

辣椒转录组 SNP 挖掘及多态性分析

刘 峰^{1,3}, 谢玲玲², 弭宝彬¹, 欧阳娴¹, 茆振川³, 邹学校^{1,*}, 谢丙炎^{3,*}

(¹ 湖南农业科学院蔬菜研究所, 长沙 410125; ² 湖南农业科学院西瓜甜瓜研究所, 长沙 410125; ³ 中国农业科学院蔬菜花卉研究所, 北京 100081)

摘 要: 利用 SNP 分析软件从辣椒 (*Capsicum annuum* L.) 251 068 条 Unigenes 中筛选出 18 159 个 SNP, 其中有 1 781 个 SNP 位点被匹配在 1 291 个注释基因上, 基因功能分类和代谢途径分析表明, 其中有 853 个基因参与初生代谢 (28.7%)、细胞代谢 (17.3%)、生物合成过程 (15.7%), 另有 125 个 (9.7%) 基因序列参与新陈代谢途径, 53 条 (4.1%) 序列参与次生代谢产物合成途径, 31 条 (2.4%) 序列参与植物激素合成途径。EST-SNP 序列中 4 172 条 (22.9%) 满足设计 CAPS 引物条件, 为了验证 EST-SNP 正确性, 并选取了 15 对 CAPS 引物对 5 份辣椒材料进行扩增, 结果发现有 8 对 (53.3%) 引物表现出多态性。表明筛选出这些 EST-SNP 标记可作为辣椒基因分型、图谱构建等的候选分子标记。

关键词: 辣椒; 转录组; EST-SNP; CAPS

中图分类号: S 641.3

文献标志码: A

文章编号: 0513-353X (2014) 02-0343-06

SNP Mining in Pepper Transcriptome and the Polymorphism Analysis

LIU Feng^{1,3}, XIE Ling-ling², MI Bao-bin¹, OUYANG Xian¹, MAO Zhen-chuan³, ZOU Xue-xiao^{1,*}, and XIE Bing-yan^{3,*}

(¹Institute of Vegetables, Hunan Academy of Agricultural Sciences, Changsha 410125, China; ²Institute of Wa-termelon and muskmelon Hunan Academy of Agricultural Sciences, Changsha 410125, China; ³Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China)

Abstract: In this study, we investigated 251 068 unigenes from two transcriptome in pepper (*Capsicum annuum* L.) using SNP finding soft. In total, 18 159 SNP were identified from these SNP-containing unique ESTs. Among, 1 781 SNP located at 1 291 annotation genes, analysis of GO (Gene Ontology) showed that 853 ESTs were classified, among primary metabolic process (28.7%), cellular metabolic process (17.3%), biosynthetic process (15.7%) being main. In KEGG map, 125 ESTs involved in taken part in metabolic pathways, and 53 participated in biosynthesis of second metabolites, and 31 involved in biosynthesis of plant hormones. Moreover, a total of 4 172 (22.9%) sequences were successfully designed primers of CAPS marker. Eight out of 15 primer pairs selected at random showed polymorphism among 5 different pepper varieties. The results indicated that those CAPS markers from EST-SNP will be more usable in map structure, analysis of genetic polymorphism in pepper.

Key words: pepper; *Capsicum annuum* L.; transcriptome; EST-SNP; CAPS

收稿日期: 2013-08-15; **修回日期:** 2014-01-17

基金项目: 国家自然科学基金项目 (31101425); 国家重点基础发展计划项目 (2009CB119000); 国家现代产业技术体系建设专项资金项目 (CARS-25-B-01); 公益性行业 (农业) 科研专项 (2011303018)

* 通信作者 Author for correspondence (E-mail: zou_xuexiao@163.com; xieby@mail.caas.net.cn)

