

紫贻贝 EST-SNP 的筛选及多态性检测

李宏俊¹, 梁瑜², 邢坤³, 苏浩¹, 高祥刚¹, 隋立军¹, 赫崇波^{1*}

(1. 辽宁省海洋水产科学研究院, 辽宁省海洋水产分子生物学重点实验室, 辽宁 大连 116023;

2. 大连海洋大学生命科学与技术学院, 辽宁 大连 116023;

3. 大连海洋大学海洋工程学院, 辽宁 大连 116023)

摘要: 利用 EST 数据库开发 SNP 是筛选 SNP 标记的重要途径。本研究利用 EST 数据库开发紫贻贝的 SNP 标记, 利用 QualitySNP 软件对紫贻贝已有的 19 709 条 EST 序列中含有 4 条以上同源序列的重叠群进行分析, 在含有 4 条以上同源序列的 963 个重叠群中, 筛选得到候选 SNP 位点 4 833 个, SNP 的平均频率为 129.1 bp, 其中 C/T 和 A/G 突变较多, 分别占总数的 28.8% 和 27.4%。根据候选 SNP 位点设计 30 组引物, 通过等位基因特异性 PCR 结合溶解曲线分析, 在 30 个野生个体中进行基因分型验证, 6 组引物 (20%) 没有扩增产物, 10 组引物 (33%) 没有多态性, 14 组 (47%) 引物具有二等位基因多态性, 稀有等位基因的频率为 0.083 ~ 0.446, 观测杂合度和期望杂合度的范围分别为 0.166 7 ~ 0.615 4 和 0.155 4 ~ 0.503 2。序列比对结果显示, 14 组引物中的 12 个 SNP 位点位于基因编码区, 并且全部为同义突变。研究结果表明, EST 数据库中存在大量的 SNP 位点, 差异溶解曲线法是一种高效便捷的 SNP 分型方法, 所筛选的 SNP 标记可用于紫贻贝遗传学分析。

关键词: 紫贻贝; 单核苷酸多态性; 表达序列标签; 等位基因特异性 PCR; 溶解曲线

中图分类号: Q 347; S 917

文献标识码: A

单核苷酸多态性 (single nucleotide polymorphism, SNP) 是第三代 DNA 分子标记, 指在基因组水平上由单个核苷酸变异引起的 DNA 序列多态性^[1-2]。研究表明, 人类基因组中每 1000 个碱基对存在 1 个 SNP 位点^[3], 在文蛤 (*Meretrix meretrix*) 和海湾扇贝 (*Argopecten irradians*) 中分别达到 156 bp 和 118 bp 存在 1 个 SNP 位点^[4-5], 长牡蛎 (*Crassostrea gigas*) 基因组中非编码区 SNP 的密度甚至达到 40 bp^[6]。SNP 突变包括单碱基的转换、颠换、插入或缺失等, 其中 C/T (G/A) 转换最常见, 其余三种颠换为 C/A (G/T), C/G (G/C) 和 T/A (A/T), 插入和缺失突变较少, 并且 SNP 通常表现为二等位基因多态性, 三等位或者四等位基因的 SNP 非常稀少。由于数量多、密度大、二等位基因多态、可以引起氨基酸突变和易于高通量分析, SNP 标记被广泛地

应用于遗传连锁图谱的构建、基因定位、遗传多样性研究、品种鉴定和动植物育种等领域^[7-10]。

SNP 标记的开发通常需要 SNP 位点的获得和基因分型两个步骤。在缺乏基因组序列的非模式生物中, 在表达序列标签 (expressed sequence tag, EST) 数据库中搜寻 SNP 位点已经成为获得 SNP 位点的重要途径, 目前在水产动物中已经有若干报道^[4-5, 11-13]。SNP 的基因分型方法有很多种, 传统的方法有限制性片段长度多态性 (restriction fragment length polymorphism, RFLP), 单链构象多态性 (single strand conformation polymorphism, SSCP) 和等位基因特异性杂交 (allele-specific hybridization, ASH) 等^[14], 近几年发展起来的 SNP 分型技术包括变性高效液相色谱 (denaturing high performance liquid chromatography, DHPLC)^[15], 基因芯片法^[16], TaqMan 探针法^[17] 和焦磷酸测序技

收稿日期: 2010-11-22 修回日期: 2010-12-24

资助项目: 辽宁省科学技术计划重大、重点项目 (2008203001); 辽宁省海洋与渔业科研计划项目 (200801); 国家海洋公益性行业科技专项 (200805037)

通讯作者: 赫崇波, E-mail: hechongbo@hotmail.com

术^[18]等。传统方法操作步骤繁琐、通量低,新技术往往需要先进的仪器和巨大的投资。本研究利用引物特异性 PCR 结合熔解曲线(melting curve, T_m)分析,探讨一种高效的 SNP 分型方法。

