

基于 EST 数据库进行 SNP 分子标记开发的研究进展及在猕猴桃属植物中的应用研究

周锦^{1,2}, 刘义飞³, 黄宏文^{3*}

(1. 中国科学院武汉植物园, 武汉 430074; 2. 中国科学院研究生院, 北京 100049;
3. 中国科学院华南植物园 中国科学院植物资源保护与可持续利用重点实验室, 广州 510650)

摘要: 对基于 EST 数据库开发 SNP 标记的特点、开发策略等进行了综述, 并介绍了在中华猕猴桃复合体(*Actinidia chinensis* Planch.)中开发 EST-SNP 的基本思路和初步结果, 为后续分子实验验证及其在自然居群中的应用奠定基础, 并为其它相关物种的 EST-SNP 分子标记开发提供借鉴。

关键词: 生物信息学; 猕猴桃; 遗传多样性; 表达序列标签; 单核苷酸多态性

中图分类号: Q78

文献标识码: A

文章编号: 1005-3395(2011)02-0184-11

doi: 10.3969/j.issn.1005-3395.2011.02.014

Progress on Development of EST Derived SNP Markers and Its Applications in *Actinidia chinensis* Species Complex

ZHOU Jin^{1,2}, LIU Yi-fei³, HUANG Hong-wen^{3*}

(1. Wuhan Botanical Garden, Chinese Academy of Sciences, Wuhan 430074, China; 2. Graduate University of Chinese Academy of Sciences, Beijing 100049, China; 3. Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China)

Abstract: The advances in characteristics of EST derived SNPs in plants and development strategies were reviewed, and the preliminary results of developing EST-SNPs from *Actinidia chinensis* complex were introduced. The potential applications of EST-SNPs in kiwifruit population genetic studies were further discussed, and the development of EST-SNPs in other species was prospected.

Key words: Bioinformatics; *Actinidia chinensis*; Genetic diversity; EST; SNP

生物物种外在的形态变异和多样化与其内在的基因组遗传变异息息相关。基因组 DNA 序列中一个核苷酸的突变便可引起相关基因或遗传位点功能的改变, 并最终导致不同的形态和生态适应性。传统的基因组随机 DNA 分子标记(Random DNA markers, RDMs)技术代表了基因组学分析的早期阶段, 主要用于生物遗传多样性和变异的研究以及相关遗传和物理图谱的构建等^[1]。随着人类基因组计划(Human Genome Project, HGP)及其它模

式生物基因组计划的实施, 大量基因组遗传信息的积累为新一代分子标记的开发应用提供了新的平台。利用相关基因序列信息发展起来的新型功能分子标记(Functional Markers, FMs)和目标基因分子标记(Gene Targeted Markers, GTMs)与特定基因位点的功能表达密切相关, 被广泛应用于生物物种的起源、进化以及生态适应性和多样化等研究中^[3]。作为研究结构基因组和功能基因组学的桥梁和工具, 不同生物物种的各种表达序列标签

收稿日期: 2010-08-24

接受日期: 2010-10-24

基金项目: 国家自然科学基金项目(30900119; 30771479); 中国科学院重要方向性项目(KSCX2-YW-N-061)资助

作者简介: 周锦(1984 ~), 男, 硕士研究生, 从事猕猴桃生物多样性研究, email: zhousjin08@scbg.ac.cn

* 通讯作者 Corresponding author, email: huanghw@mail.scbg.ac.cn

(Expressed Sequence Tags, ESTs)在3大数据库中迅速增加,这些 EST 序列不仅对发现、克隆和定位新基因起到了重要的作用,而且为以 PCR 技术为基础开发各种新型功能分子标记提供了重要的资源^[4-5]。

在基于 EST 数据库发展起来的分子标记中,单核苷酸多态性分子标记(Single Nucleotide Polymorphism, SNP)具有位点分布广泛、遗传代表性和稳定性高且易于实现自动化分析等特点,受到广泛关注。基于表达序列标签发展起来的单核苷酸多态性分子标记(EST-SNP)在遗传图谱构建、重要性状基因定位、比较遗传作图、遗传多样性分析和品种鉴别、分子标记辅助选择育种等方面发挥了重要作用。目前对基于现有 EST 数据库信息进行相关 SNP 标记的开发和应用的研究并不多,其潜在的生物信息学方法的应用及标记定位分离技术仍需进一步的探讨。为此,本文介绍了 EST-SNP 分子标记的特点,以及基于现有的 EST 数据库资源在猕猴桃属(*Actinidia*)植物中进行 SNP 标记开发的策略分析及初步结果,以期为其它物种的 SNP 标记开发研究提供借鉴。

1 EST-SNP 的特点

SNP 是指基因组 DNA 中某一特定核苷酸位置上发生的转换、颠换、插入、缺失等变化,其中任何一种等位基因在群体中出现的频率不少于 1%^[6],但也有 cDNA 中频率低于 1% 的单核苷酸变异^[7]。表达序列标签(Expressed Sequence Tags, EST)是指通过对 cDNA 文库随机挑取的克隆进行大规模测序所获得的 cDNA 5'或 3'端长度为 150~500 bp 的序列^[8],它代表了特定组织或特定时期基因表达的特征^[9]。随着功能基因组学的飞速发展,出现了以特定 EST 区段内单个核苷酸变异为基础的分子标记,称之为 EST-SNP 标记,其主要来自基因组的特定表达序列部分,因而有别于随机基因组 SNP 位点。

SNP 分子标记有无法比拟的优势。首先,单核苷酸多态性本身是生物遗传变异的根本原因,位于基因内部的 SNP 可能直接影响相关基因的表达水平和蛋白质结构,对于研究生物体的形态或性状变异以及适应性进化具有先天的优势^[10-12]。其次,由于 SNP 位点多呈现二等位基因,在进行基因分型

