李斌等（2004）的研究发现蜜蜂 EST 中微卫星六碱基重复序列占主导地位，其次是二碱基重复序列。因此微卫星类型在不同物种间分布存在差异。

褐飞虱 EST-SSRs 主要重复基元以 1 ~ 3 碱基为主，占总 EST-SSR 的 95% 以上。在单碱基重复基元中，A/T 是占优势的重复基元，这与花生、大豆和油菜中的单重复基元以 A/T 为主相似（李小白等，2007；柳展基等，2008；陈相艳等，2009）。在二相重复类型中，AG/CT 重复基元出现的频率最多，与报道的油菜（占二相重复的 84.04%）、花生（占总 SSR 的 23.44%）和大豆（占总 SSR 的 23.46%）相似。但与曼氏血吸虫（二相重复以 AC/TG 为主，占二相重复的 49%；唐远菊等，2007）和赤拟谷盗中 $(CA)_n$ 是二相重复中出现频率最高的重复基元（张琳琳等，2008）的二相重复主要基元不同。三相重复中，AAG/CTT 重复基元占绝对优势，与报道的油菜（占三相重复的 35.71%）、花生（占总 SSR 的 15.51%）和大豆相似（占总 SSR 的 11.03%）。但与曼氏血吸虫（三相重复以 AAT/TTA 为主，占三相重复的 31%；唐远菊等，2007）和赤拟谷盗〔$(CTA)_n$ 是三相重复中出现频率最高的重复基元；张琳琳等，2008〕的三相主要重复基元不同。在褐飞虱 EST-SSRs 中未查找到 GC 重复基元，GC 重复基元在多数植物中也很难见到（Gao *et*

*al.*, 2003)，同时在曼氏血吸虫(唐远菊等，2007)和赤拟谷盗中(张琳琳等，2008)GC 重复基元仅以非常低的频率出现。

通过采用不同的查找标准对褐飞虱 EST-SSRs 进行查找，研究发现随着查找标准的提高，即将不同类型的重复基元的重复数提高，不同重复类型的 SSR 随着查找标准的上升数量急剧下降。同时褐飞虱 EST-SSRs 在不同的查找标准下的分布频率也不相同。由于微卫星的定义不同，也就是设定的查找标准不一致，不同物种中 EST-SSRs 出现的频率和分布特征出现大的差异。La Rota 等(2005)通过设定不同的 SSR 长度，研究水稻、大麦和黑麦 EST-SSRs 的分布频率及特点的变化情况，当把水稻最小 SSR 长度标准由 12 bp 增加到 30 bp 时，EST-SSRs 的频率从 50% 减少到 1%，同时二核苷酸重复的数量变得与三核苷酸重复基本接近，重复基元的主导类型也由 CCG 变为 AG 重复。因此，在进行物种间特定 EST-SSRs 频率和分布特征比较时，只有在相同或相似的查找参数条件下得到的结果才具有可比性。

含有 SSR 的序列是否具备可用性的一个首要的前提是该序列要具备足够长的侧翼序列，从而才能对其进行引物设计和下一步的检验和验证。本研究首次通过对含有 SSR 的 EST 序列进行侧翼序列的长度统计分析，从而验证其可用性。统计结果分析表明两端侧翼序列均≥100 bp 的序列随着查找标准的提高所占的比例下降，最高的比例仅为 77.28%。因此在进行 ESR-SSR PCR 引物设计时，首先要统计分析出两侧翼序列均具有足够长度的序列，然后对这些序列进行下一步的引物设计等实验，从而避免时间和精力上的浪费。对于从 EST 中查找 SSR 的研究工作来讲，在进行 SSR 引物设计之前，选出具有足够侧翼长度的序列是必须和必要的。本研究的统计分析能为其他物种的相关研究提供借鉴和参考。

