长牡蛎(*Crassostrea gigas*)EST 串联重复序列的组成和分布

张琳琳[1,2], 李 莉[1], 张国范[1]

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

摘要: 长牡蛎(*Crassostrea gigas*)串联重复序列分析研究较少, 为了研究其在基因组转录本的基本结构特征并为长牡蛎中遗传多样性研究中提供有益的信息, 利用 NCBI 公共数据库中的 57 139 条长牡蛎 ESTs 序列对串联重复序列结构类型、分布、丰度等进行系统比较分析。分析结果表明: 1) 长牡蛎 EST 中共有小卫星串联重复序列 8 392 个, 在大于 100 bp 重复类型中, 162~167bp 含量最高; 2)长牡蛎 EST SSR 含量丰富, 1 954 个位点是 EST-SSR 标记开发的候选资源; 3)EST-SSR 重复数目和重复类型在 5′UTR, CDS 和 3′UTR 具有显著差异, CDS 区承受更大的选择压力。

关键词: 长牡蛎(*Crassostrea gigas*); EST; 串联重复序列; SSR; 选择压力
中图分类号: Q954 文献标识码: A 文章编号: 1000-3096(2011)04-0009-06

长牡蛎(*Crassostrea gigas*)也称太平洋牡蛎, 具有体型大、生长快、产量高、适应性强等优点, 在我国北部沿海大面积养殖, 是我国双壳贝类养殖中规模大、产量高的养殖品种之一。长牡蛎作为冠轮动物超门的模式种, 在大片段文库和遗传图谱的构建, 表达谱差异分析, 杂种优势探讨等方面进行了较详细的研究[1-4], 但基于大规模数据的重复序列方面的研究相对较少[5-6]。本文主要对长牡蛎 EST 进行串联重复序列结构类型, 分布, 丰度等的比较分析。

串联重复序列是指 1~200 个碱基左右的核心重复单位, 以头尾相串联的方式重复多次所组成的重复序列。它们在基因组中有着基因表达调节, 群体遗传多样性分析等重要作用, 与多种疾病相关[7]。 而简单序列重复, 即微卫星(Simple sequence repeat, SSR), 更是广泛地应用于遗传连锁图谱构建[8-9]和物种基因组结构的分析[10]。虽然长牡蛎大规模系统的基因组测序工作还没有完成, 但 NCBI 上公布了大量的长牡蛎 EST(Expressed sequence tags, 表达序列标签)数据。所谓 EST 是指通过对 cDNA 文库随机挑取克隆进行大规模测序所获得的 cDNA 的 5’或 3’端序列, 长度一般为 150~500bp。研究表明长牡蛎 EST 中存在大量重复序列, 可用于 SSR 标记的开发[6], 这为从 EST 中寻找并分析串联重复序列提供了依据。通过物种间和物种内串联重复序列的比较, 研究转录本的结构特征, 分析其串联重复序列特别是 SSR 的分布特征和可能的功能, 将有助于了解基因组的起源和进化, 同时更好地发挥这些序列在串联重复序列标记方面的应用。

截至 2009 年 11 月 1 日, 在 NCBI 数据库中已登录了 57 139 条长牡蛎 ESTs, 但未有对上述 57 139 条 EST 全面的串联重复序列的报道。本研究旨在对现有长牡蛎 EST 中的串联重复序列信息进行结构类型, 分布和丰度比较分析, 以明确长牡蛎串联重复序列的发生频率和特点。同时分析了 SSR 在全长 cDNA 中的分布特点, 以探讨长牡蛎转录本的结构和进化压力。本研究有助于促进串联重复序列特别是 SSR 标记在基因组结构进化和长牡蛎遗传育种中的应用。

1 材料与方法

1.1 长牡蛎 EST 序列的下载和预处理

从 NCBI 库中下载 57 139 条长牡蛎 ESTs(2009-11-01), 过滤长度小于 100 bp 的序列并与 UniVec (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html) 比对去除载体序列, 在去掉 3’末端的 PolyA 后, 得到 56 968 条序列。利用 Sequence Assembly Program, CAP3[11]对上述序列进行初步聚类, 采用的参数为重叠长度阈值 N>30, 重叠的一致性百分比 N>90。