研究时往往只需对 SNP 位点进行 +/- 的分析,易实现自动化的高通量分析^[13]。最后,相对于传统的分子标记,如 AFLP 和 SSR 等,SNP 具有高遗传稳定性和位点丰富等特点^[14],其应用范围更加广泛,是第三代分子遗传标记的典型代表^[15]。由于不同 SNP 位点在染色体上的连续分布,同一染色体上所有 SNP 位点可作为一个整体或一个单倍型(Haplotype)进行遗传。基于单倍型的分析比基于单个 SNP 分析可提供更多的生物学信息,且在分析 SNP 与表型相关性时更为有效^[16]。作为一种新型的功能分子标记,来自表达序列标签的 EST-SNP 除具备传统的 SNP 标记的优势外,还可能与功能基因表达有直接或间接的关系,从而强化了 SNP 标记在遗传研究中的应用。同时由于 EST-SNP 来自转录区,具有较高的保守性,在比较不同物种基因组时非常有利,因而被广泛应用于比较基因组学、进化基因组学和候选基因的筛选等方面^[17]。最近, Rajeev 等^[18]提出,分子标记的定性特征(Qualitative nature of data, QND)也可作为选择的标准之一。QND 是由文档记录的便捷性、标记的质量和标记的可重复性 3 个重要指标综合决定的,其中文档记录的便捷性和标记的可重复性为两个固定的指标。相比而言,EST-SNP 分子标记这两个固定指标的评分要高于 EST-SSR 和 EST-AFLP 分子标记,表明 EST-SNP 分子标记在现代居群与进化遗传学研究中的优势。

2 EST-SNP 分子标记的开发策略

2.1 数据库和生物信息学的利用

全基因组 SNP 的开发策略依赖于基因组草图搜索法,即通过比较基因组中不同染色体的测序结果发现 SNP 位点^[19-20],这种开发策略多应用于人类和其他模式生物中,对其它大多数非模式物种的使用仍存在困难。随着美国国立生物技术信息中心(National Coalition Building Institute, NCBI)、欧洲分子生物学实验室(European Molecular Biology Laboratory, EMBL)和日本 DNA 数据库(DNA Data Bank of Japan, DDBJ)中存储的 DNA 序列数据的迅速增加,这 3 大数据库成为寻找 SNP 新的重要平台。这些数据库允许任何机构和个人随时提交各种序列,从而导致某一特定区段的数据在数据库中存在多拷贝,形成不同遗传位点的多态性^[21]。采用

生物信息学分析软件,对这些冗余序列进行自动识别并进行 SNP 标记位点的发掘利用已经成为一种简单有效且廉价的 SNP 位点开发的新策略,且可减少工作的盲目性,省去许多重复测序工作,实现高效和低成本。

基于数据库和生物信息学手段开发 SNP 分子标记的途径主要有两种:一种是基于序列标签位点(Sequence Tagged Site, STS)和基因组序列的 SNP 筛查;另一种是利用已经公布的大量的 EST 序列资源进行筛查检测 SNP 位点^[22]。相对 EST 数据资源而言,目前已经公布的基因组和 STS 序列数量仍然不够,所以 EST-SNP 是目前基于生物信息学方法开发 SNP 标记的主要手段。

目前,从玉米(*Zea mays*)、拟南芥(*Arabidopsis thaliana*)、水稻(*Oryza sativa*)、大麦(*Hordeum vulgare*)等植物中已开发了大量 EST-SNP^[23-27]。EST-SNP 的开发可以分为 3 个主要的步骤:第一步是 EST 数据的获取与前期处理,从数据库中直接获取的 EST 序列往往包含一些低质量的小片段(<100 bp),同时还带有少量载体序列及末端存在 polyA/T“尾巴”的序列,这会影响到信息的分析,所以开发标记之前应去除这些“噪音”;第二步是 EST 数据的聚类 and 拼接,用聚类软件对 EST 序列进行聚类 and 拼接,得到多序列聚类簇(Multi-member cluster),用于发掘单位点的多态性;最后是制定有效的筛选策略,由于 EST 测序往往只是进行单次测序,而测序方法本身固有的局限性使得碱基间测序正确率各不相同,可能对基于生物信息学的 SNP 发掘造成干扰,因此,需要利用有效的判别标准和算法以排除假阳性位点。

2.2 用于 EST-SNP 分析的软件

EST 预处理软件 Cross_match 是 Phil Green 编写的拼接程序包的一部分,用于序列拼接 and 比对。它是一款非常好的除去 NCBI 数据库中载体序列和一些过长重复序列片段的工具,可以从 <http://www.phrap.org> 网址下载。和 BLAST 相比,它运行的速度稍慢但是准确性更高。RepeatMasker 软件也可以除去一些重复序列,以提高 EST 组装的准确率,可以从 <http://www.repeatmasker.org> 下载。这两个软件通常捆绑在一起用于 EST 序列的前期处理。

序列拼接聚类软件 聚类和拼接的目的就

是将来自同一个基因或遗传位点上具有重叠部分的 EST 整合至单一的簇(cluster)中。通过聚类和拼接,可产生较长的一致性序列,降低数据的冗余,纠正测序错误,用于检测选择性剪切。目前用的最多的软件是 CAP3 和 Phrap。CAP3 是 Huang 等^[28]开发的一套用于序列拼接的软件,它能应用正反向信息更正拼接错误并连接重叠群(Contigs);在序列拼接中应用读取序列的质量信息自动截去序列 5' 端和 3' 端的低质量区域,产生 Consed 程序可读的 ACE 格式拼接结果文件,这款软件适用于小的数据集或 EST 的拼接。另一款软件 Phrap 是 Phred-Phrap-Consed 软件包的一部分,Phrap 软件用哈密尔顿通路的拼接算法,产生准确的一致性序列并进行质量估计,因而具有较高的鉴别和处理重复序列的能力^[29]。和 CAP3 相比,Phrap 的运行速度更快,占用的计算机资源更少,适合大规模数据集或 EST 序列的拼接。不过,Phrap 软件对序列的敏感性不如 CAP3,在对小规模数据集拼接过程中可能会忽视序列的微小差异而造成 EST 的过度拼接^[28-29]。

SNP 分析软件 目前已有许多利用基因组数据库进行自动分析 SNP 位点的软件。这些软件主要有两种类型:一种是针对 EST 序列测序质量进行分析,利用碱基判断软件重新分析 EST 测序曲线^[30-31],再用统计学方法区分有效的 SNP 信息和随机序列误差及错误,如 Phred/Phrap/PolyBayes 程序等。PolyBayes 软件^[32-33]在基因组序列和峰图文件的支持下能够有效识别旁系同源基因,减少 SNP 的假阳性。然而,目前公共数据库中许多 EST 序列通常不附加测序曲线,因而 PolyBayes 软件的应用具有较大的局限性。另一种类型是利用序列比对的冗余性进行 SNP 位点的发掘,针对公共数据库中大量 EST 序列相继推出了无需测序曲线图的 EST-SNP 开发软件,如 AutoSNP^[34]和 SniPper^[35]等。

AutoSNP 是目前使用较多的软件。两条同源的 EST 序列在同一碱基位点均出现测序错误的可能性很小,只要 SNP 突变碱基在 EST 数据库中出现的次数超过 1 次便可判断假阳性结果。另外,几个相邻单碱基位点变异所构成的单倍型模块通常作为一个整体遗传下来,分析几个相邻的 SNP 位点是否满足单倍型的要求,作为筛选 SNP 的一个标准。通过这两条筛选措施,AutoSNP 能够有效地