DNA 分子标记技术是基因定位克隆, 遗传图谱构建等研究工作的重要工具, 目前已广泛应用于作物分子辅助育种及种质资源多样性分析等领域(Chao et al., 2008), 特别是单核苷酸多态性(Single Nucleotide Polymorphism, SNP)研究备受关注(Gunderson et al., 2005; Syvanen, 2005)。研究发现, 生物个体中 SNP 数量比其他类型分子标记更为丰富、更具有多态性, 如玉米基因组中 SNP 频率为每 100 bp 就出现一个 SNP(Tenaillon et al., 2001), 水稻为 89 bp(Nasu et al., 2002), 葡萄为 47 bp(Lijavetzky et al., 2007)。这些海量的 SNP 将成为基因分型及遗传关联分析更为理想的分子标记(Syvanen, 2005; Karchin, 2009)。

随着测序技术的发展, 转录组测序已成为研究生物发育阶段、细胞类型、环境互作等复杂分子机制的重要手段, 同时也是鉴定多态性 SNP 分子标记的重要数据源。基于转录组挖掘 SNP 分子标记还具有两个显著优点, 一是公共数据库 EST 序列丰富, 二是这些潜在的 SNP 变异位于功能基因上可能直接与植物表型相关(Ganal et al., 2009)。目前基于转录组及其它 EST 序列开发的 SNP 标记已被广泛运用于玉米、大豆、水稻等(Li et al., 2009; Blair et al., 2013; Frascaroli et al., 2013)作物的遗传图谱构建和遗传多样性分析中。

辣椒(*Capsicum annuum* L.)是我国重要蔬菜之一, 其 SNP 分子标记的开发日益受研究者们重视, 如 Jung 等(2010)基于 COSII 引物扩增测序挖掘辣椒 SNP, Nicolai 等(2012)利用 Roch 454 Yolo Wonder 转录组测序和 CM334 Illumina GA II 测序数据挖掘辣椒 SNP。为了进一步丰富辣椒分子标记类型及数量, 本研究中利用甜椒‘HDA149’与辣椒‘9704A’转录组测序数据以及公共数据库辣椒 EST 序列进行 SNP 分析, 并基于 CAPS(Cleaved Amplified Polymorphic Sequence)对不同辣椒材料进行 SNP 多态性分析。

1 材料与方法

1.1 植物材料及其 DNA 提取

选用甜椒‘HDA149’(法国 INRA)、牛角椒‘9704B’(湖南省蔬菜研究所)、湘研 1 号(湖南省蔬菜研究所)、博辣 5 号(湖南省蔬菜研究所)、CM334(墨西哥 INIA)为试料。每份材料选 20 粒种子, 30 °C 恒温箱中育苗, 用 CTAB 法提取供试材料 DNA, -80 °C 保存备用。

1.2 辣椒 EST-SNP 数据分析

利用‘HDA149’幼根组织和‘9704B’花蕾组织转录组数据以及 NCBI(<http://www.ncbi.nlm.nih.gov>) All database 数据库中所有辣椒 EST 序列, 使用 Trilily 软件首先对所有序列进行拼接, 然后使用 Samtools 工具分析上述两个转录组序列 SNP 位点(<http://samtools.sourceforge.net/>)。

1.3 辣椒多态性 SNP-EST 序列功能分析

利用 Blast2go(<http://www.blast2go.org>)程序, 对多态性 SNP-EST 序列进行功能注释($E < 10^{-10}$)和 GO 分类(geneontology, <http://www.geneontology.org>)及 KEGG 分析(kyoto encyclopedia of genes and genomes, <http://www.genome.jp/kegg>)。

1.4 辣椒 SNP 位点 CAPS 标记引物设计与扩增

利用 perl 脚本程序查找 EST-SNP 突变位点的内切酶识别位点, 对于没有酶切位点的 EST-SNP 序列运用 SNP2CAPS 程序进行酶切识别位点碱基转换, 并设计引物。引物设计标准: (1) PCR 产物大小为 90~600 bp; (2) 引物长度为 23~25 bp; (3) 退火温度为 55~63 °C, 上、下游引物退火温

