长牡蛎(*Crassostrea gigas*)17 个 EST-SNP 标记的开发*

王绍宗[1, 2] 李 莉[2] 亓海刚[2] 张国范[2①]

(1. 中国科学院研究生院 北京 100049; 2. 中国科学院海洋研究所 青岛 266071)

提要 利用长牡蛎已有的 EST 序列数据库, 筛选得到候选 SNP 位点共计 1140 个。根据候选 SNP 位点共设计引物 82 组, 通过片段长度差异等位基因特异性 PCR (fragment length discrepant allele specific PCR, FLDAS-PCR)的分型方法, 在一野生群体中进行检测和验证, 结果共有 17 个 SNP 候选位点显示多态性。研究结果表明, 通过基于 EST 数据库的 SNP 开发, 可以有效弥补某些海洋生物因基因组学滞后影响 SNP 标记开发的现状。

关键词 长牡蛎, 单核苷酸多态性, 表达序列标签, 片段长度差异等位基因特异性 PCR

中图分类号 Q75

单核苷酸多态性(single nucleotide polymorphisms, SNPs)在水产动物的群体遗传学研究、分子标记辅助育种和生物进化等研究领域有着极为广阔的前景(刘福平等, 2008)。目前 SNP 标记开发一般情况下是通过基因组测序, 利用测序获得的序列之间的比对得到的。因此, 对于已经开展大规模基因组测序的物种来说, SNP 常伴随着基因组序列的拼接过程而产生(杨仑等, 2004; 陈炜等, 2001)。在缺乏基因组序列时, 利用 EST 数据库开发 SNP 标记是目前筛选 SNP 的重要途径。在世界范围内有数个不同科研团体对源于 mRNA 表达的 cDNA (coding DNA)文库进行测序, 由于大多数基因不仅在一种组织中表达, 相同基因就可能被数次测序, 这样就造成了 EST 序列数据库的明显冗余。鉴于 EST 序列数据有较大冗余性, 经过聚类, 作者将来源于同一基因的序列放在同一个簇之中, 通过比对来源于同一基因的冗余 EST 序列就可以获得不同个体间的序列多态信息, 包括候选 SNP 位点。这些候选 SNP 位点进而可以用实验方法加以验证(李延恩等, 2007; He *et al*, 2003)。

在非模式生物中, 通过 EST 数据库开发 SNP 标记是发现 SNP 的一个重要途径, 目前已在水产动物中有若干报道(Qi *et al*, 2008), 在长牡蛎中也有相关报道(Quilang *et al*, 2007)。其基本思路与借助基因组序列开发 SNP 类似, 即经序列聚类后进行拼接, 在重叠区域寻找候选的 SNP 位点。本文采用常规的序列聚类和拼接方法获得 EST 簇, 通过鉴别多条同源序列中的突变位点来确定候选 SNP。候选 SNP 的筛选依赖于合适的检测及分型方法, 目前 SNP 的检测方法有多种, 基于 PCR 技术的常用方法主要有限制性酶切片段长度多态性(restriction fragment length polymorphism, RFLP)、单链构象多态性(single-strand conformation polymorphism, SSCP)以及 *Taq*Man 探针技术和焦磷酸测序技术等(Kwok, 2001; 李玉梅等, 2007)。本实验对候选 SNP 位点的筛选采用的片段长度差异等位基因特异性 PCR(FLDAS-PCR)(黄代新等, 2005; Muriel *et al*, 2007)进行检测。

1 材料与方法

1.1 候选长牡蛎 SNP 位点的获得

1.1.1 长牡蛎 EST 序列的获得 在美国国立生物

* 国家重点基础研究发展计划(973)项目, 2010CB126402 号; 国家自然科学基金项目, 40730845 号, 国家公益性行业(农业)科研专项资助, nyhyzx07-047 号; 中国科学院知识创新工程领域前沿项目资助, 2008—2010。王绍宗, 硕士研究生, E-mail: 6187460@163.com