收稿日期: 2010-03-29; 修回日期: 2010-05-20
基金项目: 国家重点基础研究发展计划项目(973 计划, 2010CB126402); 国家自然基金项目(40730845); 中国博士后基金(20090461270)
作者简介: 张琳琳(1985-), 博士研究生, 研究方向: 海洋生物学; 李莉, 通信作者, E-mail: lili@qdio.ac.cn

1.2 长牡蛎 EST 串联重复序列的分析

利用 Tandem Repeat Finder (TRF)[12]对预处理的 EST 进行串联重复序列寻找, 比对参数(匹配, 不匹配, 插入缺失)为 2, 7, 7, 最小比对分值 30, 重复单位最大长度 500。过滤掉重复序列长度不足 15bp 的重复序列。如果同一位置出现的不同重复序列预报, 本研究取重复序列长度最大的类型。长牡蛎的 *Hind*III 卫星序列的多序列比对采用 DNAMAN5.2.2 (Lynnon Biosoft Company)。

1.3 长牡蛎 EST 5'UTR, 3'UTR 和编码区 CDS SSR 分析

从 NCBI 库中下载 644 条长牡蛎蛋白质序列对应的 EST 序列, 手工筛选出含有编码区全长和 5'UTR, 3'UTR 的序列, 共 80 条。分别使用 TRF 分析其 5'UTR, 3'UTR 和 CDS 中 SSR 的分布情况。

2 结果

2.1 长牡蛎 EST 卫星重复序列的分析

在处理后的长牡蛎 EST 中共有 10 997 条串联重复序列(397 019 bp), 其中小卫星重复序列(7~436 bp)有 8 392 条, 共 335 207 bp, 占分析 EST 序列的 1.58% (图 1a, b, c)。重复序列单元总数目和重复类型间有一定规律性。重复序列单元总数目较多集中到 7~12 bp, 其中 9 bp 重复单元数目最多, 为 3 067 个重复单元, 其次是 8 bp, 10 bp, 11 bp, 12 bp, 7 bp。从 13 bp 重复类型开始, 重复单元数目降至 1 000 以下。随着重复单元长度的不断增加, 重复单元数目大致上不断减少。在 24~50 bp 重复之间, 重复单元数目波动相对较大。重复单元长度大于 55 bp 的区域中, 在 63 bp 时出现一个峰, 重复单元数目为 32.4, 其他的重复单元类型相应的重复单元数目均小于 25 bp。重复单元长度大于 300 bp 的只有 3 个重复类型, 相应的重复单元总数目为 6.6。另一方面, 串联重复序列平均拷贝数与重复类型并没有表现出线性关系, 而是呈现不规律性的波动(图 1 d)。

在长串联重复序列的分析中(本文中指串联重复序列的长度大于 100bp 的重复类型), 162~167bp 重复单元呈现一个明显的峰(图 1 c)。将此部分序列提出, 分析发现与长牡蛎的 *Hind*III satellite DNA 具有保守性(图 2)。

2.2 长牡蛎 EST 简单串联重复序列的分析

长牡蛎 EST 中含有丰富的 SSR, 共 2 602 个, 61 744 bp, 占分析序列总碱基的 0.29%(表 1)。重复序列数目表现为六碱基重复序列>单碱基>二碱基>三碱基>五碱基>四碱基, 分别为 851, 805, 307, 258, 240 和 141。重复序列长度、简单重复序列类型与拷贝数的研究过程中, 发现重复序列单元长度与平均