度相差不大于 5 ℃；(4) GC 含量为 40%~60%；尽量避免引物 Hairpin、Dimer、False Priming、Cross Dimer 情况出现。

PCR 扩增反应体系：2.5 mmol·L⁻¹ MgCl₂、1 U *Taq* 酶、800 μmol·L⁻¹ dNTP、10 mmol·L⁻¹ 10 × PCR buffer、0.2 μmol·L⁻¹ 引物、模板 DNA 50 ng，补 ddH₂O 至 25 μL。PCR 扩增程序：94 ℃预变性 3 min；94 ℃变性 30 s，最佳退火温度退火 30 s，72 ℃延伸 30 s，35 个循环；72 ℃延伸 5 min。

1.5 SNP 酶切与电泳检测及扩增产物测序

用限制性核酸内切酶酶切 PCR 产物，酶切体系包括 10 U·μL⁻¹ 限制酶 0.5 μL、1 μL buffer、5.5 μL ddH₂O、3 μL PCR 产物，置于酶切最适温度 8~10 h，酶切完全后用 2%~4%的琼脂胶对酶切产物进行电泳检测。回收 PCR 产物纯化后连接到 PMD-18 T 载体上，转化 TOP10 感受态细胞，经阳性检测后送测序公司测序。

2 结果与分析

2.1 辣椒转录组序列组装及 SNP 在 EST 序列中的特征

两个辣椒材料 ‘HDA149’、‘9704B’ 经 illumina 转录组测序分别获得 4.5 G、5 G 原始 reads (表 1)，NCBI 上共获得辣椒 EST 序列约 150 Mb，经 Trilily 软件组装共获得 251 068 条 Unigenes，总覆盖长度为 110.5 Mb，其中，‘HDA149’、‘9704B’ 分别占总数据量的 41.2%和 38.7%，NCBI 数据占 20.1%。拼接的所有 Unigenes 平均长度为 450 bp，较大的 Unigenes (> 500 bp) 平均长度为 820 bp，而最大的拼接序列达到了 4 030 bp。经 Samtools SNP 软件分析，共检测到 18 159 个 SNP 位点，其中 1 384 个变异位点超过两个核苷酸变异，约占 13.1%。4 种碱基变异频率以 T/C、G/A 最高，分别占 30.3%、23.5%，而 A/T、G/C、G/T、A/C 变异频率分别为 13.6%、11.5%、10.8%、10.3%。

表 1 ‘HDA149’ 与 ‘9704B’ 辣椒转录组测序数据
Table 1 Data summary of ‘HDA149’ and ‘9704B’ transcriptome sequencing

高通量测序数据 Illumina sequence terms	原始序列读数 Raw reads	总碱基数 Total nucleotides	重叠群 Contigs	非重复序列 Unigenes
HDA149	45 360 970	4 092 579 900	574 820	103 440
9704B	50 459 344	4 541 340 960	552 794	97 613

2.2 EST-SNP 序列功能分析

以 NCBI 数据库基因注释为参考对 18 159 个含有的 EST-SNP 序列进行 BLAST 比对，共有 5 483 条 (30.1%， $E < -10$) EST 序列被匹配，其中有 1 781 个 SNP 位点匹配到 1 291 个注释基因上，并有 853 个基因被 GO (Gene Ontology) 分类 (图 1)，功能分析发现其主要参与初生代谢 (28.7%)、细胞代谢 (17.3%)、生物合成过程 (15.7%)。

另外一部分 EST-SNP 功能还涉及各种生物 (非生物) 胁迫 (刺激) 反应，信号转导、次生代谢以及一些特异的细胞过程，如蛋白质折叠修饰等。在 KEGG 分析中，有 125 个基因序列 (9.7%) 参与新陈代谢途径，53 条序列 (4.1%) 参与次生代谢产物合成途径，另外还有 31 条序列 (2.4%) 参与植物激素合成途径。