① 通讯作者: 张国范, 研究员, 博导, E-mail: gfzhang@qdio.ac.cn

收稿日期: 2009-04-09, 收修改稿日期: 2009-06-26

技术信息中心(NCBI)的主页, 以 *Crassostrea gigas* 为关键词搜索长牡蛎的 EST 数据库, 选择 FASTA 格式, 采用批量下载模式, 将全部序列保存至单一文本文档。

1.1.2 EST 序列的聚类和拼接　聚类过程中采用 MEGABLAST 工具对 EST 中的一致序列进行同源比较, 设定的聚类阈值为：2 条 EST 序列的重叠区超过 40 个碱基(Zhang *et al*, 2000), 并采用 Vector NTI (InforMax, Inc.)综合软件包中的 ContigExpress 模块进行序列的拼接, 从而使符合聚类条件的若干个 EST 合并成一簇。

1.1.3 从下载的 EST 文件中设置筛选 SNP 的标准利用聚类后的 EST 簇通过人工判断进行 SNP 的筛查。由于客观上测序错误的存在, 因而只在含有不少于 4 条的 EST 簇中寻找 SNP。去除多态性位点及模糊位点, 并且当一个突变位点在所有 EST 序列中出现的频率大于 20%以上时, 作者认为是一个候选 SNP 位点, 可以进行下一步的检测和验证。

1.2 长牡蛎候选 **SNP** 的验证和功能预测

1.2.1 DNA 模板制备及鉴定　用于 DNA 模板扩增的 33 个野生长牡蛎于 2008 年 9 月采自日本东京湾。提取后用 0.8%琼脂糖凝胶电泳检测其质量, 用紫外分光光度计估算浓度和纯度, 调整浓度至 30ng/μl, 4℃保存备用。

1.2.2 FLDAS-PCR 引物的设计方式及 FLDAS-PCR 原理　利用 Primer Premier 5.0、Oligo 6.0 软件进行引物设计。设计两条长度不同、3′末端分别与 SNP 两个等位基因碱基配对的上游引物, 同时在两个等位基因特异性引物 3′末端第 3 或第 4 位碱基引入错配以增加特异性减少非特异性扩增, 另外在 5′末端添加不同长度的碱基序列尾巴, 下游为公用引物(Muriel *et al*, 2007)。扩增 DNA 片段长度为 100—200bp(图 1)。引物由上海生工公司合成。

1.2.3 DNA 扩增及 SNP 分型　在 25μl 反应体系中, 使用 50—80ng DNA、1U *Taq* DNA 聚合酶、10mmol/L Tris-HCl (pH=8.0)、50mmol/L KCl、2mmol/L $MgCl_2$、200mmol/L dNTP, 三种引物每种 0.5mmol/L。PCR 循环参数设置为：94℃ 3min, 94℃ 1min, 最适退火温度下 30s, 72℃ 30s, 共 30 个循环, 循环结束后 72℃延伸 10min, 4℃保存。扩增产物经非变性 12%的聚丙烯酰胺凝胶电泳, E.B 染色后显带, 由于上游引物的长度不同, 如果候选 SNP 位点被证实：纯合子为单一条带, 杂合子则显示为两条大小相差 8bp 的条带(黄代新等, 2005)(图 2)。

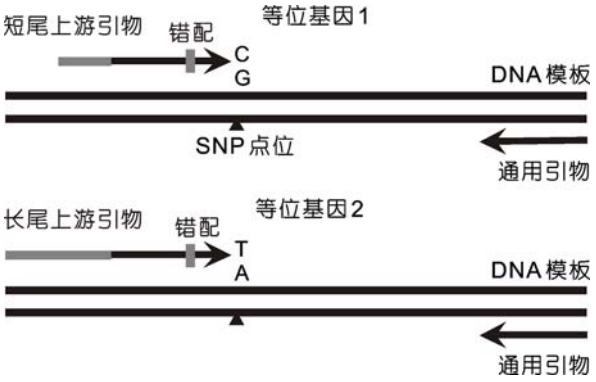

图 1　引物设计示意图(Muriel *et al*, 2007)