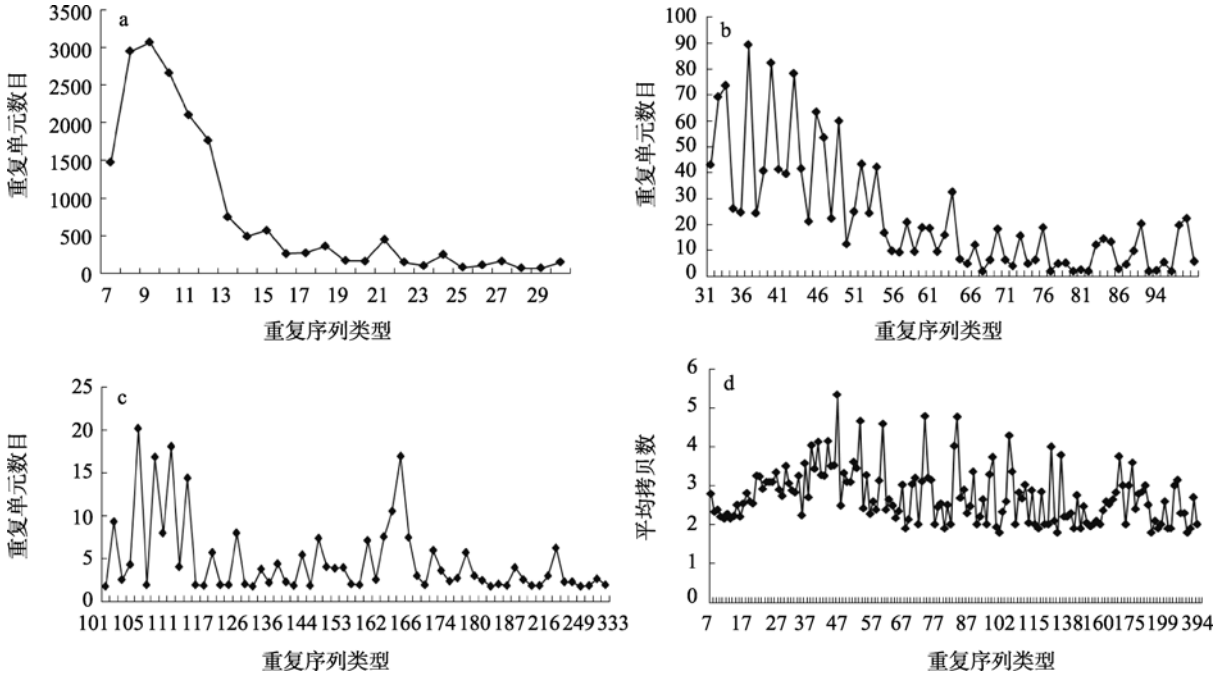

图 1 不同串联重复序列类型在长牡蛎中的拷贝数特征

Fig. 1 The copy number of tandem repeats in the pacific oyster ESTs

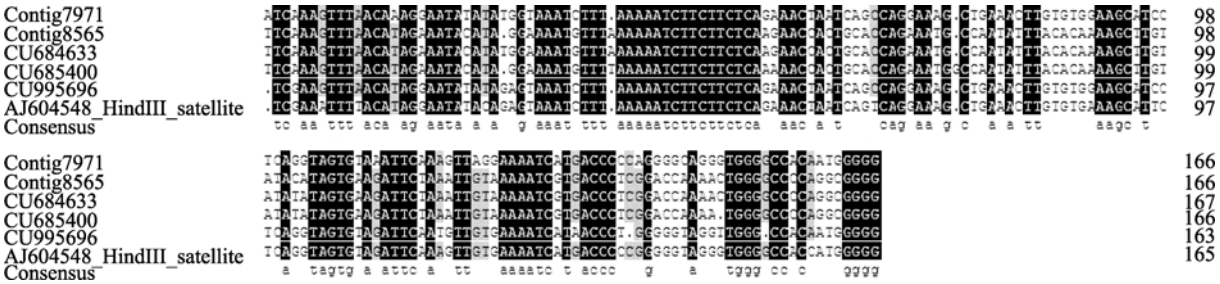

图 2　长牡蛎 *Hind*III 卫星序列的多序列比对

Fig. 2　Alignment of multiple *Hind*III satellites in Pacific oyster

AJ604548_*Hind*III_satellite 为 NCBI 下载的牡蛎的 *Hind*III 卫星序列, NCBI 号为: AJ604548

AJ604548_*Hind*III_satellite is the oyster *Hind*III satellite with NCBI accession AJ604548

拷贝数成反比。另一方面, 相同重复单元长度不同重复类型的重复序列数目、重复序列长度和平均拷贝数有很大的差别。每种重复单元类型代表其互补或顺序不同的所有重复单元, 如 ATC 代表 ATG/TGA/GAT/CAT/ATC/TCA 6 种重复类型。由于四碱基、五碱基、六碱基重复序列的重复类型较多, 我们用 AT 的百分比代替分析 SSR 的分布特征和结构[9]。研究发现, A 串联重复远远大于 T。对于 G 串联重复最大拷贝数为 974, 是因为 NCBI 号为 FP000596 的序列低质量测序, 在除去此序列的影响后, G 串联重复序列的最大拷贝数为 26。不同的重复序列重复类型重复序列的拷贝数目不同, 如二碱基重复中, AG 的重复序列数目远远大于 AT、AC 和 GC。相同重复单元长度不同重复类型的平均拷贝数也有很大差别, 并且与该重复类型的重复序列数目无关, 如 ATC 重复类型的重复序列数目约为 ACT 的 30 倍, 但 ACT 重复类型的平均拷贝数大于 ATC 重复类型。