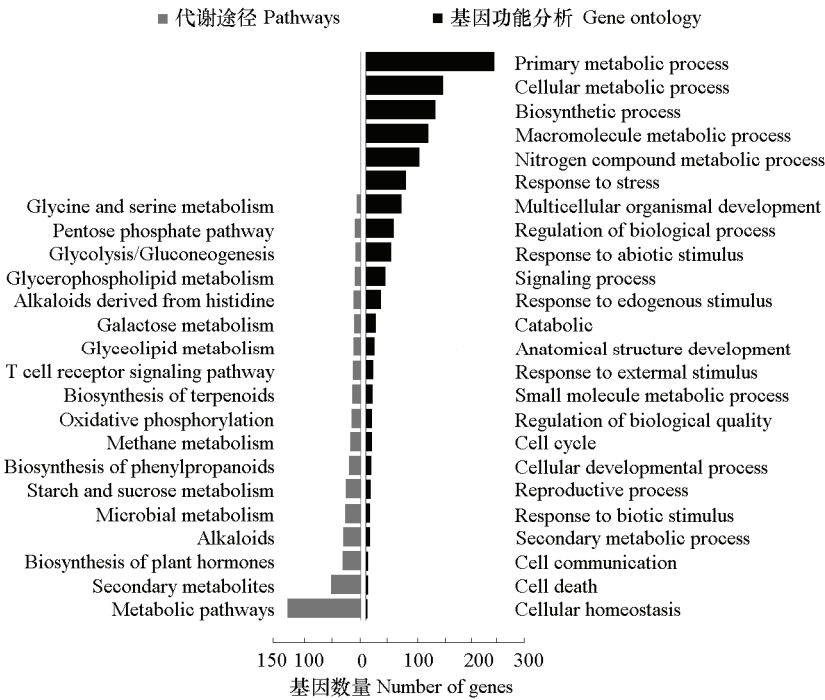


图 1 辣椒 EST-SNP GO 分析及代谢途径
Fig. 1 Gene Ontology and pathways of EST SNP in pepper

2.3 EST-SNP CAPS 标记引物设计及多态性分析

利用程序共获得 EST-SNP 酶切位点 7 034 个, 约占总 EST-SNP 的 38.7%, 其中可以直接酶切 EST-SNP 位点 1 183 个, 约占 6.5%, 经 EST-SNP 位点两边碱基转换后可酶切的 SNP 位点 5 851 个约占 32.2%, 其中能满足设计 CAPS 引物条件的 EST-SNP 序列为 4 172 条, 约占总 EST-SNP 的 22.9%。为了进一步验证 CAPS 引物的有效性以及在不同材料中的多态性, 共选取了 15 对 EST-SNP CAPS 引物对 5 份辣椒材料进行扩增。电泳检测表明有 12 对引物具有单一条带的扩增产物, 随后扩增产物经酶切检测发现其中 8 对引物表现出多态性 (表 2, 图 2)。