紫贻贝(*Mytilus galloprovincialis*),俗称壳菜、海虹,自然分布于我国辽宁、山东、浙江等省沿海,因其繁殖力强、生长速度快、环境适应力强和营养价值高^[19],是世界主要贝类养殖品种之一。早在 20 世纪 70 年代,我国紫贻贝的全人工育苗就获得成功,随着对紫贻贝遗传多样性的调查检测、遗传选育和改良以及遗传图谱构建等研究工作的不断深入,对紫贻贝分子标记的需求越来越迫切。目前,已报道的紫贻贝分子标记仅有 17 对微卫星标记^[20-22],相比微卫星,SNP 标记具有稳定性高和易于自动化分析等优点。为配合紫贻贝遗传学工作的开展,从 EST 数据库中筛选出 14 组紫贻贝 SNP 标记,并估算了这些位点的主要遗传参数。

1 材料与方法

1.1 实验材料及基因组 DNA 的提取

实验所用活体样本于 2010 年 5 月采自大连近海,共 30 只,活体解剖,取肌肉组织分装到样品袋中,液氮冷冻,保存于 $-80\text{ }^{\circ}\text{C}$ 冰箱中备用。利用常规“酚-氯仿”的方法^[23]提取贻贝基因组

DNA,紫外分光光度计定量, $-20\text{ }^{\circ}\text{C}$ 保存备用。

1.2 EST-SNP 位点的查找和引物设计

采用软件 Cap 3^[24]对 NCBI 注册的 19 709 条紫贻贝 EST 进行序列拼接,利用软件 QualitySNP^[25]在含有 4 条以上 EST 的拼接序列(contig)中搜寻 SNP 位点,在重叠区最少出现两次的 SNP 作为候选位点进行引物设计。利用软件 Primer Premier 5.0 (Premier Biosoft International)设计引物, T_m -shift 引物设计原则参照 WANG 等^[26]的方法。对于任意 1 个 SNP 位点,设计 2 条正向等位基因特异性引物和 1 条反向引物,正向引物 3'末端碱基对应 SNP 的 2 个等位基因(图 1)。正向引物的退火温度一般在 $58.3\text{ }^{\circ}\text{C}\sim 63.5\text{ }^{\circ}\text{C}$ (最适 $59\text{ }^{\circ}\text{C}\sim 62\text{ }^{\circ}\text{C}$),长度在 $14\text{ }^{\circ}\text{C}\sim 27\text{ }^{\circ}\text{C}$ (最适 $15\text{ }^{\circ}\text{C}\sim 22\text{ }^{\circ}\text{C}$),反向引物的退火温度一般在 $62\text{ }^{\circ}\text{C}\sim 75\text{ }^{\circ}\text{C}$ (最适 $63\text{ }^{\circ}\text{C}\sim 70\text{ }^{\circ}\text{C}$),长度在 $22\text{ }^{\circ}\text{C}\sim 30\text{ }^{\circ}\text{C}$ (最适 $22\text{ }^{\circ}\text{C}\sim 27\text{ }^{\circ}\text{C}$),反向引物 3'末端一般位于正向引物下游 20 bp 以内。两条长度不同的 GC 尾巴分别加到等位基因特异引物的 5'端,为了增加熔解温度差异,长 GC 尾巴加到具有高熔解温度 3'端碱基(G 或 C)的引物,短 GC 尾巴加到低熔解温度 3'端碱基(A 或 T)的引物。为了增加等位基因引物的特异性,在正向引物的 3'末端加入 1 个人为错配碱基^[27]。

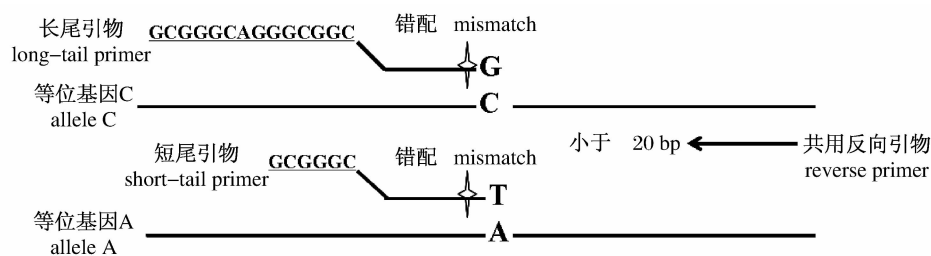


图 1 T_m -shift 法引物设计示意图

Fig. 1 Schematic explanation of the primer design for T_m -shift method

1.3 PCR 扩增和熔解曲线分析

PCR 反应总体积为 $20\text{ }\mu\text{L}$,内含 $10\text{ }\mu\text{L}$ 的 $2\times$ SYBR[®] Premix Ex Taq[™] (Takara), 3 条引物各 $0.3\text{ }\mu\text{L}$ ($10\text{ }\mu\text{mol/L}$) 和 50 ng 的模板 DNA。PCR 反应在 Mx 3000P (Stratagene) 荧光定量扩增仪上进行。PCR 反应为 40 个循环,每个循环包括: $94\text{ }^{\circ}\text{C}$ 变性 15 s, $60\text{ }^{\circ}\text{C}$ 退火 30 s, $72\text{ }^{\circ}\text{C}$ 延伸 20 s;首次循环前变性 3 min;循环结束后进行 $70\text{ }^{\circ}\text{C}\sim 95\text{ }^{\circ}\text{C}$ 熔解曲线分析。利用每个个体 PCR 产物熔解曲