尽管 EST-SSRs 有很多的优越性，但它自身也存在着弊端。由于 EST 是长约 150～500 bp 的基因表达序列片段，因此与传统的 SSR 标记相比较，EST-SSRs 标记的多态性比来自于基因组的 SSR 的多态性低。这也是 EST-SSRs 的在实际应用中的一大缺陷。Eujayl 等(2002)研究发现在小麦中，EST-SSRs 的多态性低于基因组 SSRs，仅为 25%，在大麦中，来自于 3′-UTR 的 EST-SSR 的多态性比自于 5′-UTR 的 EST-SSRs 的多态性高。正是由于 EST-SSRs 的多态性低，因此需要对大量的位点进行引物设计并验证各自的多态性以便找到足够的多态位点。微卫星以逐步突变模型(stepwise mutation model，SMM)为主进行突变(Ohta and Kimura，1973)，因此只有重复基元的重复数足够多，找到足够数量的多态位点才更有可能性。本研究通过列出不同查找方式下 EST-SSRs 的分布特征和频率，目的是比较不同的查找标准下得到的 EST-SSRs 数量差异。因此在实际应用中，应根据研究工作的实际需要，按照严格的条件对 EST-SSRs 进行筛选，从而避免过多精力和财力的浪费。

总之，尽管 EST-SSRs 存在一些不足，但随着 EST 计划开展和 EST 数据资源将不断丰富，EST-SSRs 标记的建立对于加速物种资源的开发利用、遗传资源评价、物种间比较作图、绘制遗传图谱、阐明物种起源和人工选择的历史过程等研究都具有重要的意义。

## 参 考 文 献 (References)


Chambers GK, MacAvoy ES, 2000. Microsatellites: Consensus and controversy. *Comparative Biochemistry and Physiology Part B*, 126: 455 - 476.

Chen QQ, Zhan XJ, Lan JY, Huang Y, 2008. Study progress in development of EST (expressed sequence tags). *Chinese Agricultural Science Bulletin*, 24(9): 72 - 77. [陈全求, 詹先进, 蓝家样, 黄云, 2008. EST 分子标记开发研究进展. 中国农学通报, 24(9): 72 - 77]

Chen XY, Li W, Dai HY, Zhang LF, 2009. Analysis of SSR information in EST resource of soybean (*Glycine max*). *Soybean Science*, 28(3): 394 - 399. [陈相艳, 李伟, 戴海英, 张礼凤, 2009. 大豆 EST 资源的 SSR 信息分析. 大豆科学, 28(3): 394 - 399]

Eujayl I, Sorrells ME, Baum M, Wolters P, Powell W, 2002. Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. *Theoretical and Applied Genetics*, 104: 399 - 407.

Gao LF, Tang JF, Li HW, Jia JZ, 2003. Analysis of microsatellites in major crops assessed by computational and experimental approaches. *Molecular Breeding*, 12: 245 - 261.

Huang X, Madan A, 1999. CAP3: a DNA sequence assembly program. *Genome Research*, 9: 868 - 877.

Jin JQ, Cui HR, Chen WY, Lu MZ, Yao YL, Xin Y, Gong XC, 2006. Data mining for SSRs in ESTs and development of EST-SSR marker in tea plant (*Camellia sinensis*). [金基强, 崔海瑞, 陈文岳, 卢美贞, 姚艳玲, 忻雅, 龚晓春, 2006. 茶树 EST-SSR 的信息分析与标记建立. 茶叶科学, 26: 17 - 23]

La Rota M, Kantety RV, Yu JK, Sorrells ME, 2005. Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. *BMC Genomics*, 6: 23.