EST-SSR 在标记应用时, 多是以 PCR 为基础的, 对 SSR 两侧的侧翼序列有一定长度的要求。因此, 本研究统计了简单重复序列两侧的侧翼序列不低于 30bp 的微卫星位点, 统计表明长牡蛎有 1 954 个简单重复序列位点符合要求, 这些位点是微卫星标记开发的候选。

2.3　长牡蛎 cDNA 5'UTR, 3'UTR 和编码区 CDS SSR 分析

对挑选的含有 5'UTR, 3'UTR 以及完整的编码区的 80 条长牡蛎序列分析发现 UTR 区域 SSR 长度所占的比例(0.005 和 0.0026)远远大于 CDS 区域 SSR 所占的比例(0.0011)(表 2)。因为 5'UTR 序列总长度相对较少, SSR 重复单元数目的关系为: 5'UTR < CDS < 3'UTR, 分别为 19, 32.9 和 64.3。此外, cDNA 的位置对简单串联重复序列的重复类型具有选择性。5'UTR 区域只含有单碱基重复单元, CDS 区域只含有三碱基倍数重复单元(三碱基/六碱基), 3'UTR 所含的重复单元类型较为丰富, 含有单碱基, 二碱基和五碱基重复单元。

3　讨论

3.1　长牡蛎 EST 中串联重复序列类型丰富, 162~167bp 类型含量高

从 NCBI 上下载的长牡蛎的 EST 序列中含有丰富的串联重复序列类型。覆盖从 1~436bp 重复类型的 152 种。对长牡蛎 100bp 的重复类型中 162~167 范围的峰值的分析表明, 14 个重复序列中有 5 个与长牡蛎的 *Hind*III 卫星序列具有高的相似度。南极贝(*Adamussium colbecki*)中曾报道了一个 170bp 重复单元的卫星序列, 占基因组序列的 0.2%[13]。该卫星序列之后又被证明在牡蛎中具有中间的保守性, 与哺乳动物的 CENP-B box 具有保守性, 并被用来做牡蛎物种分类的标记[14]。

3.2　长牡蛎 SSR 分布广泛, 1 594 个候选位点

在简单重复序列中, 从单碱基重复到六碱基重复均覆盖大多数重复序列类型。不同的简单重复序列类型的拷贝数目有很大差异。在二碱基重复中, AG 的重复序列数目高达 221, AT 和 AC 均不超过 50, GC 最少为 0, 这与前人的报道相一致[8, 15-16]。在三碱基重复序列中, ATC 重复序列数目最多为 73 次, 其次为 AAC, AAT, AAG, AGG, 其他的类型重复次数均小于 15 次, 这与之前在栉孔扇贝中的报道类似[15]。在四、五、六碱基重复序列中, 我们发现第二高 AT 百分比的重复序列类型拥有更高的重复序列数目, 这与家蚕中的报道相一致[9]。从引物设计的角度考虑,

表 1　长牡蛎 EST 微卫星重复序列的数目、长度和拷贝数特征
Tab. 1　The number, length, and copy number of SSR in the Pacific Oyster EST

