文章编号：1674－5566(2012)06－0945－06

大口黑鲈两个亚种EST数据库分析

景燕娟[1,2]，白俊杰[1]，李胜杰[1]，于凌云[1]，蔡　磊[1,2]

(1.中国水产科学研究院珠江水产研究所 农业部热带亚热带水产资源利用与养殖重点实验室，广东 广州　510380；
2.上海海洋大学 水产与生命学院，上海　201306)

摘　要：应用新一代高通量测序技术Roche 454对大口黑鲈(*Micropterus salmoides*)北方亚种和佛罗里达亚种进行转录组测序并建立ESTs数据库，结果得到北方亚种ESTs序列468 671条，佛罗里达亚种ESTs序列332 322条，两亚种ESTs序列的平均长度分别为306.5 bp和304.4 bp。将得到的高质量序列进行拼接，大口黑鲈北方亚种和佛罗里达亚种共得到contig序列总数分别为42 056条和35 743条，平均长度分别为612.6 bp和588.2 bp。将大口黑鲈北方亚种和佛罗里达亚种的数据库合并，通过与蛋白数据库比对后，共78 938条EST序列被注释，根据Gene Ontology(GO)信息，对序列按照分子功能、细胞组成、生物学过程进行分类。通过对大口黑鲈合并的EST库信息进行大通量SSR和SNP位点的发掘，发现了含SSRs和SNPs的序列分别为25 469和8 547条。从EST数据库中随机选取75条contigs进行北方亚种和佛罗里达亚种的序列比较，结果表明两个亚种EST序列同源性为99.2%。

研究亮点：大口黑鲈EST数据库的建立将为开展大口黑鲈功能基因调控机制研究、分子标记开发以及两个亚种的比较基因组学等研究提供基础资料。相关内容国内还未见报道，具有较好的创新性。

关键词：大口黑鲈；北方亚种；佛罗里达亚种；罗氏454测序；表达序列标签
中图分类号：S 917
文献标志码：A

表达序列标签(expressed sequence tags，ESTs)技术是将mRNA反转录成cDNA并克隆到质粒或噬菌体载体构建成cDNA文库，从中随机挑取克隆并对其3′或5′端进行单轮测序。构建的转录组数据可为寻找新的编码蛋白和非编码蛋白信息和对基因组序列的功能注释、组织特异性基因表达谱分析、连锁图谱的构建、SSR、SNP的研究提供帮助[1－3]。现在，大规模cDNA测序工作在水产领域已有了研究和应用，如鲤[4](*Ctenopharyngodon idella*)、斑点叉尾鮰[5](*Ictalurus furcatus*)、尖吻鲈[6](*Lates calcarifer*)、马氏珠母贝[7](*Pinctada martensii*)等已经完成了测序工作。大口黑鲈(*Micropterus salmoides*)俗称加州鲈，广泛分布于美国、加拿大等淡水水域，按照地理分布和形态学方面的不同，可分为北方亚种和佛罗里达亚种[8]。20世纪70年代我国台湾从原产地引进此鱼，并于1983年引进到广东省，现已推广到全国各地，已成为我国重要的淡水养殖品种之一[9]。本研究以大口黑鲈北方亚种和佛罗里达亚种的脑、肌肉和肝脏组织为材料，通过新一代高通量的Roche 454测序仪获得了大量的ESTs序列，并对其进行生物信息学分析。本研究结果将为开展大口黑鲈功能基因调控机制研究、分子标记开发以及两个亚种的比较基因组学等研究提供基础资料。

1　材料与方法

1.1　材料

大口黑鲈北方亚种(*M. salmoides salmoides*)与佛罗里达亚种(*M. salmoides floridanus*)均取自广州珠江水产研究所良种基地。分别挑选6月龄大口黑鲈北方亚种和佛罗里达亚种各15尾，

收稿日期：2012-07-06　　**修回日期：**2012-08-12
基金项目：农业部“948”项目(2010－G12)；农业部公益性行业科研专项(200903045)；国家科技支撑计划项目(2012BAD26B03)
作者简介：景燕娟(1987—)，女，硕士研究生，研究方向为水生生物遗传育种。E-mail：jingyanjuan66@163.com
通讯作者：白俊杰，E-mail：jjbai@163.net