开发 SNP 位点,但是无法完全消除旁系同源基因的干扰,一些测序错误导致的假阳性 SNP 也可能出现。根据单倍型内 SNP 的相似系数辨别旁系同源基因或部分同源基因,在 AuotSNP 软件的策略基础上开发出 QualitySNP 软件^[36],它能有效地排除旁系同源基因的干扰,并解决测序错误引起的假阳性问题。Tang 等^[36]用该软件成功从马铃薯(*Solanum tuberosum*)两个品种 161202 条 EST 序列中开发出 17745 个 SNP,用重测序方法验证,有 94.3% 得到确认。

AutoSNP 和 QualitySNP 软件都可以通过建立自己的数据库和公共数据库相关联,完成基因序列的注释工作。用户也可以使用一些常用的注释工具如 BLAST 等将拼接而成的一致性序列和公共数据库已知的非冗余蛋白质数据库(nr)、非冗余核酸数据库(nt)、蛋白质功能结构域(domain)及基序(motif)等二级数据库进行同源性比对,从而赋予这些一致性序列一定的基因功能信息。

2.3 EST-SNP 位点的验证方法

用生物信息学方法开发的 SNP 位点不可避免的有假阳性产生,需要通过其它的实验方法检测之后才能使用。目前 SNP 的检测分析技术非常多,最直接的方法是进行目标片段测序,也是最具信服力的标准,但费用较高,不适合批量 SNP 位点的检测。传统的检测技术是进行小通量 SNP 分型测试,例如酶切扩增多态性序列(Cleaved amplified polymorphic sequence, CAPS)或 dCAPS 法^[37]、单链构象多态性(single strand conformation polymorphis, SSCP)、等位基因特异性 PCR(A allele specific PCR, AS-PCR)^[38]等。这些技术虽然难以实现大规模高通量 SNP 位点的筛选和自动化检测,但只需常规的实验试剂和设备,可以较方便快捷地完成少量 SNP 位点的检测工作。基于荧光标记的 SnaPshot 方法、基质辅助激光解吸电离-飞行时间质谱法(MALDI-TOF)^[39]和高分辨溶解曲线分析(high resolution melting, HRM)^[40]都适用于中等通量 SNP 位点的检测验证,对设备仪器有不同要求,针对不同的实验条件,可以选择合适的方法加以应用。此外,SNP 芯片是一种通量很高的检测 SNP 的手段,一次可同时检测百万个 SNP,是大规模分析的最主要手段之一^[41]。最新推出的 Affymetrix SNP 6.0 它可以检测 931946 个 SNP,主要应用于大规模人群

的关联性分析^[42]。另外,新出的 Illumina 芯片技术完全有别 Affymetrix 的原位合成专利技术,用微珠为载体,可以产生高密度的探针分布,特别适合大规模的 SNP 检测,但成本较高。

3 EST-SNP 分子标记的应用

生物进化及多样化与基因组的遗传变异紧密相关,同时外在环境的选择作用会对特定基因的进化方式产生影响^[43],对基于编码区标记的基因组扫描更容易分析这些特定基因的分子适应过程。Marie 等^[44]选择数量性状有差异的白云杉(*Picea glauca*) 6 个自然居群,分析了 345 个表达基因的 534 个 EST-SNP 位点,通过计算居群间的 SNP 发生频率和估计异常值鉴定出自然选择下的候选基因,并用贝叶斯方法分析位点差异,结果表明这种分离位点和居群影响的分析方法对于鉴定选择压力下的候选基因是非常有效的。利用物种间 DNA 序列尤其是编码序列的保守性^[45],通过比较 EST-SNP 分子标记在不同物种基因组中的分布,可以揭示染色体的共线性,从而对不同物种的基因组结构及基因组演化历程进行精密分析。Jose 等^[46]用生物信息学方法,利用 SGN 数据库中番茄(*Lycopersicon esculentum*) 3 个品种的 EST 数据开发出 SNP 标记,并对野生种和栽培种的序列进行差异分析,用 McDonald-Kreitman 分析法^[47]和以密码子为基础的最大似然法^[48-49]分析序列的进化与适应、自然选择的过程,揭示了一些与野生种的进化及栽培种的驯化过程相关的基因和基因家族,为中性进化学说提供了一些支持的依据。

在居群遗传学研究中,利用 EST-SNP 分子标记可以分析同源 DNA 序列的遗传分化程度,进而了解物种的遗传多样性和系统进化的关系。Gary 等^[50]用几种主要农作物的 350 多万条 EST 开发出十多万候选 SNP 位点,并分析比较了不同农作物中 SNP 的分布密度、连锁程度,结果表明异源多倍体小麦(*Triticum aestivum*)和八倍体甘蔗(*Saccharum officinarum*)序列的 SNP 多态性最高,其次是四倍体玉米,二倍体水稻、高粱(*Sorghum bicolor*)等最低。毛新国等^[51]用生物信息学方法开发出 1296 个小麦 SNP,通过比较 3 个基因组供体的 EST-SNP 数量和利用 22 个基因构建的进化树进行系统进化分析,结果表明小麦的 3 个基因组供

体中 A、D 基因组亲缘关系比较近,而与 S 基因组的的关系比较远。Schmid 等^[52]利用拟南芥种间的 STS 和 EST 比对发掘 SNP,对 12 份拟南芥材料进行系统分类学分析,并用邻近归并法(Neighbor-Joining)构建系统树,揭示了其中两个生态型 Cvi-O 与 Col-O 之间存在较大遗传距离。

植物遗传多样性评价是种质资源保护、开发和利用的基础,利用分子标记可极大地提高发掘优异自然种质资源的效率。功能性 EST-SNP 分子标记是直接对目标基因区的扩增,有利于对已知功能基因进行克隆定位^[53]。Feng 等^[54]用 EST-SNP 分子标记构建了白菜型油菜(*Brassia campestris*)的连锁图谱,通过与拟南芥比较,鉴定出控制花期的 *BrFLC1* 和 *BrFLC2* 基因,并在早花期亲本系的启动子中鉴定出 3 个与 *BrFLC2* 基因低表达相关的 SNP 位点。Yang 等^[55]用番茄(3 个品种的 EST 数据,在 44 个基因中发掘了 101 个候选 SNP 位点,经检测证明 83% 是真实的,将 46 个 SNP 定位在染色体的特定区域,证明 2 个 SNP 与果实颜色基因座紧密连锁。与性状相关基因紧密连锁标记的获得,提高了育种的选择性与预见性,将加快新品种的选育进程。Umemoto 等^[56]报道水稻编码区的淀粉酶 II a (*SSIa*)基因位点存在 5 个 SNPs,其中 1 个 SNP 能引起氨基酸编码的改变,靠近 C 末端的两个 SNPs 能引起氨基酸的变化,可能是引起淀粉酶 II a 基因活性改变的关键因素,从而为培育中等直链淀粉含量的水稻恢复系提供了理论指导。Shu 等^[57]用生物信息学方法将大豆(*Glycine max*) EST 序列联配到大豆基因组序列上,发掘出 EST-SNP 位点 537 个,有效促进了大豆的遗传育种研究。