重复类型		重复序列数目(条)	重复序列数目占微卫星序列数目百分比(%)	重复序列长度(bp)	重复序列长度占微卫星序列总长度百分比(%)	最大拷贝数	最小拷贝数	平均拷贝数
单碱基	A/T	603	23.17	16 491	26.71	153	15	27.3
	G/C	202	7.76	5 017	8.13	1046	15	24.8
小结		805	30.93	21 508	34.84	1046	15	26.7
二碱基	AT	50	1.92	1 075	1.74	27	7.5	10.8
	AC	36	1.38	810	1.31	17	8	11.3
	AG	221	8.49	8 329	13.49	55	7.5	18.8
小结		307	11.79	10 214	16.54	55	7.5	16.6
三碱基	AAC	48	1.84	1 479.3	2.4	40.7	5	10.3
	AAG	39	1.5	943.5	1.53	21.3	5	8.1
	AAT	40	1.54	765.3	1.24	9.7	5	6.4
	ACC	14	0.54	327.3	0.53	18	5	7.8
	ACT	6	0.23	189.9	0.31	20	5	10.6
	AGG	26	1	519.3	0.84	12.7	5	6.7
	ATC	73	2.81	2 005.5	3.25	17.7	5	9.2
	CAG	9	0.35	203.1	0.33	15	5	7.5
	CGA	2	0.08	38.1	0.06	7.7	5	6.4
	CGC	1	0.04	30.9	0.05	10.3	10.3	10.3
小结		258	9.93	6 502.2	10.54	40.7	5	8.4
四碱基	0%AT	2	0.08	36.8	0.06	5	4.2	4.6
	25%AT	6	0.23	187.2	0.3	11.8	4	7.8
	50%AT	33	1.27	935.2	1.51	27.5	3.8	7.1
	75%AT	59	2.27	1 528.8	2.48	21.2	3.8	6.5
	100%AT	41	1.58	880	1.43	11.5	3.8	5.4
小结		141	5.43	3 568	5.78	27.5	3.8	6.3
五碱基	0%AT	0	0	0	0	0	0	0
	20%AT	10	0.38	233	0.38	7.4	3	4.7
	40%AT	12	0.46	217	0.35	7.2	3	3.6
	60%AT	26	1	474	0.77	5.4	3	3.6
	80%AT	113	4.34	2 068	3.35	6.6	3	3.7
	100%AT	79	3.04	1 589	2.57	18.4	3	4
小结		240	9.22	4581	7.42	18.4	3	3.8
六碱基	0%AT	7	0.27	120.6	0.2	3.5	2.5	2.9
	16.7%AT	53	2.04	969.6	1.57	6.3	2.5	3
	33.3%AT	46	1.77	819	1.33	7.2	2.5	3
	50%AT	160	6.15	3 036.6	4.92	10	2.5	3.2
	67.7%AT	223	8.57	3 906.6	6.33	8.7	2.5	2.9
	83.3%AT	241	9.26	4301.4	6.97	6.7	2.5	3
	100%AT	121	4.65	2 217	3.59	10.3	2.5	3.1
小结		851	32.71	15 370.8	24.91	10.3	2.5	3
总结		2 602	100	61 744	100	1046	2.5	12.7

表 2　长牡蛎 EST 简单重复序列 5'UTR, 3'UTR 和 CDS 特征
Tab. 2　The distributions of 5'UTR, 3'UTR and CDS of SSR in the Pacific Oyster EST

重复类型	SSR 重复单元数目			SSR 长度占所处区域序列总长度比例(%)		
	5'UTR	CDS	3'UTR	5'UTR	CDS	3'UTR
1	19	0	17	0.26	0	0.06
2	0	0	40.5	0	0	0.31
3	0	30.4	0	0	0.1	0
4	0	0	0	0	0	0
5	0	0	6.8	0	0	0.13
6	0	2.5	0	0	0.01	0
小结	19	32.9	64.3	0.26	0.11	0.5

有 1 594 个位点为微卫星标记开发的候选位点，该结果为进一步开发长牡蛎 EST-SSR 标记奠定了基础。

3.3　长牡蛎 EST 的 CDS 区承受更大的选择压力

CDS 区域简单串联重复序列相对较少，这与编码区受到的选择压力大于 UTR 区域有关，而编码区的重复序列类型为三碱基和六碱基，这两种碱基类型均为编码氨基酸的密码子数目 3 的倍数，这更说明了非 3 倍数的简单重复序列对编码区具有破坏作用，而自然选择将这部分破坏的简单重复序列淘汰了，这与水稻中的报道相一致[17]。在本研究中，编码区三碱基重复序列的类型为 ACA, GAA 和 GAT 重复，推测该三种重复类型可能与串联重复数目具有一定联系，其进一步研究可能需要使用更多的全长 cDNA 才能得出更明确的结论。

参考文献:

[1] Cunningham C, Hikima J, Jenny M J, et al. New resources for marine genomics: bacterial artificial chromosome libraries for the Eastern and Pacific oysters (*Crassostrea virginica* and *C. gigas*)[J]. Mar Biotechnol (NY), 2006, 8(5): 521-533.