取大脑、肌肉、肝脏组织，并将两个亚种的这些组织分别混合，用液氮冷冻，-80 ℃保存备用。

1.2 研究方法

1.2.1 总 RNA 提取和 cDNA 反转录

总 RNA 提取参照 XU 等[10]的方法。以 1 μg 总 RNA 为模板，采用 SMART™ PCR cDNA Synthesis Kit（Clontech）反转录合成 cDNA，然后采用 PCR Advantage Ⅱ polymerase（Clontech）对 cDNA 进行扩增，扩增条件为 95 ℃ 1 min；94 ℃ 15 s，65 ℃ 30 s，68 ℃ 3 min，18 个循环。最后采用 PureLink™ PCR Purification kit（Invitrogen）去除体系中小于 300 bp 的片段。

1.2.2 454 文库构建和测序

应用新一代高通量的 Roche 454（GS FLX Titanium System）测序仪分别对两个亚种的 5 μg cDNA 样品测序。双链 cDNA 打断为 300 ~ 800 bp 的片段后，两端添加特异性衔接子 A 和 B，变性为单链连接到磁珠上，经 Emulsion PCR 富集后，置于 PicoTiterPlate 板上，上机测序。

1.2.3 序列处理与拼接

原始序列采用 SeqClean 和 Lucy 软件去掉文件制备及测序过程中所用的接头序列、低度复杂序列、头尾低质量区域，以及最终长度小于 100 bp 的序列。将得到的序列进行 2 次 cap3 拼接，第 1 次控制质量分数中断点为 15 次，第二次控制相似性在 95% 以上。GC 含量分析窗口移动值为 51 bp。

2 结果与分析

2.1 EST 序列和 EST 序列拼接

采用 Roche 454 GS FLX 高通量测序仪测序获得的大口黑鲈北方亚种和佛罗里达亚种 EST 序列分别为 468 671 条和 332 322 条，两亚种 EST 序列的平均长度分别为 306.5 bp 和 304.4 bp。两个亚种的 EST 序列长度分布见图 1。从图 1 中可以看出，北方亚种 EST 小于 100 bp 的 EST 序列占 9.9%，而佛罗里达亚种占 10.7%，北方亚种获得的 EST 序列质量稍高。

将得到的两个亚种 EST 数据库中的序列进行各自拼接，分别得到北方亚种的 contigs 序列数为 42 056 条，佛罗里达亚种的为 35 743 条，然后将两个数据库混合拼接共得到大口黑鲈 contig 序列总数为 50 736 条。北方亚种和佛罗里达亚种的 contig 平均长度分别是 612.6 bp 和 588.2 bp，最小的 contig 序列长度都是 42 bp，最大的 contig 序列长度分别是 8 216 bp 和 8 689 bp。没有进入拼接的 EST 序列分别为 93 740 条和 69 343 条。图 2 展示了 contig 序列长度的分布情况。两个亚种的 contig 序列长度在 300 ~ 600 bp 之间，均具有较高的频率分布。

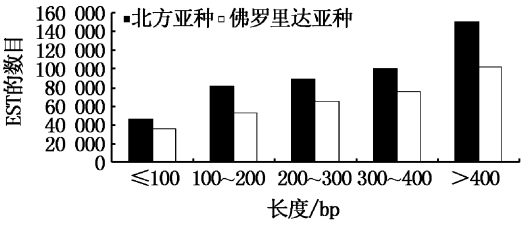

图 1 两个亚种 EST 有效长度分布

Fig. 1 Length distribution of two subspecies EST effective length

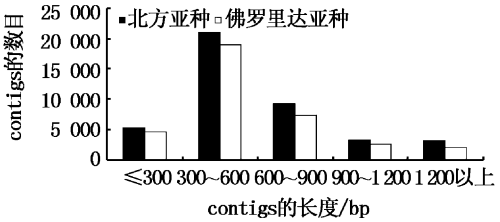

图 2 拼接后两个亚种 contig 长度分布

Fig. 2 Length distribution of two subspecies EST assembly

2.2 GO 分类

GO 是一个国际标准化的基因功能分类体系，提供了一套动态更新的标准词汇表（controlled vocabulary）和严格定义的概念描述来全面地概括任何生物体中基因和基因产物的属性。GO 总共有 3 个本体（ontology），分别是基因发挥的分子功能（molecular function）、细胞组分（cellular component）、参与的生物过程（biological process），GO 的基本单位是词条、节点，每个词条、节点都属于一个本体。