线峰图确定其基因型,利用软件 GENEPOP 4.0^[28]进行群体遗传分析,计算观测杂合度(observed heterozygosity, H_o) 和期望杂合度(expected heterozygosity, H_e),利用 U 检验进行 Hardy-Weinberg 平衡验证。

利用 NCBI 的 BLASTX 程序,对含有 SNP 的 EST 序列进行功能预测,匹配序列同源性达到期望值 $< 1e^{-10}$ 时判定为同源序列。利用 ORF finder^[29]推测 EST 正确的开发阅读框,确定 SNP

在遗传密码子的位置,与标准密码子比对确定该突变是否引起氨基酸改变。

2 结果

2.1 紫贻贝 EST 数据库中含有 SNP 的数量和特征

通过 GenBank 搜索,目前紫贻贝共有 19 709 条 EST 序列。经过软件 Cap 3 序列比对后得到单一序列 5 994 条和重叠群 2 486 个(表 1),含有 4 条以上同源序列的重叠群共有 963 个,利用软件 QualitySNP 在这些重叠群中共发现候选 SNP 位点 4 833 个,其中包括转换 2 714 个(C/T,1 391;A/G,1 323)和颠换 1 748 个(A/T,655;A/C,550;T/G,335;C/G,208),C/T 和 A/G 突变较多,分别占总数的 28.8% 和 27.4%,转换与颠换的比值为 1.6:1,插入突变 295 个,三态和四态的 SNP 位点 76 个,紫贻贝 EST 序列平均每 129.2 bp 存在 1 个 SNP 位点。

2.2 候选 SNP 位点的基因分型

利用 QualitySNP 在含有 4 条以上序列的重叠群中共发现 4 833 个 SNP 位点,根据 SNP 位点的分布频率和侧翼序列质量,利用 Primer Premier 5.0 设计引物 30 组,其中 6 组(20%)引物没有扩增产物,10(33%)组引物没有多态性,14 组(47%)SNP 标记在 30 个个体中显示多态性(表 2),SNP 引物 *MgSNP200* 扩增产物的代表性溶解曲线见图 2。所检测的 14 组 SNP 均呈现二等位基因多态性,稀有等位基因的频率为 0.083 ~ 0.446,观测杂合度和期望杂合度的范围分别为

0.166 7 ~ 0.615 4 和 0.155 4 ~ 0.503 2,经过 *U* 检验,除 *MgSNP53*、*MgSNP235* 和 *MgSNP662*,其余各位点均符合哈迪-温伯格平衡($P > 0.05$)。

表 1 紫贻贝 EST 数据库中 SNP 数量和特征
Tab.1 Summary of EST-SNPs from *M. galloprovincialis*

参数 parameter	数值 value
EST 总数 number of ESTs	19 709
单一序列 singletons	5 994
重叠群 contigs	2 486
2 ~ 3 ESTs	1 523
4 ~ 5 ESTs	416
6 ~ 10 ESTs	316
11 ~ 20 ESTs	151
>20 ESTs	97
含有 SNP 重叠群 SNP-containing contigs	717
SNP 总数 number of total SNPs	4 833
转换 transition	2 714
C↔T	1 391
A↔G	1 323
颠换 transversion	1 748
A↔T	655
A↔C	550
T↔G	335
C↔G	208
插入突变 indel	295
其他类型* other types	76
SNP 频率 SNP frequency	129.2 bp

注: * 其他类型包括三等位基因和四等位基因突变。

Notes: * other types include tri- and tetra-allelic mutations.