Fig.1　Principle of designing primers of FLDAS-PCR

1.2.4 含有长牡蛎 SNP 的 EST-contig 功能预测借助 NCBI 网站 BlastX 程序, 将含有 SNP 的 contig 按照 6 种可能的读码方式翻译后与公共数据库中的蛋白质序列逐一搜索, 当 P 值小于 1E-5 时认为是有效的匹配(Quilang *et al*, 2007), 推测出该 EST 编码的蛋白质序列。同时在结果输出界面可以确定 EST 的可能的开放阅读框, 推算出 SNP 在遗传密码子中的位置, 与标准密码子表比较确定这个突变的性质。

2　结果与分析

2.1 基于 **EST** 数据库开发的候选 **SNP** 数目及类型统计

通过检索, 目前长牡蛎共有 EST 序列 29000 余条。经过序列聚类后得到 EST 簇 4548 个, 含有 4 条和四条以上同源 EST 序列的 EST 簇共有 1079 个。经过后期的人工比对筛选, 共有 313 个 EST 簇可供 SNP 开发使用。

在 313 条拼接形成的 EST-contig 中统计得到候选 SNP 共计 1140 个, 平均一条 contig 含有候选 SNP 3 个, 碱基置换类型 A/G(C/T)、A/C(G/T)、A/T、C/G 的位点数目分别为 464、360、105、211 个, 分别占总数的 40.7%、31.6%、9.2%、18.5%, 碱基置换类型模糊以及三态、四态的位点有 214 个。在不同的 EST 片段中, SNP 的分布密度差别比较大, 从 0.13%—2.5%不等。

2.2 候选 **EST-SNP** 位点的 **FLDAS- PCR** 验证及生物统计学分析

选取以上筛选的 82 个候选 EST-SNP 位点, 设计引物 82 组, 利用 FLDAS- PCR 扩增产物, 并利用 12%的非变性聚丙烯酰胺凝胶进行基因分型检测, 结果共有 17 个 SNP 候选位点显示多态性(表 1), 因此从 EST 库中筛选的候选 SNP 位点在本实验中证实率大

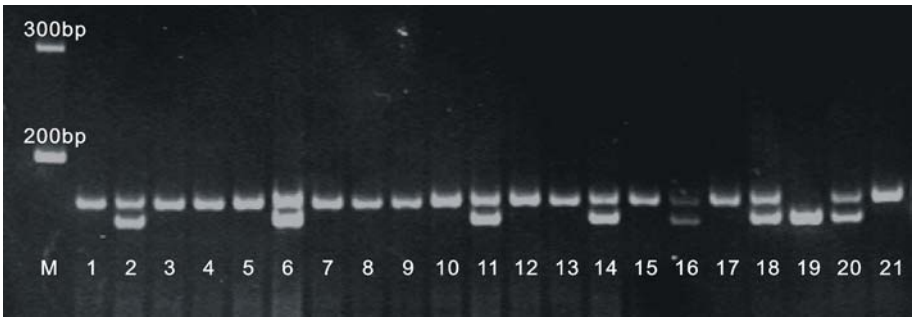

图 2 FLDAS-PCR 对 SNP 位点 S11 的分型结果

Fig.2 The genotyping result of the S11 SNP site using FLDAS-PCR

注：图中所示有三种不同的基因型，其中 2、6、11、14、16、18、20 为杂合子，其余为纯合子，第 19 条带显示的是次要等位基因(minor allele), M 表示 Marker