4 EST-SNP 标记开发前景展望

通过 EST 数据库开发 SNP 并不是将来自不同个体的同源 EST 进行简单比对找出差异碱基,实际操作过程中会面临许多问题,需要制定有效的措施加以解决。构建 cDNA 文库时,各种反转录酶和 DNA 聚合酶会不可避免地引入错配^[58]。同时双脱氧核苷酸终止测序法本身固有的局限性使得同一读序碱基间测序正确率不同,特别在 EST 序列两端测序质量普遍较低。为避免这个问题,需要在筛选时选择高质量的序列区域的位点进行比对分析。再比如,在序列比对和拼接过程中,旁系同源基因

(Paralogous gene)容易被错误当成等位基因进行比对拼接,导致开发的 SNP 存在假阳性。因此需要设计相应的算法排除非目标序列的干扰。

大部分开发的 EST-SNP 位点在二倍体植物中能够得到有效检测和证实。但自然界中许多重要植物都包含多倍体,如油菜(*Brassica napus*)、棉花(*Gossypium hirsutum*)、烟草(*Nicotiana tabacum*)等都是异源四倍体,普通小麦等是六倍体。二倍体植物的 SNP 位点检测方法应用于多倍体植物中可能存在困难,虽然测序的方法可以鉴定同源多倍体真实的 SNP 位点^[59],但阵列杂交技术在分析杂合多倍体时却有很大困难。现在的分型技术无法区分 SNP 位点是在异源多倍体染色体之间还是染色体之内^[60]。所以怎样在异源多倍体物种中鉴定 SNP 仍然是一个难题。这也是 EST-SNP 等分子标记在多倍体植物中应用的瓶颈。

随着生物信息学算法的完善,新的分析 SNP 的软件不断产生和成熟,软件和分析策略已不再是有效开发 SNP 标记位点的最大制约因素,而大量相关 EST 数据库的不断积累和发展便显得更加关键。大量数据的积累有利于发掘基因组上更多相关基因本身遗传变异的 SNP 位点,有助于进行特异数量遗传性状与个体分子标记的关联性分析,也叫全基因组扫描(Whole-genome Scanning, WGS)、全基因组水平的关联性分析(Genome-wide Association Studies, GWAS)或关联性遗传(Association Genetics)^[61]。另外,目前 SNP 位点的鉴定和验证技术也在不断发展,SNP 芯片等新的高通量检测技术的出现和发展成熟,将会极大地促进利用生物信息学方法开发 SNP 分子标记。据报道,Affymetrix 公司最新开发了 1 张经过 40 次测序验证总共包含 11500 个人类 SNP 的测序芯片,而高密度的水稻 SNP 芯片也即将问世^[42]。可以预见,大规模基因挖掘和功能鉴定的热潮即将到来,基于 EST 数据库开发 SNP 标记生物信息策略将会被越来越多地使用。

5 猕猴桃属植物开发 EST-SNP 标记的可行性分析

猕猴桃属(*Actinidia* Lindl.)隶属于猕猴桃科(Actinidiaceae Hutch.),有 55 种 20 变种,约 75 个分类群^[62]。其栽培品种也比较多,主要是基于中华猕

猴桃(*A. chinensis*)复合体的两个变种,即中华猕猴桃(*A. chinensis* var. *chinensis*)和美味猕猴桃(*A. chinensis* var. *deliciosa*)培育而来。猕猴桃属植物主要分布在东亚地区,中国为其分布中心,少数种分布于南亚地区,其自然地理分布广泛,从北纬 50°到赤道附近都有,纵跨泛北极和热带亚热带植物区系^[63]。目前,对猕猴桃属植物的基因组信息了解甚少,许多与重要经济性状相关的遗传位点的遗传方式仍不清楚^[64]。开发 SNP 分子标记对于分析猕猴桃重要特征性状的遗传机制、功能基因的定位和克隆以及优异种质资源的发掘和遗传育种等有重要作用。

目前有关猕猴桃 EST 数据库的 SNP 分析的报道较少,新西兰的 Crowhurst 等^[65]对猕猴桃进行了 SNP 的筛选,但并没有探讨开发的 SNP 位点的发生规律、分布特点、碱基变化类型等。本文将探讨

从 NCBI 数据库的中华猕猴桃复合体 EST 序列进行 SNP 分子标记开发的可行性,同时介绍利用 Dave 的 PERL 语言程序 AutoSNP 2.0 筛查猕猴桃 EST 数据库中 SNP 信息的大概思路,为开发新型猕猴桃 EST-SNP 分子标记及其应用奠定理论基础。

5.1 中华猕猴桃复合体 EST-SNP 开发方法及策略
EST 序列分析 截至 2009 年 11 月,NCBI 数据库中 共有猕猴桃 EST 序列 132577 条,其中中华猕猴桃有 47379 条,美味猕猴桃有 57751 条(表 1),分别占总数的 35.74% 和 43.56%。但无重叠序列的数量还不到序列总数的五分之一,说明中华猕猴桃复合体的 EST 有较高的冗余性,适合进行 EST-SNP 分析,为此制定了利用序列冗余性的策略,用 AutoSNP 软件来分析中华猕猴桃复合体的 EST 序列。

表 1 中华猕猴桃复合体 EST 数量信息
Table 1 Information of EST database of the *Actinidia chinensis* species complex

	cDNA 文库数量 Number of cDNA library	序列总数 Number of EST sequences	无重叠序列 Number of singletons	%	平均 EST 长度 Average length of EST
中华猕猴桃 <i>A. chinensis</i> var. <i>chinensis</i>	21	47379	7962	16.8	558
美味猕猴桃 <i>A. chinensis</i> var. <i>deliciosa</i>	14	57751	11691	20.2	453
合计 Total	35	105130	19653	18.7	
平均 Mean				500	