表 2 紫贻贝的单核苷酸多态性标记
Tab.2 Single nucleotide polymorphism (SNP) markers for *M. galloprovincialis*

位点 locus	引物 primer(5'-3')	变异类型 mutation type	次要等位基因频率 minor allele frequency	H_o	H_e	P_{HW}
<i>MgSNP53</i>	AS1: <u>gcgggcaggcggc</u> AATAAAGTACGACGTGGTTGAC [C]GC AS2: <u>gcgggc</u> AATAAAGTACGACGTGGTTGACA [T]T CR: AGGTGCTGTGAACTGACCAGATGAT	C/T	T 0.250	0.214 3	0.381 8	0.016 6*
<i>MgSNP164</i>	AS1: <u>gcgggcaggcggc</u> TGATTGGGAAGATGTTTCGAC [G]C AS2: <u>gcgggc</u> TGATTGGGAAGATGTTTCGA [T]AT CR: GGTACACTGGATCGAGTCGTAGC	C/T	0.100	0.200 0	0.183 1	0.581 3
<i>MgSNP200</i>	AS1: <u>gcgggcaggcggc</u> AGGTGGAGATTTACCAGAGGAGGC AS2: <u>gcgggc</u> AGGTGGAGATTTACCAGAGGATAT CR: AGTTCTCGTCCCTCAAATTTATTGCC	C/T	C 0.433	0.466 7	0.499 4	0.714 5

· 续表 2 ·

位点 locus	引物 primer(5'-3')	变异类型 mutation type	次要等位 基因频率 minor allele frequency	H_o	H_e	P_{HW}
<i>MgSNP235</i>	AS1: <u>gcgggcagggcggc</u> CAATGTCGCTAGTGAACACGCCA C G AS2: <u>gcgggc</u> CAATGTCGCTAGTGAACACGC T AAA CR: CGATTTGGTTGAAACTTTTTCACAG	G/A	A 0.268	0.178 6	0.399 4	0.002 6*
<i>MgSNP283</i>	AS1: <u>gcgggcagggcggc</u> CATTCCAGAAAATACCAGC G CCC AS2: <u>gcgggc</u> CATTCCAGAAAATACCAGCA T CA CR: TGGCAGTTTGATCGTTTAAAAGTTG	C/A	C 0.393	0.571 4	0.485 7	0.340 7
<i>MgSNP347</i>	AS1: <u>gcgggcagggcggc</u> TGAATCATTTCAGAGAATGAT G AC AS2: <u>gcgggc</u> TGAATCATTTCAGAGAATGAAT A T CR: GAACTTTGTCACGGATGTTTGGTA	C/T	T 0.125	0.250 0	0.222 7	0.487 4
<i>MgSNP410</i>	AS1: <u>gcgggcagggcggc</u> ACAATGGCGGTCCAAGAGGA C GC AS2: <u>gcgggc</u> ACAATGGCGGTCCAAGAGGAT A T CR: TGTTCAGTATCCAATGGTGTAG	C/T	T 0.350	0.433 3	0.462 7	0.722 6
<i>MgSNP662</i>	AS1: <u>gcgggcagggcggc</u> GGAACCAGTACAAGGAAGTAAA G AC AS2: <u>gcgggc</u> GGAACCAGTACAAGGAAGTAAATAT CR: AGTATCTTCTACCACGGTAACGACG	C/T	C 0.296	0.615 4	0.434 4	0.029 0*
<i>MGSNP671</i>	AS1: <u>gcgggcagggcggc</u> GAATTACAAAATAAGTACCAA C GTG AS2: <u>gcgggc</u> GAATTACAAAATAAGTACCAAT A TT CR: CATGTTTACACGTGTGTTTCTATTG	G/T	T 0.205	0.227 3	0.333 0	0.117 4
<i>MGSNP672</i>	AS1: <u>gcgggcagggcggc</u> TTGCTGTCCGAGACATGAG G CAG AS2: <u>gcgggc</u> TTGCTGTCCGAGACATGAGA T AA CR: CCTTTGTAACTTCTTTGATGACTCC	G/A	T 0.183	0.233 3	0.304 5	0.182 2
<i>MGSNP907</i>	AS1: <u>gcgggcagggcggc</u> TTCCCATCCAAGTCTCTCATA G C AS2: <u>gcgggc</u> TTCCCATCCAAGTCTCTCATAAT CR: GTTGCTTTCTGTTCTCACAACCTTG	C/T	T 0.100	0.200 0	0.183 1	0.581 3
<i>MGSNP1224</i>	AS1: <u>gcgggcagggcggc</u> GGGTTCAATATGCAATGACTGG C TC AS2: <u>gcgggc</u> GGGTTCAATATGCAATGACTG T ATT CR: CATCTGACAAATAGCACATGGTTCA	C/T	T 0.411	0.392 9	0.492 9	0.273 6
<i>MGSNP1506</i>	AS1: <u>gcgggcagggcggc</u> CAAGAAGGAAATCAAGGACATG C TG AS2: <u>gcgggc</u> CAAGAAGGAAATCAAGGACAT A TTA CR: CAGCTACCAATAATGTCCCGTCATA	G/A	A 0.446	0.464 3	0.503 2	0.676 4
<i>MGSNP2016</i>	AS1: <u>gcgggcagggcggc</u> GGCGAGATTATAAAGAGGGA G TC AS2: <u>gcgggc</u> GGCGAGATTATAAAGAGG T ATTT CR: CATTACCAAGCCAATGTTACCCCTG	C/T	T 0.268	0.392 9	0.399 4	0.929 4
<i>MGSNP2210</i>	AS1: <u>gcgggcagggcggc</u> CACCCACCATGCCAGGGAATGG C CG AS2: <u>gcgggc</u> CACCCACCATGCCAGGGAATGG T AA CR: CCACTTTTATCCGCCAGTGAAGTC	G/A	A 0.083	0.166 7	0.155 4	0.658 5