将大口黑鲈两个亚种 EST 库合并，拼接成 contig，将拼接后的序列和没有拼接进去的序列 singleton 与蛋白库进行比对，有 78 938 条 EST 序列被注释了功能。将这些被注释功能的基因序列进行 GO 分类分析。其中，有 2 584 条 EST 序列归入“分子功能”中，注释了 14 个亚类；929 条

EST 序列归入“细胞组分”中,注释了 5 个亚类;5 424条 EST 序列归入“生物过程”中,注释了 16 个亚类。图 3 是亚类注释的数量。

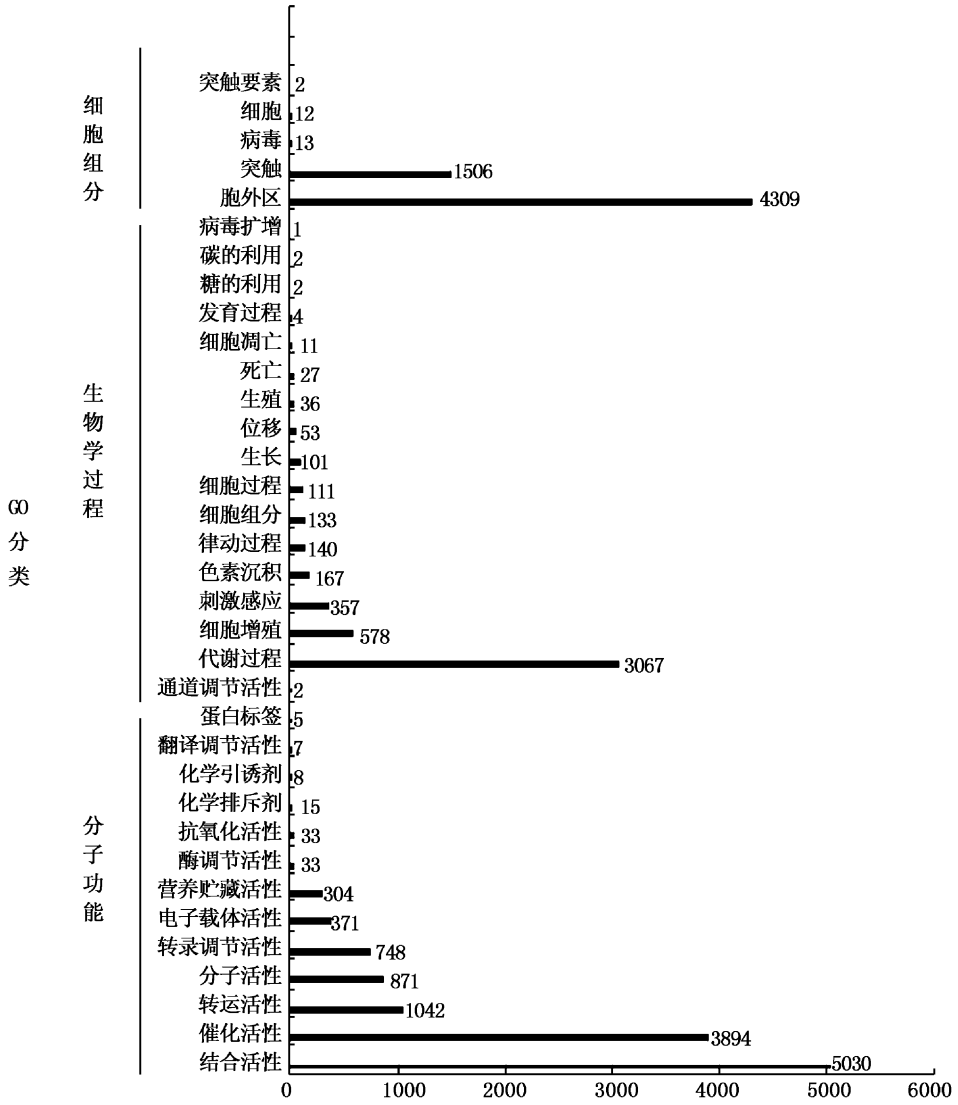

图 3　基因 GO 功能注释分类
Fig. 3　GO function annotations

2.3　大口黑鲈 EST-SSRs 的分布特征

在北方亚种和佛罗里达亚种合并的 EST 库中,共 25 469 个 EST-SSRs,142 种重复基元。其中有 8 234 个 SSRs 存在于拼接后的 contig 序列中,46.15% 有功能注释;17 235 个存在于没有进入拼接的 singlet 序列中,15.78% 有功能注释。在所有重复基元中,二核苷酸重复基元出现的频率最高为 32.00%,其次分别是三、五和四核苷酸重复基元(表 1)。

表 1　大口黑鲈 SSRs 中不同重复基元出现的频率
Tab. 1　Occurrence frequency of different repeat motifs in the largemouth bass SSR

重复基元类型	数量	频率/%	最多的重复基元(数量,百分比)
二核苷酸	8 154	32.15	AC/TG(5 762,75.66%)
三核苷酸	7 877	30.93	AGG/TCC(2 662,33.79%)
四核苷酸	3 674	14.43	AAAT/TTTA(571,15.54%)
五核苷酸	5 764	22.63	AAAAC/TTTTG(731,12.68%)

2.4 EST－SNP 出现的频率及突变类型