EST-SNP 的识别 利用 Cross_match 和 RepeatMasker 软件对所有的 EST 序列进行预处理,利用 D2cluster 聚类软件进行原始聚类,参数设置为“80% similarity”,然后利用 CAP3 软件组装重叠群,参数设置为“minmatch 100, overlap similarity 0.97”,即序列比对重叠区域最少要有 100 个碱基而且相识度达到 97%。Blastx 用来帮助处理非冗余序列数据库,以鉴定不同重叠群的 EST 序列。BLASTHITS 设置为程序默认参数。AutoSNP 软件是 Perl 语言程序,会自动分析在 D2cluster 和 CAP3 产生的 Fasta 或者 Ace format 格式数据,利用数据库中序列的冗余性比对差异分析出 SNP 位点。整个开发的流程如下:从 NCBI 数据库中下载中华猕猴桃复合体 EST→预处理 EST 序列→EST 序列拼接聚类→利用 SNP 碱基的冗余性和单倍型策略,分析 SNP 位点→候选 SNP 位点。

5.2 EST-SNP 分析初步结果
5.2.1 SNP 的发现
中华猕猴桃复合体的 EST 序列中含有至少两条比对序列的片段重叠群共有 4060 个,涉及的序列 28801 条,共发现 8886 个单核苷酸多态性位点。而含有至少 4 条比对序列的重叠群中有 4893 个 SNP 位点,其中含有 SNP 的重叠群为 1272 个,占含至少 4 条比对序列的重叠群总数的 60.3%。
对重叠群中含有的 SNP 数量的统计结果表明,构成重叠群的 EST 序列越多,产生的 SNP 位点数量也越多(图 1)。这同 Chris 等^[27]对大麦 EST 序列的研究结果基本一致。但是具体每个类型重叠群所含的 SNP 位点数并没有明显的规律,这可能和不同物种的多态性位点分布不同有关。
利用 AutoSNP 软件对于真实 SNP 位点分析判定标准,在含有较多位点的重叠群类型中可以利用

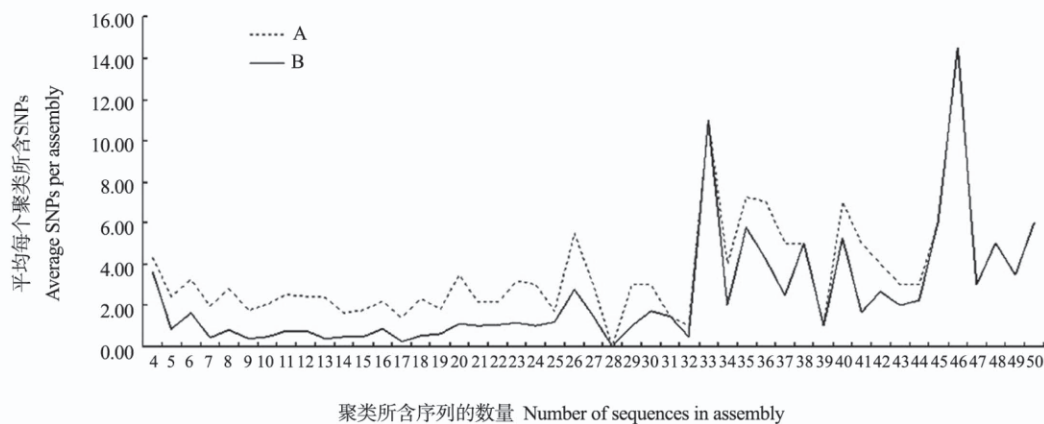


图 1 重叠群中的 SNP 位点平均数

Fig. 1 Average single nucleotide polymorphisms (SNPs) per assembly

A. SNP 数量占有所有重叠群数量的百分比 Percentage of SNP to all assembly;B. SNP 数量占含 SNP 位点重叠群数量的百分比 Percentage of SNP to all assembly with SNPs.

相邻 SNP 位点的单倍型共分离得到一批真实度更高的位点,SNP 数目越多含有的信息量越大^[66]。中华猕猴桃和美味猕猴桃的 EST 序列分析表明,聚类数量较大的重叠群更容易得到真实度较高的 SNP 位点。

5.2.2 SNP 丰富度

含 EST 数目大于 2 的簇(重叠群)的一致性序列长度共计 3101544 bp,中华猕猴桃复合体 cDNA 中 SNP 发生的频率为 0.29 (100 bp)⁻¹。Newcomb 等^[67]报道苹果(*Malus domestica*)的 150000 条 EST 中,SNP 发生频率为 0.14 (100 bp)⁻¹。可见,中华猕猴桃复合体的 EST-SNP 的发生频率明显比近缘的

苹果高,可能是由于中华猕猴桃复合体的多倍化特性导致旁系同源和部分同源基因的干扰,或重新测序产生的 EST 序列质量问题导致的。

为了获得真实度较高的 SNP 位点,规定至少有 4 条 EST 序列才能形成 1 个聚类群。聚类 and 拼接后产生的一致性序列长度为 2045019 bp,共有 4893 个 SNP 位点,SNP 发生频率为0.24 (100 bp)⁻¹。另外从表 2 可以看出,SNP 在 EST 序列上呈现不均匀的分布。

5.2.3 碱基变化及插入缺失分析

对 EST 数目大于 2 的簇中产生的 8886 碱基突变类型进行分析,结果表明有 50 个多等位基因位

表 2 重叠群所含 SNP 的数量以及频率

Table 2 The number and frequency of SNP in the contigs

重叠群所含序列的数量	SNP 数量	一致性序列长度	SNP 频率
Number of sequences in each contig	Number of SNP	Total consensus length (bp)	SNP frequency (per 100 bp)
2	1759	569173	0.31
3	2234	487352	0.46
4	3225	629676	0.51
5	243	219200	0.11
6 ~ 10	447	483436	0.09
11 ~ 20	303	473527	0.06
21 ~ 30	110	99267	0.11
31 ~ 40	116	34170	0.34
41 ~ 50	102	31699	0.32
> 50	347	74,044	0.47
合计 Total	8886	3101544	0.29

点,4259 个转换类型,3733 个颠换类型(图 2)。转换类型明显高于颠换,这与前人^[68-69]的研究结果一致。转换类型特别是 C-T 转换发生的频率很高,可能是由于 C(胞嘧啶)常以甲基化的形式存在,脱氨后即成为胸腺嘧啶(T)^[70],其它类型的碱基变化的原因尚不清楚。除了转换和颠换的类型,还有 844 个插入和缺失的类型,插入和缺失的发生频率是 0.027 (100 bp)⁻¹,大约 10 个碱基突变之中就有 1 个插入缺失的变化。