约为 20%(17/82)。所得的基因型数据利用 PopGene32 (Yeh *et al*, 2000)统计分析各位点上的基因型分布频率、期望杂合度 H_e、观测杂合度 H_o、平均期望杂合度 $\bar{H}_e$、平均观测杂合度 $\bar{H}_o$，并进行哈代-温伯格(HW)平衡、连锁不平衡检验，结果显示以上 17 个 SNP 位点均含有两个等位基因，次要等位基因(minor allele)的出现频率的分布范围为 0.091—0.470，期望杂合度分布区间为 0.088—0.506，观测杂合度分布区间为 0.091—0.667，平均期望杂合度 H_e 为 0.349，平均观测杂合度 H_o 为 0.355。通过哈代-温伯格平衡、连锁不平衡检验，结果显示只有 S04、S06、S10 三个位点的 $P<0.05$，不符合 HW 平衡，其余 14 个位点没有明显的连锁不平衡被证实。通过以上分析可以看出，得到的大多数 SNP 位点能够用于长牡蛎的群体遗传学分析。

2.3 测序分析确认结果

对部分 SNP 位点，随机挑选已经分型的个体，PCR 产物经胶回收测序。测序由上海尼桑生物科技有限公司完成。结果表明，分型结果与测序分析结果完全一致，证明了本实验方法的可靠性。

2.4 长牡蛎 **SNP** 位点的功能预测

借助 NCBI 网站 BlastX 程序对含有 17 个 SNP 的 EST 的共同序列进行 BlastX 分析，当 *P* 值小于 1E-5 时认为是有效的匹配，并据此推测其潜在功能。比对结果显示 17 条 EST-contig 的共同序列中，除 S07 所在的 contigE 值(0.005)大于 1E-5 之外，其余皆符合有效匹配。涉及的基因主要包括管家基因(ribosomal protein、tubulin、myosin 等)、呼吸链蛋白基因(NADH dehydrogenase 等)、生理系统调控蛋白(matrilin、selenoprotein T、cAMP-responsive element binding protein 等)等(表 2)。在结果输出界面观察 EST 的可能的开放阅读框，与标准密码子表比较确定突变的性质，并没有找到错义 SNP(nsSNP)。

3 讨论

鉴于目前长牡蛎基因组研究相对滞后，因此 EST 数目有限，但仍旧能够通过聚类和拼接找到一定数目的候选 SNP。但由于 EST 序列数目较少，可能会导致部分分布频率较低的 SNP 被误判为测序错误而被忽略，虽然在不同的 EST 区段中 SNP 分布密度差别较大(本研究中分布频率在 0.13%—2.5%之间不等)，但总体而言，SNP 在编码区域中一般分布较少，而多数生物在非编码区域可能存在更多的突变位点。Curole 等(2005)指出在长牡蛎基因组中大约每 40bp 就存在 1 个 SNP 位点，从而可推断出在整个长牡蛎核基因组中 SNP 分布密度不会低于 1%。因此有理由认为在长牡蛎的基因组中，SNP 可能呈现较高频率的分布。

在筛查候选 SNP 的过程中，在聚类时采用 MEGABLAST 工具对一致序列进行序列同源比较，并用 Vector NTI 软件的 ContigExpress 模块程序对每一 EST 簇进行拼接检验。这一聚类策略能够降低测序错误带来的影响，并可以有效识别基因家族成员，避免选择性剪接的干扰。此外本实验设置较高的筛选条件，并通过人工判断去除明显的测序错误引入的不合理的插入缺失，从而基本上可以最大程度的排除序列错误带来的假阳性，但是由于本研究不能结合由原始峰图所评估出的测序质量情况进行判断，不能够严格筛选出现于不少于两个文库中的 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，从而增加了假阳性的机率。再者由于长牡蛎基因组中存在大量的内含子序列，在基因表达的时候可能剪切去除过内含子序列，因此 EST 序列中缺少相应的内含子序列，从而得不到目的扩增片段。基于以上几个原因，从 EST 库中筛查得到的 SNP 候选位点在实验中的证实率不是很高，只有大约 20% (17/82)。