在大口黑鲈 EST 库中，按照 reads 与 contig 对位排列分析大于等于 0.9，SNP 数≥5，质量值≥20 的条件，北方亚种和佛罗里达亚种都有 8 547个 SNP 位点，按照 reads 与 contig 对位排列分析大于等于 0.95，SNP 数≥5，质量值≥20，两个亚种都有 2755 个 SNP 位点。本研究从第一种情况对 8 547 个 SNP 位点进行分析，结果显示北方亚种和佛罗里达亚种的碱基置换有 7 054 个（82.53%），碱基转换占 17.40%，碱基颠换占 75.30%，各个类型的置换见表 2。其中，一个基因位点测出来 3 种碱基的有 502 个，4 种碱基的有 13 个。碱基插入或缺失有 1 476 个（17.27%）。经过计算得出大口黑鲈北方亚种和佛罗里达亚种编码区的 SNP 发生频率分别为 0.059 5% 和 0.084 5%（单核苷酸多态位点数与测到的高质量的碱基总数的比值），平均每 1 700 bp和 1 200 bp 就有一个 SNP 位点。

表 2 大口黑鲈 EST-SNP 的类型及数量

Tab.2 The type and number of mutation in EST-SNP of *Micropterus salmoides*

SNP 类型	碱基转换		碱基颠换			
	A/T	C/G	A/C	A/G	C/T	G/T
数量	584	643	621	1 990	2 027	674
百分比	8.28%	9.12%	8.80%	28.21%	28.74%	9.55%
转换和颠换各占百分比	17.40%		75.30%			

2.5 大口黑鲈北方亚种和佛罗里达亚种基因差异的比较

从大口黑鲈北方亚种 EST 数据库中随机选取 75 条 contigs，然后在佛罗里达亚种数据库中找到与之对应的序列。用软件 Vector NTI Suite 8（invitrogen）将两个亚种比较，结果显示，EST 序列的同源性大于或等于 99.0% 有 67 条，占随机选取 contigs 的 89.3%（图 4），平均同源性为 99.2%，表明两个亚种 cDNA 差异微小。

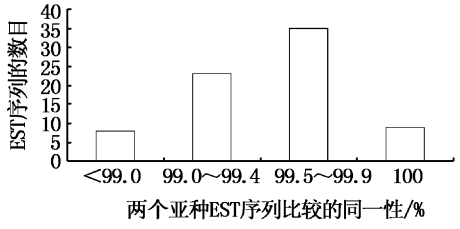

图 4 北方亚种和佛罗里达亚种 ESTs 序列同源性分布

Fig.4 ESTs sequence matching rate distribution of the northern subspecies and florida subspecies