5.2.4 基因注释

选择至少含有 9 条比对序列的重叠群中产生的一致性序列(共 691 条)进行基因注释,把这些

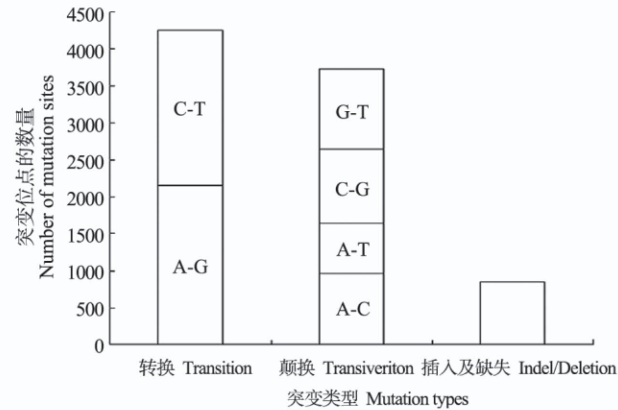


图 2 碱基突变分析
Fig. 2 Mutation of single nucleotide polymorphisms (SNPs)

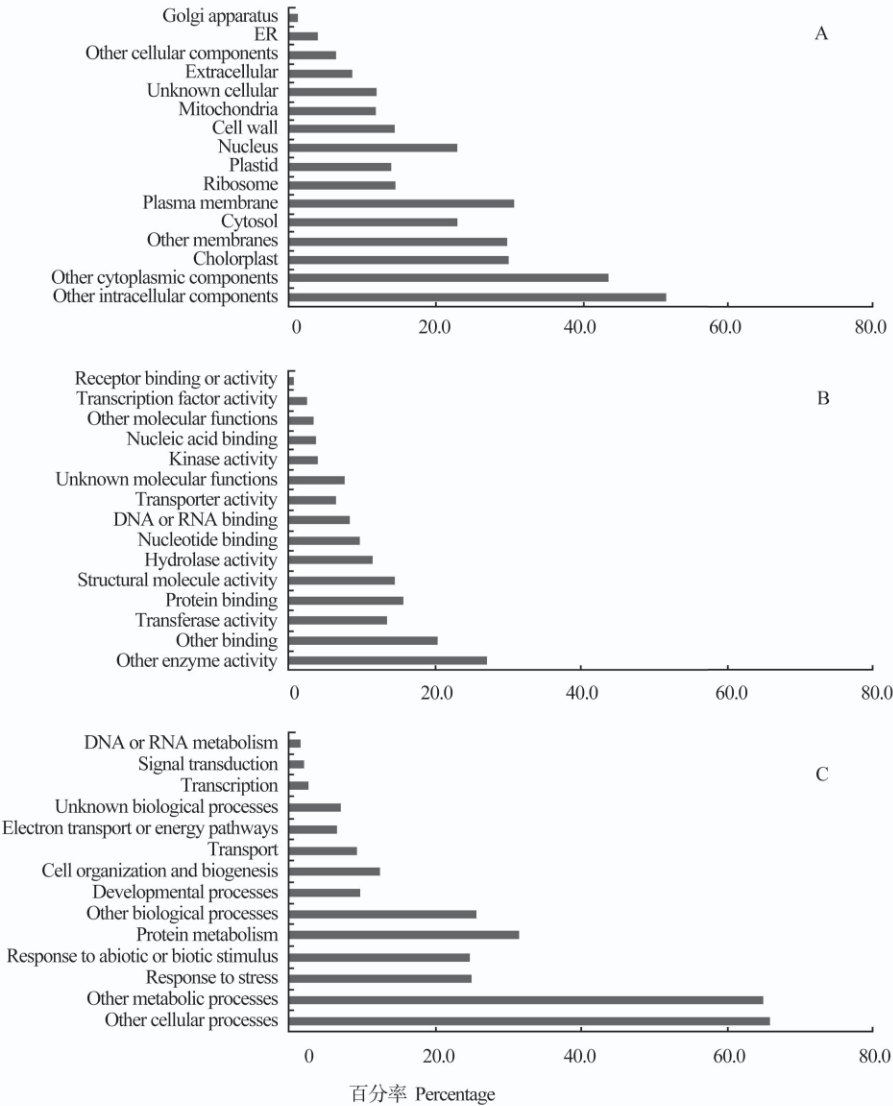


图 3 功能分类

Fig. 3 Functional categorization for loci

A: 细胞成分 GO cellular component (N=455); B: 生物过程 Biological process (N=460); C: 分子功能 Molecular function (N=464).

一致性序列和模式植物拟南芥数据库(TAIR)进行同源比对,共得到注释信息 5018 条。这些基因注释按功能分类可分为细胞成分、生物过程和分子功能 3 个部分,图 3 反映了基因注释的数量和涉及的基因总数 N 的百分比。

借助 AutoSNP 和其它相关生物信息学软件对中华猕猴桃复合体的 EST-SNP 分子标记开发进行初步的可行性分析和研究,成功开发 8886 个多态性位点,其中 4 条以上比对序列产生的高质量位点 4893 个,证明在该复合体中进行 EST-SNP 标记开发是可行的。同时,对中华猕猴桃复合体可预测的 SNP 位点的统计分析表明,SNP 位点在 EST 序列上呈现不均匀的分布,转换类型高于颠换,几个相邻的 SNP 位点组成的单倍型多见于较大规模的重叠群中,这些为猕猴桃 EST-SNP 的后续分子实验验证及其在自然居群中的应用奠定了基础,并为在其它物种中进行有效 EST-SNP 分子标记的开发提供了宝贵的经验和理论指导。