基因注释表明这些多态性 EST-SNP 涉及到植物的生长发育和抗逆等过程。为了进一步验证 CAPS

表 2 EST-SNP CAPS 引物及基因注释
Table 2 Annotation and CAPS marker of EST-SNP

编号 ID	基因 ID Unigene ID	引物 Primers	内切酶 Enzyme	基因注释 Annotation
I	Unigene17576	GATTCCCATATTCAGGAACGA GAAGGCTTGAGCTTCACTGTGT	HinfIII	乙烯受体 Ethylene receptor
II	Unigene19983	ATTATTTCAAATTGTTCAAGTCGA TTCTTCACTACTCCTACAGGCTTG	HinfIII	ATP 合成酶亚基 ATP synthase subunit
III	Unigene21587	GCGGTGAACAGTTACAAAAGGA ACCTTAGTCTTCCCGAAGTGGTC	BamH I	受光调控的短下胚轴发育蛋白 LSH10
IV	Unigene23269	GCTAATCCAGAGGATCAAGTCGA ATGCACCTCTCTTACCTTTTCGC	Sal I	热激蛋白 Heat shock protein
V	Unigene25494	TACATCGGGGCTGTGACACTGCA GTAACCCATTGTCCCTCTTGCTGC	Pst I	胁迫诱导受体 Stress-induced receptor
VI	Unigene25924	GCATGATCAGGTGCTTCATCGAT GAACAAAGCACTTGAAAGTGGG	Pvu I	盐诱导蛋白 Salt-inducible protein
VII	Unigene28451	GTGGTTCATGAGTTCAGACTGCA AGTAGCTGCTCTTCTGTCTCTCC	Pst I	PPR 重复蛋白 PPR repeat protein
VIII	Unigene32387	GATCCAAATAGAAGCCGTCATA CTAGATTCAAGCTTGCCGTGGT	Nde I	NBS 抗病蛋白 NBS resistance protein

引物的正确性, 对部分 EST-SNP 酶切后 DNA 带型为野生型 (Wt)、突变型 (Hm)、杂合型 (Hz) 的扩增产物进行了测序。其中图 3 为 Unigene28451 CAPS 引物扩增 ‘9704B’、‘HDA149’、‘CM334’ 材料获得的 3 种基因型产物测序峰图, 测序分型结果与扩增产物酶切分型结果相吻合 (图 2, VII), 表明本研究筛选出 EST-SNP 及其转化的 CAPS 标记能准确地进行基因分型。

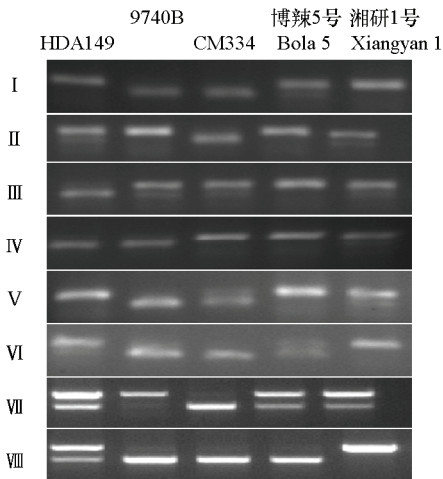


图 2 不同 CAPS 引物在辣椒材料中的扩增酶切结果

Fig. 2 Results of enzyme digestion of PCR product in different peppers

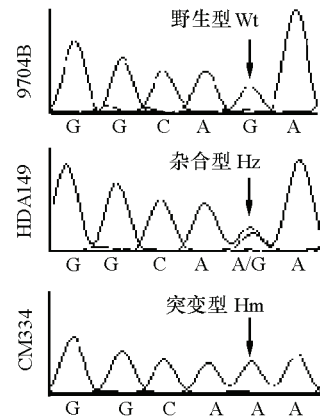


图 3 CAPS 引物 VII (Unigene28451) SNP 位点测序结果

Fig. 3 Sequence results of CAPS marker in different genotypes

3 讨论

本研究中从辣椒转录组和 EST 公共数据库产生的 251 068 条 Unigenes 中共鉴定出 18 159 个 SNP, 总覆盖长度为 110.5 Mbp, SNP 频率为 1/5.8 kb, 与 YCM334/Taeann 辣椒之间的 SNP 频率 (1/6.2 kb) (Lu et al., 2011) 相当, 但比其他物种的 SNP 频率低 (Lijavetzky et al., 2007), 这种频率差异主要与研究材料的遗传背景差异有关, 当 SNP 频率越高表明遗传背景差异越大 (van Tassell et al., 2008)。本研究中 SNP 位点碱基变异类型以 T/C 最高占 30.3%, 与人、大豆、玉米、大麦、小麦等物种中的 SNP 碱基变异类型相似 (Huang & Madan, 1999; Chao et al., 2008; Sato et al., 2011)。