Li B, Xia QY, Lu C, Zhou ZY, 2004. Analysis of microsatellites derived from bee ESTs. *Acta Genetica Sinica*, 31(10): 1 089 -

1 094. ［李斌，夏庆友，鲁成，周泽扬，2004. 蜜蜂 EST 中的微卫星分析. 遗传学报，31(10)：1 089－1 094］

Li JM，Li HJ，Chai SC，Li XQ，Li LH，2007. Studies on genetic diversity of genomic-SSR and EST-SSR molecular marker in common wheat. *Journal of Anhui Agricultural Science*，35(26)：8 173－8 175. ［李建明，李洪杰，柴守诚，李秀全，李立会，2007. 普通小麦 Genomic-SSR 和 EST-SSR 分子标记遗传差异. 安徽农业科学，35(26)：8 173－8 175］

Li XB，Zhang ML，Cui HR，2007. Analysis of SSR information in EST resource of oilseed rape. *Chinese Journal of Oil Crop Sciences*，29(1)：20－25. ［李小白，张明龙，崔海瑞，2007. 油菜 EST 资源的 SSR 信息分析. 中国油料作物学报，29(1)：20－25］

Liu ZJ，Sun P，Bu X，2008. Analysis of SSR information in EST resource of peanut. *Journal of Peanut Science*，37(4)：6－11. ［柳展基，孙萍，步迅，2008. 花生 EST 资源的 SSR 信息分析. 花生学报，37(4)：6－11］

Noda H，Kawai S，Koizumi Y，Matsui K，Zhang Q，Furukawa S，Shimomura M，Mita K，2008. Annotated ESTs from various tissues of the brown planthopper *Nilaparvata lugens*：a genomic resource for studying agricultural pests. *BMC Genomics*，9：117.

Ohta T，Kimura M，1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research*，22：201－204.

Schlötterer C，1998. Microsatellite. In：Hoelzel AR ed. Molecular Genetic Analysis of Populations：A Practical Approach. IRL Press，Oxford. 237－261.

Selkoe KA，Toonen RJ，2006. Microsatellites for ecologists：a practical guide to using and evaluating microsatellite markers. *Ecology Letters*，9：615－629.

Tang YJ，Luo HL，Nie K，2007. Analysis of microsatellites from *Schistosoma mansoni* ESTs. *Chinese Journal of Preventive Veterinary Medicine*，29(8)：629－633. ［唐远菊，罗洪林，聂奎，2007. 曼氏血吸虫 EST 中的微卫星分析. 中国预防兽医学报，29(8)：629－633］

Varshney RK，Graner A，Sorrells ME，2005. Genic microsatellite markers in plants：Features and applications. *Trends in Biotechnology*，23：48－55.

Varshney RK，Thiel T，Stein N，Langridge P，Graner A，2002. *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cell Mol. Biol. Lett.*，7(2A)：537－546.

Wang CB，Guo WZ，Cai CP，Zhang TZ，2006. Characterization，development and exploitation of EST-derived microsatellites in *Gossypium raimondii* Ulbrich. *Chinese Science Bulletin*，51(3)：316－320. ［王长彪，郭旺珍，蔡彩平，张天真，2006. 雷蒙德氏棉 EST-SSRs 分布特征及开发与利用. 科学通报，51(3)：316－320］

Xin Y，Cui HR，Lu MZ，Yao YL，Jin JQ，Lim YP，Choi SY，2006. Data mining for SSRs in ESTs and EST-SSR marker development in Chinese cabbage. *Acta Horticuhurae Sinica*，33(3)：549－554. ［忻雅，崔海瑞，卢美贞，姚艳玲，金基强，林容杓，崔水莲，2006. 白菜 EST-SSR 信息分析与标记的建立. 园艺学报，33(3)：549－554］

Zane L，Bargelloni L，Patarnello T，2002. Strategies for microsatellite isolation：a review. *Molecular Ecology*，11(1)：1－16.

Zhang LL，Wei CM，Lian ZM，Kong GY，2008. Abundance of microsatellites in the entire genome and EST of *Tribolium castaneum*. *Chinese Bulletin of Entomology*，45(1)：38－42. ［张琳琳，魏朝明，廉振民，孔光耀，2008. 赤拟谷盗全基因组和 EST 中微卫星的丰度. 昆虫知识，45(1)：38－42］

（责任编辑：袁德成）