注:单下划线表示 GC 尾巴;方框表示人为错配碱基;AS1 and AS2:等位基因特异引物;CR:反向公用引物; H_o :观测杂合度; H_e :期望杂合度; P_{HW} :哈迪-温伯格检验(*代表 $P < 0.05$)。

Notes: Tail sequences are underlined and mismatches are boxed; AS1 and AS2: allele-specific primers, CR: common reverse primer; H_o : observed heterozygosity, H_e : expected heterozygosity, P_{HW} values represent significant values for exact tests of deviation from Hardy-Weinberg equilibrium(* represents $P < 0.05$).

2.3 SNP 引物的功能预测

利用 NCBI 网站的程序对含有 14 个 SNP 的重叠序列进行功能预测, 比对结果显示 14 条重叠序列均找到有效匹配(表 3), 涉及到的基因主要包括外套膜基因 4、(mantle gene 4) 亲环素 A(cyclophilin A)、

溶菌酶(lysozyme) 和小热激蛋白(small heat shock protein) 等。利用 ORF finder 确定正确的开放阅读框, 结果显示位点 *MGSNP671* 和 *MGSNP2210* 位于非编码区, 其余位点均位于密码子的第三位, 与标准密码子表比对后确认均为同义突变。

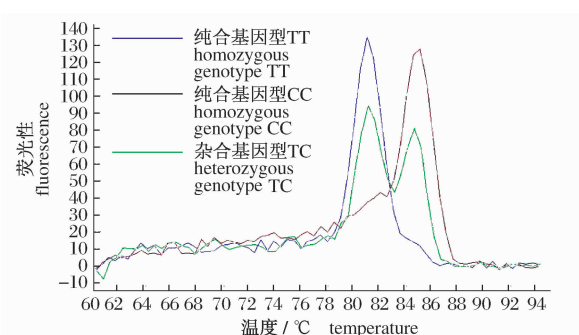


图2 *MgSNP200* 扩增产物的熔解曲线

Fig. 2 Melting curves of *MgSNP200* amplified products

表3 紫贻贝单核苷酸标记引起的同义突变和功能注释

Tab.3 Synonymous codon substitutions and gene functions related to the detected SNP loci from *M. galloprovincialis*

位点 locus	氨基酸 amino acid	GenBank 索引号 accession no.	基因功能 gene function
<i>MgSNP53</i>	Leu: CTC→CTT	EH662762	mantle gene 4
<i>MgSNP200</i>	Asp: GAT→GAC	FL593736	cyclophilin A
<i>MgSNP235</i>	Lys: AAG→AAA	FL497812	inhibitor of apoptosis 2
<i>MgSNP283</i>	Pro: CCC→CCA	FL500048	cathepsin L-associated protein
<i>MgSNP347</i>	Pro: TAT→TAC	FL499261	metalloproteinase inhibitor 3
<i>MgSNP410</i>	Cys: TGC→TGT	EH662817	lysozyme
<i>MgSNP662</i>	Tyr: TAT→TAC	FL496487	nuclease-sensitive element-binding protein
<i>MGSNP671</i>	-	EH662488	myticin c precursor
<i>MGSNP672</i>	Gln: CAA→CAG	EH663443	elongation factor 1 alpha
<i>MGSNP907</i>	Asn: AAC→AAT	EH662510	60s ribosomal
<i>MGSNP1224</i>	Leu: CTT→CTC	EH663211	placenta-specific gene 8
<i>MGSNP1506</i>	Leu: TTA→TTG	FL492272	ribosomal protein s16
<i>MGSNP2016</i>	Phe: TTC→TTT	EH662805	fibrinogen c domain-containing protein 1
<i>MGSNP2210</i>	-	FL496056	small heat shock protein

注: - 表示非编码区的 SNP。

Notes: - indicates non-coding SNP.