[2] Hubert S, Hedgecock D. Linkage maps of microsatellite DNA markers for the Pacific oyster *Crassostrea gigas*[J]. Genetics, 2004, 168(1): 351-362.

[3] Fleury E, Huvet A, Lelong C, et al. Generation and analysis of a 29,745 unique Expressed Sequence Tags from the Pacific oyster (*Crassostrea gigas*) assembled into a publicly accessible database: the Gigas Database[J]. Bmc Genomics, 2009, 10: 341.

[4] Hedgecock D, Lin J Z, DeCola S, et al. Transcriptomic analysis of growth heterosis in larval Pacific oysters (*Crassostrea gigas*)[J]. Proc Natl Acad Sci U S A, 2007, 104(7): 2313-2318.

[5] Wang Y, Guo X. Development and characterization of EST-SSR markers in the eastern oyster *Crassostrea virginica*[J]. Mar Biotechnol (NY), 2007, 9(4): 500-511.

[6] Wang Y, Ren R, Yu Z. Bioinformatic mining of EST-SSR loci in the Pacific oyster, *Crassostrea gigas*[J]. Anim Genet, 2008, 39(3): 287-289.

[7] Richard G F, Kerrest A, Dujon B. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes[J]. Microbiol Mol Biol Rev, 2008, 72(4): 686-727.

[8] Toth G, Gaspari Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis[J]. Genome Res, 2000, 10(7): 967-981.

[9] Prasad M D, Muthulakshmi M, Madhu M, et al. Survey and analysis of microsatellites in the silkworm, Bombyx mori: frequency, distribution, mutations, marker potential and their conservation in heterologous species[J]. Genetics, 2005, 169(1): 197-214.

[10] Subramanian S, Mishra R K, Singh L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions[J]. Genome Biol, 2003, 4(2): R13.

[11] Huang X Q, Madan A. CAP3: A DNA sequence assembly program[J]. Genome Research, 1999, 9(9): 868-877.

[12] Benson G. Tandem repeats finder: a program to analyze DNA sequences[J]. Nucleic Acids Res, 1999, 27(2): 573-580.

[13] Canapa A, Barucca M, Cerioni P N, et al. A satellite DNA containing CENP-B box-like motifs is present in the antarctic scallop *Adamussium colbecki*[J]. Gene,

2000, 247(1-2): 175-180.

[14] Lopez-Flores I, de la Herran R, Garrido-Ramos M A, et al. The molecular phylogeny of oysters based on a satellite DNA related to transposons[J]. Gene, 2004, 339: 181-188.

[15] Zhang L, Chen C, Cheng J, et al. Initial analysis of tandemly repetitive sequences in the genome of Zhikong scallop (*Chlamys farreri* Jones et Preston)[J]. DNA Seq, 2008, 19(3): 195-205.

[16] Li Y C, Korol A B, Fahima T, et al. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review[J]. Mol Ecol, 2002, 11(12): 2453-2465.

[17] Zhang Z and Xue Q. Tri-nucleotide repeats and their association with genes in rice genome[J]. Biosystems, 2005, 82(3): 248-256.

Bioinformatics data mining of EST- tandem repeats of the Pacific Oyster (*Crassostrea gigas*)

ZHANG Lin-lin[1,2], LI Li[1], ZHANG Guo-fan[1]

(1. Institute of Oceanology, the Chinese Academy of Sciences, Qingdao 266071, China; 2. Graduate University, the Chinese Academy of Sciences, Beijing 100049, China)

Received: Mar., 29, 2010
Key words: *Crassostrea gigas*; EST; tandem repeat; SSR; selective pressure

Abstract: Large scale analysis of tandem repeats in the Pacific Oyster is underdeveloped on the level of ESTs. The analysis of tandem repeats will be useful to a variety of applications in the study of transcripts characteristics and population genetics of the Pacific Oyster. We analyzed tandem repeats in the Pacific Oyster based on 57 139 ESTs downloaded from NCBI. The major results were as follows: 1) we obtained 8 392 minisatellite sequences, which were much more widely dispersed in the repeat types of 162~167 bp (>100bp); 2) ESTs were abundant, and 1 954 EST-SSRs were of the potential for designing primers specific to flanking sequences; 3) the repeat counts and repeat types were significantly different for 5'UTR, 3'UTR, and CDS; CDS undergoes much greater selective pressure.

(本文编辑: 张培新)