致谢 感谢澳大利亚昆士兰大学的 Dave Edwards 博士和美国亚利桑那大学的 Michael Shane Barker 博士对实验技术指导。

参考文献

- [1] Sunnucks P. Efficient genetic makers for population biology [J]. Trends Ecol Evol, 2000, 15(5): 199–203.
- [2] Boguski M S. The turning point in genome research [J]. Trend Biochem Sci, 1995, 20(8): 295–296.
- [3] Andersen J R, Berstedt T L. Functional markers in plants [J]. Trends Plant Sci, 2003, 8(11): 554–560.
- [4] O'Brien S J, Womack J E, Lyons L A, et al. Anchored reference loci for comparative genome mapping in mammals [J]. Nat Genet, 1993, 3: 103–112.
- [5] Hieter P, Boguski M. Functional genomics: It's all how you read it [J]. Science, 1997, 278(5338): 601–602.
- [6] Jia Y Y(贾玉艳), Chen H(陈宏). Molecular genetic marker of single nucleotide polymorphism and its application [J]. J Yellow Cattle Sci(黄牛杂志), 2003, 29(1): 42–45.(in Chinese)
- [7] Brooker A J. The essence of SNPs [J]. Gene, 1999, 234(2): 177–186.
- [8] Luo M(骆蒙), Jia J Z(贾继增). Progress in expressed sequence tags (EST) project of plant genome [J]. Progr Biochem Biophys (生物化学与生物物理进展), 2001, 28: 494–497.(in Chinese)
- [9] Adams M D, Kelley J M, Goeayne J D, et al. Complementary DNA sequencing: Expressed sequence tags and human genome project [J]. Science, 1991, 252(5013): 1651–1656.
- [10] Butcher L M, Davis O S, Craig I W, et al. Genome wide quantitative trait locus association scan of general cognitive ability using pooled DNA and 500K single nucleotide polymorphism microarrays [J]. Genes Brain Behav, 2007, 7(4): 435–446.
- [11] Kiyohara C, Yoshimasu K. Genetic polymorphisms in the nucleotide excision repair pathway and lung cancer risk: A meta-analysis [J]. Inter J Med Sci, 2007, 4: 59–71.
- [12] Lazarus R, Vercelli D, Palmer L J, et al. Single nucleotide polymorphisms in innate immunity genes: Abundant variation and potential role in complex human disease [J]. Immunol Rev, 2002, 190: 9–25.
- [13] Rafalski A. Applications of single nucleotide polymorphisms in crop genetics [J]. Curr Opin Plant Biol, 2002, 5(2): 94–100.
- [14] Marshall B, Leelayuwat C, Degli-Esposti M A, et al. New major histocompatibility complex genes [J]. Hum Immunol, 1993, 38(1): 24–29.
- [15] Moreno-Vazquez S, Ochoa O E, Faber N, et al. SNP-based codominant markers for a recessive gene conferring resistance to corky root rot (*Rhizomonas suberifaciens*) in lettuce (*Lactuca sativa*) [J]. Genome, 2003, 46(6): 1059–1069.
- [16] Salisbury B A, Pungliya M, Choi J Y, et al. SNP and haplotype variation in the human genome [J]. Mutation Res, 2003, 526: 53–61.
- [17] Kantety R V, Rota M L, Matthews D E, et al. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat [J]. Plant Mol Biol, 2002, 48: 501–510.
- [18] Varshney R K, Chabane K, Hendre P S, et al. Comparative assessment of EST-SSR, EST-SNP and AFLP markers for evaluation of genetic diversity and conservation of genetic resources using wild, cultivated and elite barleys [J]. Plant Sci, 2007, 173: 638–649.
- [19] Suh Y, Vijg J. SNP discovery in associating genetic variation with human disease phenotypes [J]. Mutation Res, 2005, 573: 41–53.
- [20] Wang S L, Sha Z X, Sonstegard T S, et al. Quality assessment parameters for EST-derived SNPs from catfish [J]. BMC Genom, 2008, 9(1): 450–461.
- [21] Wang J, Zou Q, Guo M Z. Mining SNPs from EST sequences using filters and ensemble classifiers [J]. Genet Mol Res, 2010, 9(2): 820–834.
- [22] Guryev V, Koudijs M J, Berezikov E, et al. Genetic variation in the zebrafish [J]. Genom Res, 2006, 16(4): 491–497.
- [23] Barker G, Batley J, Sullivan H O, et al. Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP [J]. Bioinformatics, 2003, 19: 421–422.
- [24] Useche F J, Gao G, Hanafey M, et al. High throughput identification database storage and analysis of SNPs in EST sequence [J]. Genom Inform, 2001, 12: 194–203.
- [25] Iwata H, Ujino-Ihara T, Yoshimura K, et al. Cleaved amplified polymorphic sequence markers in sugi and their locations on a linkage map [J]. Theor Appl Genet, 2001, 103: 881–895.
- [26] Yang L(杨仑), Shen W B(沈文彪), Chen H(陈虹), et al. Developing putative rice SNPs through bioinformatics approaches [J]. Chin J Rice Sci(中国水稻科学), 2004, 18: 185–191.(in Chinese)