3 讨论

单核苷酸多态性是动植物基因组中最常见的遗传变异,候选 SNP 的搜寻主要有两种策略:第一种是利用包含大量冗余序列的 EST 数据库;第

二种是依赖于基因组测序中的重叠区域^[30]。本研究利用 EST 文库搜索的方法,在紫贻贝中发现 4 833 个候选 SNP 位点,利用等位基因 PCR 结合溶解曲线分析,成功获得 14 组紫贻贝 SNP 标记。由于 EST 序列来源于表达基因,所以通过上述方法获得的 SNP 直接和功能基因相关,适用于与特定表型进行关联分析。

在本研究中,SNP 在紫贻贝 EST 序列中的分布频率较高,达到 129.2 bp,由于基因组的变异高于 EST 序列,可以肯定贻贝基因组中 SNP 的密度更高。SNP 的分布频率在各个物种中差别很大,人类基因组大约每隔 1 kb 存在 1 个 SNP 位点^[31],玉米基因组中 104 bp 分布 1 个 SNP^[31],鳕 EST 重叠群每 100 bp 存在 0.79 个 SNP^[32],在贝类中,海湾扇贝 EST 重叠群每 118.2 bp 存在 1 个 SNP^[4],长牡蛎编码区和非编码区 SNP 的频率分别达到 60 和 40 bp^[6]。紫贻贝的 SNP 频率与海湾扇贝和牡蛎基本一致,表明双壳贝类基因组的 SNP 密度普遍较高。紫贻贝 EST 序列中的 SNP 大部分来源于转换(2 714/4 833),部分来源于颠换(1 748/4 833),插入突变和其他类型突变较少(371/4 833),其中 C/T 突变类型最多,C/G 突变类型最少,与大西洋鲑^[33]和海湾扇贝^[4]的实验结果类似,研究表明 CpG 二核苷酸上的 C 碱基是基因组中最易发生突变的位点,可自发脱去氨基形成 T 碱基^[34]。本研究中位于基因编码区的 12 个 SNP 标记均为同义突变,这与海湾扇贝和鲍的研究结果类似^[4,12],暗示贝类基因组中的同义突变远远高于非同义突变,同义突变虽然不改变编码蛋白的结构与功能,但是可以通过改变基因表达量间接影响生物体表型^[35],紫贻贝同义突变 SNP 的功能有待进一步研究。

本研究从获得的 SNP 位点中选取 30 条等位基因频率分布均匀的序列来设计引物,然后进行 PCR 产物的熔解曲线分析来检测这些引物能否作为 SNP 引物。其中 6 组引物没有扩增产物可能是由于选取的 SNP 引物序列处于跨越内含子的两个外显子的拼接处,或者在转录的修饰区中,这样可能引起一端或两端的引物无法和基因组 DNA 模板结合而使扩增失败。另外 10 组引物有扩增产物,但是没有多态性,说明通过生物信息学分析获得的候选 SNP 位点存在部分假阳性。虽然本实验设置了较高的筛选条件,在含有 4 条以

上同源序列的重叠群中,稀有等位基因出现大于 2 次的 SNP 才被认定为候选位点,但是本研究不能结合原始测序峰图所评估出的测序质量情况进行判断,不能严格筛选不同文库中的 SNP,同时由于 EST 序列通常较短,进行聚类分析时很可能错误地将同源性较高的非等位基因归为一簇^[36],从而增加了假阳性的概率。WANG 等^[32]验证鳃 EST-SNP 标记时认为重叠群包含的同源序列数目和稀有等位基因频率是影响成功率的最重要两点因素,重叠群包含的序列越多、稀有等位基因频率越高,引物开发的成功率越高,因此,开发 EST-SNP 标记时应该尽量选择重叠群大和等位基因频率分布均匀的位点设计引物。

SNP 分型技术是目前国际上的研究热点,本研究验证了一种先进的单管 SNP 基因分型技术,不需要进行传统的限制性内切酶消化,因特异性引物的扩增产物存在熔解曲线差异,利用荧光定量 PCR 仪,可以同时确定多个个体的基因型,极大提高了检测效率,可以解决 SNP 检测的高速度问题。在特异性引物的 3'端引入 1 个人为错配碱基,使特异性大大提高,降低了 SNP 分型的假阳性率^[27]。但熔解曲线差异法每个反应只能检测 1 个 SNP 位点,通量有限,适用于分析位点数目较少的群体遗传学研究,多重 SNP 的检测技术有待于进一步研究。

参考文献:

- [1] PICOULT-NEWBERG L, IDEKER T E, POHL M G, *et al.* Mining SNPs from EST databases [J]. *Genome Res*, 1999, 9(2) : 167 - 174.
- [2] CHO R J, MINDRINOS M, RICHARDS D R, *et al.* Genome-wide mapping with biallelic markers in *Arabidopsis thaliana* [J]. *Nat Genet*, 1999, 23 (2) : 203 - 207.
- [3] FRAZER K A, BALLINGER D G, COX D R, *et al.* A second generation human haplotype map of over 3.1 million SNPs [J]. *Nature*, 2007, 449 : 851 - 861.
- [4] LI R H, LI Q, KONG L F. Characterization of expressed sequence tag-derived single-nucleotide polymorphisms in the bay scallop *Argopecten irradians irradians* [J]. *Fish Sci*, 2009, 75 (6) : 1389 - 1400.
- [5] LI H J, ZHU D, GAO X G, *et al.* Mining single nucleotide polymorphisms from EST data of hard clam *Meretrix meretrix* [J]. *Conserv Genet Res*, 2010, 2(S1) : 69 - 72.
- [6] SAUVAGE C, BIERNE N, LAPEGUE S, *et al.* Single nucleotide polymorphisms and their relationship to codon usage bias in the Pacific oyster *Crassostrea gigas* [J]. *Gene*, 2007, 406(1 - 2) : 13 - 22.
- [7] SMITH C T, TEMPLIN W D, SEEB J E, *et al.* Single nucleotide polymorphisms provide rapid and accurate estimates of the proportions of US and Canadian Chinook salmon caught in Yukon River fisheries [J]. *North Am J Fish Manage*, 2005, 25 : 944 - 953.
- [8] YAMAMOTO M, NARUKAWA J, OKUDA K K, *et al.* Construction of a single nucleotide polymorphism linkage map for the silkworm, *Bombyx mori*, based on bacterial artificial chromosome end sequences [J]. *Genetics*, 2006, 173(1) : 151 - 161.
- [9] RENGMARK A H, SLETTAN A, SKAALA O, *et al.* Genetic variability in wild and farmed Atlantic salmon (*Salmo salar*) strains estimated by SNP and microsatellites [J]. *Aquaculture*, 2006, 253 (1 - 4) : 229 - 237.
- [10] DU Z Q, CIOBANU D C, ONTERU S K, *et al.* A gene-based SNP linkage map for pacific white shrimp, *Litopenaeus vannamei* [J]. *Anim Genet*, 2009, 41(3) : 286 - 294.
- [11] HUBERT S, TARRANT B J, HIGGINS B, *et al.* Development of single nucleotide polymorphism markers for Atlantic cod (*Gadus morhua*) using expressed sequences [J]. *Aquaculture*, 2009, 296(1 - 2) : 7 - 14.
- [12] QI H G, LIU X, ZHANG G F, *et al.* Mining expressed sequences for single nucleotide polymorphisms in Pacific abalone *Haliotis discus hannai* [J]. *Aqua Res*, 2009, 40(14) : 1661 - 1667.
- [13] LI H J, HE C B, YANG Q, *et al.* Characterization of single nucleotide polymorphisms from expressed sequence tags of Chinese mitten crab *Eriocheir sinensis* [J]. *Aquat Biol*, 2010, 11(2) : 193 - 199.
- [14] KWOK P Y. Methods for genotyping single nucleotide polymorphisms [J]. *Annu Rev Genomics Hum Genet*, 2001, 2 : 235 - 258.
- [15] WOLFORD J K, BLUNT D, BALLECER C, *et al.* High-throughput SNP detection by using DNA pooling and denaturing high performance liquid chromatography (DHPLC) [J]. *Hum Genet*, 2000, 107(5) : 483 - 487.
- [16] SCHMALZING D, BELENKY A, NOVOTNY M A, *et al.* Microchip electrophoresis: a method for high-speed SNP detection [J]. *Nucleic Acids Res*, 2000, 28(9) : e43.