在本研究中，在开发得到的 SNP 中并没有找到错义 SNP，这些 SNP 位点根据 EST-contig 的读码框基本都定位在密码子第三位。由遗传密码表可以看出，

除编码色氨酸和甲硫氨酸的密码子第三位不是简并碱基之外, 编码其它氨基酸的密码子在第三位都是简并碱基, 且以 R 和 Y 居多, N 次之, 而在开发的 SNP 中 R/Y 也是最多的, S、 M/K 次之。R/Y 出现的几率较高, 有研究表明可能是因为 CpG 二核苷酸上的胞嘧啶残基是基因组中最易发生突变的位点, 其中大多数发生甲基化, 可自发地脱去氨基而形成胸腺嘧啶。SNP 是物种在长期演化过程中自身遗传与外部环境选择的结果, 因而发生了变异的位点导致的多肽的变化极易导致个体的致死。同义 SNP (sSNP)虽然不改变基因编码的蛋白质, 但仍然具有不容忽视的作用, 特别在影响基因外显子剪切方面发挥重要作用。外显子剪切增强子(Exon Splicing Enhancers, ESEs)是一些位于外显子内部的短序列片段, 它们通过细胞剪切机制增强外显子的识别率。一旦 SNP 发生在某个 ESE 中时, 这些增强子可能会影响 mRNA 的剪切过程, 导致遗漏外显子。实际上, ESE 确实与某些疾病相关的 SNP 有关联, 例如, 与 BRCA1 和 BRCA2 这两个癌症基因密切相关的 SNP, 研究表明, 它们就位于这两个基因内部的 ESE 片段上(Liu *et al*, 2001)。

本实验使用 FLDAS-PCR 分型对候选 SNP 位点进行筛选时, 利用温度梯度 PCR 对模板进行扩增检验温度的影响时发现, 退火温度对产物影响不是很大, 在 55℃附近即可获得较好的扩增效果, 因此本实验方法具有对退火温度要求不是很高的优越性。在使用 FLDAS-PCR 方法筛选 SNP 位点时, 经常出现假阳性, 无法准确分辨出目的片段的基因型, 但是在上游引物 3′端的末位碱基添加人为错配碱基可以有效的降低假阳性。理论上, 只要末位碱基不配对则引物无法有效延伸, 但实际上 3′末端的单碱基错配往往达不到预期的分辨效果, 特别是当突变型相较于野生型的比率较低的时候。因此在 3′末端的倒数第 3 或第 4 个碱基处人为引入第 2 个错配碱基, 用以增加 3′末端的稳定性, 进而提高了聚丙烯酰胺凝胶电泳的分辨效率, 可以达到很好的效果。

总之, 对于目前基因组学研究尚处在初级阶段的海洋生物物种, 利用已有的 EST 数据库进行 SNP 的开发是一条重要的途径。利用片段长度差异等位基因特异性 PCR 对开发的 SNP 进行检测简单实用且成本低廉, 适合于大多数实验室。对于如何更好的克服在此过程中出现的问题, 提高 SNP 的开发效率都有待于进一步探讨和完善。

致谢 长牡蛎 EST 序列聚类分析由中国科学院海洋研究所海洋生物学重点实验室柳承璋博士和王兵博士协助完成, 本实验室吴琪同学在实验技能方面提供了无私的帮助, 谨致谢忱。