- [27] Chris D, Nikki A, Megan V, et al. Single nucleotide polymorphism discovery in barley using autoSNPdb [J]. *Plant Biotechn J*, 2009, 7: 326–333.
- [28] Huang X, Madan A. CAP3: A DNA sequence assembly program [J]. *Genome Res*, 1999, 9: 868–877.
- [29] de la Bastide M, McCombie W R. Assembling genomic DNA sequences with PHRAP [J]. *Curr Protoc Bioinform*, 2007, Chapter 11(Unit 11): 14.
- [30] Brent E, LaDeana H, Michael C W, et al. Basecalling of automated sequencer traces using Phred I Accuracy Assessment [J]. *Genom Res*, 1998, 8(3): 175–185.
- [31] Brent E, Phil G. Basecalling of automated sequencer traces using Phred II Error Probabilities [J]. *Genom Res*, 1998, 8(3): 186–194.
- [32] Marth G T, Korf I, Yandell M D, et al. A general approach to single nucleotide polymorphism discovery [J]. *Nat Genet*, 1999, 23: 452–456.
- [33] Le D L, Chagne D, Pot D, et al. Automated SNP detection in expressed sequencetags: Statistical considerations and application to maritime pine sequences [J]. *Plant Mol Biol*, 2004, 54: 461–470.
- [34] Batley J, Barker G O, Sullivan H, et al. Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data [J]. *Plant Physiol*, 2003, 132: 84–91.
- [35] Kota R, Rudd S, Facius A, et al. Snipping polymorphisms from large EST collections in barley (*Hordeum vulgare* L.) [J]. *Mol Gen Genom*, 2003, 270: 24–33.
- [36] Tang J, Vosman B, Voorrips R E, et al. QualitySNP: A pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species [J]. *BMC Bioinform*, 2006, 7: 438–443.
- [37] Konieczny A, Ausubel F M. A procedure for mapping Arabidopsis mutations using co-dominant ecotype-specific PCR-based markers [J]. *Plant J*, 1993, 4(2): 403–410.
- [38] Pui-Yan K. Single Nucleotide Polymorphisms Methods and Protocols [M]. Totowa, New Jersey, United States: Humana Press Inc, 2003: 71–84.
- [39] Ross P, Hall L, Smirnov I, et al. High level multiplex genotyping by MALDI-TOF mass spectrometry [J]. *Nat Biotechn*, 1998, 16: 1347–1351.
- [40] Montgomery J, Wittwer C T, Palais R, et al. Simultaneous mutation scanning and genotyping by high-resolution DNA melting analysis [J]. *Nat Protoc*, 2007, 2(1): 59–66.
- [41] Nishida N, Koike A, Tajima A, et al. Evaluating the performance of Affymetrix SNP Array 6.0 platform with 400 Japanese individuals [J]. *BMC Genom*, 2008, 9: 431–441.
- [42] Fang W Y(方唯意). Progress in high-throughput genotyping of SNPs [J]. *Foreign Med Sci Mol Biol Fasc(国外医学分子生物学分册)*, 2003, 25(6): 333–337.(in Chinese)
- [43] Wang Y S(王云生), Huang H W(黄宏文), Wang Y(王瑛). Recent progress in plant molecular population genetics [J]. *Hereditas(遗传)*, 2007, 29(10): 1191–1198.(in Chinese)
- [44] Marie-Claire N, Jean B, Nicolas J, et al. Scanning the genome for gene single nucleotide polymorphisms involved in adaptive population differentiation in white spruce [J]. *Mol Ecol*, 2008, 17: 3599–3613.
- [45] Ann-Christine S. Accessing genetic variation: Genotyping single nucleotide polymorphisms [J]. *Nat Rev Genet*, 2001, 2 (12): 930–942.
- [46] Jimenez-Gomez J M, Maloof J N. Sequence diversity in three tomato species: SNPs, markers, and molecular evolution [J]. *BMC Plant Biol*, 2009, 9(1): 85–96.
- [47] McDonald J H, Kreitman M. Adaptive protein evolution at the *Adh* locus in *Drosophila* [J]. *Nature*, 1991, 351(6328): 652–654.
- [48] Yang Z, Swanson W J. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes [J]. *Mol Biol Evol*, 2002, 19(1): 49–57.
- [49] Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution [J]. *Mol Biol Evol*, 1998, 15(5): 568–573.
- [50] Gary L, Barker A, Keith J E. A genome-wide analysis of single nucleotide polymorphism diversity in the world's major cereal crops [J]. *Plant Biotechn J*, 2009, 7: 318–325.
- [51] Mao X G(毛新国), Tang J F(汤继凤), Zhou R H(周荣华), et al. Wheat cSNP mining based on full-length cDNA sequences [J]. *Acta Agron Sin(作物学报)*, 2006, 32(12): 1836–1840.(in Chinese)
- [52] Schmid K J, Ramos-Onsins S, Ringys-Beckstein H, et al. A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism [J]. *Genetics*, 2005, 169(3): 1601–1615.
- [53] Zhao X(赵雪), Xie H(谢华), Ma R C(马荣才). New functional molecular markers for plants in the functional genomics era [J]. *Chin Biotechn(中国生物工程杂志)*, 2007, 27(8): 104–110.(in Chinese)
- [54] Feng L, Hiroyasu K, Kiyofumi I, et al. A *Brassica rapa* linkage map of EST-based SNP markers for identification of candidate genes controlling flowering time and leaf morphological traits [J]. *DNA Res*, 2009, 16: 311–323.
- [55] Yang W C, Bai X D, Kabelka E, et al. Discovery of single nucleotide polymorphisms in *Lycopersicon esculentum* by computer aided analysis of expressed sequence tags [J]. *Mol Breed*, 2004, 14: 21–34.
- [56] Umemoto T, Aoki N. Single nucleotide polymorphisms in rice starch synthase II a that alter starch gelatinisation and starch association of the enzyme [J]. *Funct Plant Biol*, 2005, 32(9): 763–768.
- [57] Shu Y J(束永俊), Li Y(李勇), Wu N(吴娜), et al. Mining and identification of SNP from EST sequences and conversion of CAPS markers in soybean [J]. *Acta Agron Sin(作物学报)*, 2010, 36(4): 574–579.(in Chinese)
- [58] Picoult-Newberg L, Ideker T E, Pohl M G, et al. Mining SNPs from EST databases [J]. *Genom Res*, 1999, 9: 167–174.
- [59] Li L, Paulo M J, Strahwald J, et al. Natural DNA variation at candidate loci is associated with potato chip color, tuber starch

- content, yield and starch yield [J]. *Theor Appl Genet*, 2008, 116: 1167–1181.
- [60] Somers D J, Kirkpatrick R, Moniwa M, et al. Mining single nucleotide polymorphisms from hexaploid wheat ESTs [J]. *Genome*, 2003, 46: 431–437.
- [61] McCarthy M I, Abecasis G R, Cardon L R, et al. Genome-wide association studies for complex traits: Consensus, uncertainty and challenges [J]. *Nat Rev Genet*, 2008, 9(5): 356–369.
- [62] Li J Q, Li X W, Soejarto D D. *Actinidiaceae* [M]// Wu Z Y, Raven P H, Hong D Y, et al. *Flora of China*. Vol. 12 (*Hippocastanaceae* through *Theaceae*). Beijing: Science Press, St. Louis: Missouri Botanical Garden Press, 2007: 334–360.
- [63] Huang H W(黄宏文), Gong J J(龚俊杰), Wang S M(王圣梅), et al. Genetic diversity in the genus *Actinidia* [J]. *Chin Biodiv(生物多样性)*, 2000, 8(1): 1–12. (in Chinese)
- [64] Ferguson A R, Huang H W. Genetic resources of kiwifruit: Domestication and breeding [J]. *Hort Rev*, 2007, 33: 1–121.
- [65] Ross N C, Andrew P G, Elspeth A M, et al. Analysis of expressed sequence tags from *Actinidia*: Applications of a cross species EST database for gene discovery in the areas of flavor, health, color and ripening [J]. *BMC Genom*, 2008, 9: 351–377.
- [66] Salisbury B A, Pungliya M, Choi J Y, et al. SNP and haplotype variation in the human genome [J]. *Mutat Res*, 2003, 526: 53–61.
- [67] Newcomb R D, Crowhurst R N, Gleave A P, et al. Analyses of expressed sequence tags from apple [J]. *Plant Physiol*, 2006, 141 (1): 147–166.
- [68] Garg K, Green P, Nickerson D A. Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags [J]. *Genome Res*, 1999, 9: 1087–1092.
- [69] Deutsch S, Iseli C, Bucher P, et al. A cSNP map and database for human chromosome 21 [J]. *Genome Res*, 2001, 11: 300–307.
- [70] Coulondre C, Miller J H, Farabaugh P J, et al. Molecular basis of base substitution hot spots in *Escherichia coli* [J]. *Nature*, 1978, 274: 775–780.