- [17] DELA V F M, LAZARUK K D, RHODES M D, *et al.* Assessment of two flexible and compatible SNP genotyping platforms: *TaqMan* SNP Genotyping Assays and the SNPlex Genotyping System [J]. *Mutat Res*, 2005, 573(1-2): 111-135.
- [18] FAKHRAI-RAD H, POURMAND N, RONAGHI M. Pyrosequencing: an accurate detection platform for single nucleotide polymorphisms [J]. *Hum Mutat*, 2002, 19(5): 479-485.
- [19] 苏秀榕, 张健, 李太武, 等. 两种贻贝营养成分的研究[J]. *辽宁师范大学学报: 自然科学版*, 1997, 20(3): 239-243.
- [20] PRESA P, PEREZ M, DIZ A P. Polymorphic microsatellite markers for blue mussels (*Mytilus* spp.) [J]. *Conserv Genet*, 2002, 3(4): 441-443.
- [21] VARELA M A, GONZALEZ-TIZON A, MARINAS L, *et al.* Genetic divergence detected by ISSR markers and characterization of microsatellite regions in *Mytilus* mussels [J]. *Biochem Genet*, 2007, 45(7-8): 565-578.
- [22] YU H, LI Q. Development of EST-SSRs in the Mediterranean blue mussel, *Mytilus galloprovincialis* [J]. *Mol Ecol Notes*, 2007, 7(6): 1308-1310.
- [23] 李宏俊, 刘晓, 杜雪地, 等. 海湾扇贝微卫星标记开发及其分离方式分析[J]. *海洋科学*, 2009, 33(12): 4-8.
- [24] HUANG X Q, MADAN A. CAP3: A DNA sequence assembly program [J]. *Genome Res*, 1999, 9(9): 868-877.
- [25] 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 Bioinformatics*, 2006, 7: 438.
- [26] WANG J, CHUANG K, AHLUWALIA M, *et al.* High-throughput SNP genotyping by single-tube PCR with T_m -shift primers [J]. *Biotechniques*, 2005, 39(6): 885-893.
- [27] GAUDET M, FARA A G, SABATTI M, *et al.* Single-reaction for SNP genotyping on agarose gel by allele-specific PCR in Black Poplar (*Populus nigra* L.) [J]. *Plant Mol Biol Rep*, 2007, 25(1): 1-9.
- [28] FRANCOIS R. GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux [J]. *Mol Ecol Res*, 2008, 8(1): 103-106.
- [29] ROMBEL I T, SYKES K F, RAYNER S, *et al.* ORF-FINDER: a vector for high-throughput gene identification [J]. *Gene*, 2002, 282(1-2): 33-41.
- [30] 陈伟, 张戈, 张思仲. 基于生物信息学的 SNP 候选位点搜寻方法 [J]. *遗传*, 2001, 23(2): 153-156.
- [31] TENAILLON M I, SAWKINS M C, LONG A D, *et al.* Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.) [J]. *Proc Natl Acad Sci USA*, 2001, 98(16): 9161-9166.
- [32] WANG S L, SHA Z X, SONSTEGARD T S, *et al.* Quality assessment parameters for EST-derived SNPs from catfish [J]. *BMC Genomics*, 2008, 9: 450.
- [33] HAYES B, LAERDAHL J K, LIEN S, *et al.* An extensive resource of single nucleotide polymorphism markers associated with Atlantic salmon (*Salmo salar*) expressed sequences [J]. *Aquaculture*, 2007, 265(1-4): 82-90.
- [34] SHI H, MAIER S, NIMMRICH I, *et al.* Oligonucleotide-based microarray for DNA methylation analysis: principles and applications [J]. *J Cell Biochem*, 2003, 88(1): 138-140.
- [35] RICHARD I, BECKMAN J S. How neutral are synonymous codon mutations? [J]. *Nat Genet*, 1995, 10: 259.
- [36] 王绍宗, 李莉, 亓海刚, 等. 长牡蛎 (*Crassostrea gigas*) 17 个 EST-SNP 标记的开发 [J]. *海洋与湖沼*, 2010, 41(2): 274-281.

Development of single nucleotide polymorphism markers for blue mussel (*Mytilus galloprovincialis*) using expressed sequence tags

LI Hong-jun¹, LIANG Yu², XING Kun³, SU Hao¹, GAO Xiang-gang¹, SUI Li-jun¹, HE Chong-bo^{1*}

(1. The Key Laboratory of Marine Fishery Molecular Biology of Liaoning Province,

Liaoning Ocean and Fishery Science Institute, Dalian 116023, China;

2. College of Life Science and Technology, Dalian Ocean University, Dalian 116023, China;

3. College of Marine Engineering, Dalian Ocean University, Dalian 116023, China)

Abstract: Single nucleotide polymorphism (SNP) has very broad prospects in the research fields of population genetics of aquacultural species, molecular marker-assisted breeding and biological evolution. The development of SNP markers is normally obtained through genome sequencing, by sequence comparison. However, for the species for which no large-scale genome sequencing has been carried out, using EST database is often an important way for the development of SNP markers. In this study, the SNP genotyping assays and development of SNP markers for blue mussel (*Mytilus galloprovincialis*) were conducted through expressed sequence tags (ESTs) database mining. Some 19 709 EST sequences of blue mussel from the GenBank were clustered into 2 486 contigs, of which 963 contained four or more ESTs. After manual quality filtering, 4 833 putative SNPs were identified from these SNP-containing contigs. The average putative SNP frequency was one per 129.2 bp of contig sequences. C/T and A/G with high frequency account for 28.8% and 27.4%, respectively. 31 of the putative SNPs were chosen for validation by allele specific PCR with melting temperature (T_m)-shift analysis, and 14 (47%) of them were polymorphic with the minor allele frequency ranging from 0.083 to 0.446. The observed heterozygosity and expected heterozygosity were distributed from 0.1667 to 0.6154 and 0.1554 to 0.5032, respectively. Six primers (20%) amplified no any product and 10 (33%) were monomorphic. BLASTX showed that significant hits for all 14 genotyped SNP-containing contigs, 12 of which were located in coding regions and all resulted in a synonymous substitution. The result of present study shows that ESTs could provide effective means for SNP identification in species with limited genome sequence resources, and T_m -shift analysis is a simple, efficient and reliable SNP genotyping method for non-model organisms.

Key words: *Mytilus galloprovincialis*; single nucleotide polymorphisms (SNP); expressed sequence tag (EST); allele specific PCR; melting curve

Corresponding author: HE Chong-bo. E-mail: hechongbo@hotmail.com