对 18 159 条含有的 EST-SNP 序列进行 NCBI 数据库比对, 共有 5 483 条 EST 序列被匹配约占 30%, 还有大约 2/3 的序列未被匹配, 主要源于数据库中没有辣椒转录组同源性序列或序列过短因匹配度过低而被过滤, 其它有关辣椒转录组研究也得到了类似的结果 (Lu et al., 2011; 刘峰 等, 2012)。筛选的 SNP 中有 1 781 个位点被注释到 1 291 个基因上, 未被注释的序列多为未知功能基因。经 GO 分类和 KEGG 代谢途径分析发现大部分参与初生代谢、细胞代谢、生物合成过程, 以及其它各种生物 (非生物) 胁迫 (刺激) 反应, 信号转导、次生代谢和植物激素合成等, 表明这些 EST-SNP 基因广泛地参与辣椒各种代谢反应过程, 并影响着辣椒各种农艺性状的建成, 可能正是因为功能基因 SNP 变异的丰富性导致其种类形态的多态性。

目前 SNP 作为分子标记发展出了多种 SNP 检测方法: Taqman 法、质谱法、芯片法、测序法、酶切法等 (Kim & Misra, 2007)。这些检测方法各有优缺点, 在分子育种领域如何高效、简单、快速的进行 SNP 检测显得十分重要。在辣椒 SNP 检测中 Jung 等 (2010) 选用了 AS-PCR 法进行基因分型, 该方法虽然简单易行, 但稳定性欠佳, 易产生假阳性结果。因此, 本研究中以 PCR 扩增、限制性内切酶酶切和琼脂糖电泳为检测手段的 CAPS (dCAPS) (Thiel et al., 2004) 方法对 SNP 检测。本研究能满足 CAPS 引物条件的 EST-SNP 序列有 4 172 条 (22.9%), 大部分序列未被成功设

计成 CAPS 标记引物, 主要因为 EST 序列过短或无法满足引物设计预设条件而被过滤掉, 当研究需要时可以适当放宽引物设计条件以达到研究目的。测试的 15 对 CAPS (dCAPS) 引物中有 12 引物能在 5 种不同辣椒材料中进行有效扩增, 并有 8 对引物表现出多态性, CAPS 标记转化成功率为 53.3%。表明本研究中开发的 SNP 转化为 CAPS 标记后可以相对稳定、经济地在辣椒分子遗传育种、基因定位等研究中应用。随着辣椒基因组、转录组测序研究的不断深入, 辣椒 SNP 资源将更加丰富, 对辣椒遗传育种、遗传图谱构建等研究具有更积极的促进作用。