从已随机选择的 75 条 EST 序列中再随机选择 25 条并对其设计引物，分析北方亚种和佛罗里达亚种 EST 序列的真实差异性。结果显示 25 对引物中有 23 对成功的扩增出了片段，在北方亚种和佛罗里达亚种中相一致的序列有 21 条，其余 2 条 EST 序列各存在 1 个突变位点，进一步的分析表明两个突变点没有改变编码的蛋白质，属于无义突变。

3 讨论

大口黑鲈 cDNA 数据库的建立，得到了大口黑鲈几十万条 EST 序列资源。许多学者的研究表明，数据库中 EST 数据的精确度大约为 97%[11]，而大口黑鲈的平均精确度为 99%，说明大口黑鲈的 cDNA 文库测序数据比较准确、可靠，这为基因组重复和基因组进化的分析和通过种质遗传学分析提供基础。鉴于北方亚种和佛罗里达亚种很近的亲缘关系，将两个亚种 EST 数据库合并，预计大部分重复的 EST 序列将可以拼接组装，提供大口黑鲈更为完整的转录组。经两个亚种的 EST 数据库合并拼接后的大口黑鲈 contigs 序列总数为 50 736 条，远大于北方亚种（42 056）和佛罗里达亚种（35 743）的任何一种，这可能是由以下 3 个原因引起的：（1）一些 ESTs 序列存在于北方亚种中，而不在佛罗里达亚种，反之亦然；（2）无论是北方亚种还是佛罗里达亚种，没有拼接的 singletons 在混合后会拼接成新的 contigs；（3）两个亚种之间序列变异和拼接的不同可能导致 contigs 数目的增大。

大口黑鲈北方亚种和佛罗里达亚种编码区的 SNP 发生频率分别为 0.059 5% 和 0.084 5%，平均大约每 1 700 bp 和 1 200 bp 就有一个 SNP 位点。说明大口黑鲈的 SNP 发生频率比较低，要低于草鱼中 EST 的 SNP 发生率（平均 1 000 bp 发

现 7.7 个 SNP)[12],也低于鲤的发生率(平均 1 000 bp 发现 2.9 个 SNP)[4]。与已研究的很多植物相比,大口黑鲈的 SNP 的发生频率更低。如 OSSOWSKI[13]对拟南芥的 3 个自然品系的基因组测序及重测序研究发现了超过 82 万的 SNP 位点,平均约每 1.4 kb 就会有一个 SNP。茶树 EST 中 SNP 的发生频率为每 200 bp 就有一个 SNP 位点[14],桉树[15]的 4 个种每 100 bp 大约有 3.83 ~ 7.30 个 SNP 个位点,柑橘[16]的是每 1 000 bp 有 6.1个。

点突变是 DNA 序列进化中最重要的因素之一。很多研究表明,即使是无功能的假基因,碱基突变方向也不是随机的,如碱基颠换 A/G 的发生频率比 C/T 更高。本研究结果表明,大口黑鲈 EST 序列中的 SNP 大部分来源于颠换,部分来源于转换。在 6 种碱基替换类型中,大口黑鲈的 SNP 位点以 C/T 和 A/G 的碱基颠换为主,分别占 28.21% 和 28.74%,与王丽鸳的茶树[12]的研究结果(C/T 和 A/G 的转化分别占 31.78% 和 29.10%)、张晓红[17]的桉树 EST-SNP 研究结果(A/G 和 C/T 的碱基颠换分别占 32.4% 和 35.0%)有相似趋势。有研究表明 CpG 二核苷酸的胞嘧啶(C)是人类基因组中最易发生突变的位点,其中大多数是甲基化的,可自发地脱去氨基而形成胸腺嘧啶(T),导致颠换型变异的 SNP 约占总数的三分之二。

参考文献:

[1] 刑巨斌,谢彩霞,张逸飞,等. 变态前牙鲆 cDNA 文库中 I 型微卫星的多态性分析[J]. 上海海洋大学学报,2011,20(2):167 - 172.

[2] JING D, YE Q L, WANG F S, et al. The mining of citrus EST-SNP and its application in cultivar discrimination [J]. ScienceDirect, 2010, 9(2): 179 - 190.