参 考 文 献

刘福平, 白俊杰, 2008. 单核苷酸多态性及其在水产动物遗传育种中的应用. 中国水产科学, 15(4): 704—712

李延恩, 周艳红, 2007. SNP 功能分析的生物信息学方法及其资源. 计算机仿真, 24(4): 297—300

李玉梅, 姚纪元, 吴 静等, 2007. PCR-SSCP 技术的研究及应用进展. 生物技术通讯, 6: 71—74

杨 仑, 沈文飚, 陈 虹等, 2004. 基于生物信息学的水稻候选 SNP 的发掘. 中国水稻科学, 18(3): 185—191

陈 炜, 张 戈, 张思仲, 2001. 基于生物信息学的候选位点搜寻方法. 遗传, 23(2): 153—156

黄代新, 杨庆恩, 赵贵森, 2005. 片段长度差异等位基因特异性一种改良的分型新方法. 法医学杂志, 21(1): 11—14

Curole J P, Hedgecock D, 2005. High frequency of SNPs in the Pacific Oyster genome. Plant & Animal Genomes XIII Conference W026, San Diego, 153—156

Qi H, Liu X, Zhang G, 2008. Characterization of 12 single nucleotide polymorphisms (SNPs) in Pacific abalone, *Haliotis discus hannai*. Molecular Ecology Resources, 8: 974—976

He C, Chen L, Simmons M *et al*, 2003. Putative SNP discovery in interspecific hybrids of catfish by comparative EST analysis. Anim Genet, 34: 445—448

Liu H X, Zhang M Q, Krainer A R *et al*, 2001. A mechanism for exon skipping caused by nonsense ormissense mutations in BRCA1 and other genes. Nat Genet, 27(1): 55—58

Muriel Gaudet, Anna-Giulia Fara, Maurizio Sabatti *et al*, 2007. Single-reaction for SNP Genotyping on Agarose Gel by Allele-specific PCR in Black Poplar (*Populus nigra* L.). Plant Mol Biol Rep, 25: 1—9

Kwok Pui-Yan, 2001. Methods for Genotyping Single Nuleotide Polymorphisms. Annu Rev Genomics Hum Genet, 2: 235—258

Quilang J, Wang S, Li P *et al*, 2007. Generation and analysis of ESTs from the eastern oyster, *Crassostrea virginica* Gmelin and identification of microsatellite and SNP markers. BMC Genomics, 8: 157—167

Yeh F C, Yang R, Boyle T J *et al*, 2000. POPGENE 32, Microsoft Windows-Based Freeware for Population Genetic Analysis. Molecular Biology and Biotechnology Centre, University of Alberta, Edmonton, Canada, 153—156

Zhang Z, Schwartz S, Wagner L *et al*, 2000. A greedy algorithm for aligning DNA sequences. J Comput Biol, 7: 203—214

DEVELOPMENT OF 17 SINGLE NUCLEOTIDE POLYMORPHISMS (SNPs) IN *CRASSOSTREA GIGAS* SEARCHED FROM EST DATABASE

WANG Shao-Zong[1, 2], LI Li[2], QI Hai-Gang[2], ZHANG Guo-Fan[2]
(1. *Graduate School, Chinese Academy of Sciences*, *Beijing*, 100049; 2. *Institute of Oceanology*, *Chinese Academy of Sciences*, *Qingdao*, 266071)

Abstract Some 29000 EST sequences of *Crassostrea gigas* from the GenBank were clustered into 4548 groups, of which 1079 contained 4 or more ESTs. After manual quality filtering, 313 clusters could be used for the SNP development. 82 candidate SNPs were amplified with fragment length discrepant allele specific PCR (FLDAS-PCR), of which 17 were polymorphic in a wild population. The function of the ESTs was predicted, and ORF was deduced based on BlastX analysis. The expected and observed heterozygosities of the SNPs ranged from 0.088 to 0.506 and 0.091 to 0.667, respectively. No significant linkage disequilibria were observed in most of the loci. Only three loci did not conform to Hardy-Weinberg equilibrium at the level of $P<0.05$. The results show that EST database is an important source to develop SNPs for the species whose genome research is on the primary stage.

Key words *Crassostrea gigas*, Single nucleotide polymorphism (SNP), Expressed sequence tag (EST), Fragment length discrepant allele specific PCR (FLDAS-PCR)