References

- Blair M W, Cortes A J, Penmetsa R. V, Farmer A, Carrasquilla-Garcia N, Cook D R. 2013. A high-throughput SNP marker system for parental polymorphism screening, and diversity analysis in common bean (*Phaseolus vulgaris* L.). *Theor Appl Genet*, 126 (2): 535 – 548.
- Chao S, Zhang W, Akhunov E, Sherman J, Ma Y, Luo M C, Dubcovsky J. 2008. Analysis of gene-derived SNP marker polymorphism in US wheat (*Triticum aestivum* L.) cultivars. *Molecular Breeding*, 23 (1): 23 – 33.
- Frascaroli E, Schrag T A, Melchinger A E. 2013. Genetic diversity analysis of elite European maize (*Zea mays* L.) inbred lines using AFLP, SSR, and SNP markers reveals ascertainment bias for a subset of SNPs. *Theor Appl Genet*, 126 (1): 133 – 141.
- Ganal M W, Altmann T, Roder M S. 2009. SNP identification in crop plants. *Curr Opin Plant Biol*, 12 (2): 211 – 217.
- Gunderson K L, Steemers F J, Lee G, Mendoza L G, Chee M S. 2005. A genome-wide scalable SNP genotyping assay using microarray technology. *Nat Genet*, 37 (5): 549 – 554.
- Huang X, Madan A. 1999. CAP3: A DNA sequence assembly program. *Genome Research*, 9 (9): 868 – 877.
- Jung J K, Park S W, Liu W Y, Kang B C. 2010. Discovery of single nucleotide polymorphism in *Capsicum* and SNP markers for cultivar identification. *Euphytica*, 175 (1): 91 – 107.
- Karchin R. 2009. Next generation tools for the annotation of human SNPs. *Briefings in Bioinformatics*, 10 (1): 35 – 52.
- Kim S, Misra A. 2007. SNP genotyping: Technologies and biomedical applications. *Annu Rev Biomed Eng*, 9: 289 – 320.
- Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K. 2009. SNP detection for massively parallel whole-genome resequencing. *Genome Research*, 19 (6): 1124 – 1132.
- Lijavetzky D, Cabezas J A, Ibáñez A, Rodríguez V, Martínez-Zapater J M. 2007. High throughput SNP discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a re-sequencing approach and SNPlex technology. *BMC Genomics*, 8 (1): 424.
- Liu Feng, Wang Yun-sheng, Tian Xue-liang, Mao Zhen-chuan, Zou Xue-xiao, Xie Bing-yan. 2012. SSR mining in pepper (*Capsicum annuum* L.) transcriptome and the polymorphism analysis. *Acta Horticulturae Sinica*, 39 (1): 168 – 174. (in Chinese)
- 刘 峰, 王运生, 田雪亮, 茆振川, 邹学校, 谢丙炎. 2012. 辣椒转录组 SSR 挖掘及其多态性分析. *园艺学报*, 39 (1): 168 – 174.
- Lu F H, Yoon M Y, Cho Y I, Chung J W, Kim K T, Cho M C, Cheong S R, Park Y J. 2011. Transcriptome analysis and SNP/SSR marker information of red pepper variety YCM334 and Taeon. *Scientia Horticulturae*, 129 (1): 38 – 45.
- Nasu S, Suzuki J, Ohta R, Hasegawa K, Yui R, Kitazawa N, Monna L, Minobe Y. 2002. Search for and analysis of single nucleotide polymorphisms (SNPs) in rice (*Oryza sativa*, *Oryza rufipogon*) and establishment of SNP markers. *DNA Research*, 9 (5): 163 – 171.
- Nicolai M, Pisani C, Bouchet J, Vuylsteke M, Palloix A. 2012. Discovery of a large set of SNP and SSR genetic markers by high-throughput sequencing of pepper (*Capsicum annuum*). *Genetics and Molecular Research*, 11 (3): 2295 – 2300.
- Sato K, Close T J, Bhat P, Munoz-Amatriain M, Muehlbauer G J. 2011. Single nucleotide polymorphism mapping and alignment of recombinant chromosome substitution lines in barley. *Plant Cell Physiol*, 52 (5): 728 – 737.
- Syvanen A C. 2005. Toward genome-wide SNP genotyping. *Nat Genet*, 37 (Suppl): 5 – 10.
- Tenaillon M I, Sawkins M C, Long, A D, Gaut R L, Doebley J F, Gaut B S. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proceedings of the National Academy of Sciences*, 98 (16): 9161 – 9166.
- Thiel T, Kota R, Grosse I, Stein N, Graner A. 2004. SNP2CAPS: A SNP and INDEL analysis tool for CAPS marker development. *Nucleic Acids Research*, 32 (1): 5.
- van Tassel, Smith C P T P, Matukumalli L K, Taylor J F, Schnabel R D, Lawley C T, Haudenschild C D, Moore S S, Warren W C, Sonstegard T S. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Methods*, 5 (3): 247 – 252.