[3] WOODS I G, KELLY P D, CHU F, et al. A. comparative map of the zebrafish genome [J]. Genome Research, 2000, 10: 190 - 1914.

[4] 张晓峰,杨晶,孙效文. 基于 EST 序列的鲤鱼生长相关 SNP 发掘[J]. 水产学杂志,2009,22(4):1 - 7.

[5] WANG S L, ERIC P, JASON A, et al. Assembly of 500000 inter-specific catfish expressed sequence tags and large scale gene-associated marker development for whole genome association studies[J]. Genome Biology, 2010, 11(1): R8.

[6] XIA J H, GEN H Y. Identification and analysis of immune-related transcriptome in Asian seabass Lates calcarifer[J]. BMC Genomics, 2010, 11: 356.

[7] 王爱民. 马氏珠母贝生长相关杂种优势和 EST 序列分析[D]. 上海:上海海洋大学,2010.

[8] BAILEY R M, HUBBS C L. The black basses (Micropterus) of Florida, with description of a new species [J]. University of Michigan Museum of Zoology Occasional Papers, 1949, 516: 1 - 40.

[9] BAI J J, LUTZ-CARRILLO D J, QUAN Y C, et al. Taxonomic status and genetic diversity of cultured largemouth bass Micropterus salmoides in China[J]. Aquaculture, 2008, 278(1/4): 27 - 30.

[10] XU M, ZANG B, YAO H S, et al. Isolation of high quality RNA and molecular manipulations with various tissues of Populus[J]. Russian Journal of Plant Physiology, 2009, 56(5): 716 - 719.

[11] HILLIER L D, LENNON G, BECK M, et al. Generation and analysis of 280000 human expressed sequence tags [J]. Genome Research, 1996, 6(9): 807 - 828.

[12] XU B, WANG S L, JING Y, et al. Generation and Analysis of ESTs from the Grass Carp, Ctenopharyngodon idellus [J]. Animal Biotechnology, 2010, 21: 217 - 225.

[13] OSSOWSKI S, SCHNEEBERGER K, CLARK R M, et al. Sequencing of natural strains of *Arabidopsis thaliana* with short reads[J]. Genome Research, 2008, 18(12): 2024 - 2033.

[14] 王丽鸳. 基于 EST 数据库和转录组测序的茶树 DNA 分子标记开发与应用研究. [D]. 杭州:中国农业科学院,2011.

[15] CARSTEN K, SUAT H Y, JENS M, et al. Comparative SNP diversity among four Eucalyptus species for genes from secondary metabolite biosynthetic pathways [J]. BMC Genomics, 2009, 10: 452.

[16] JIANG D, YE Q L, WANG F S, et al. The mining of citrus EST-SNP and its application in cultivar discrimination[J]. Agricultural Sciences in China, 2010, 9(2): 179 - 190.

[17] 张晓红. 桉树 EST-SNP 的开发及 EST 图谱的构建[D]. 南京:南京农业大学,2009.

Analysis of EST database for two subspecies of largemouth bass

JING Yan-juan[1,2], BAI Jun-jie[1], LI Sheng-jie[1], YU Ling-yun[1], CAI Lei[1,2]

(1. *Key Laboratory of Tropical & Subtropical Fishery Resource Application & Cultivation, Ministry of Agriculture, Pearl River Fisheries Research Institute of Chinese Academy of Fishery Sciences, Guangzhou* 510380, *Guangdong, China*; 2. *College of Fisheries and Life Science, Shanghai Ocean University, Shanghai* 201306, *China*)

Abstract: In this study, Roche 454 high-throughput technology was used to conduct transcriptome sequencing and establish ESTs database of the largemouth bass northern subspecies and Florida subspecies. The results showed that a total of 468 671 and 332 322 ESTs were generated from northern subspecies and Florida subspecies. The average length was 306.5 bp and 304.4 bp, respectively. Assembly of the largemouth bass northern subspecies and Florida subspecies' high quality ESTs resulted in 42 056 and 35 743 contigs. The average length was 612.6 bp and 588.2 bp, respectively. EST database of largemouth bass northern subspecies and Florida subspecies was merged and then compared with known protein databases. 78 938 EST sequences in total were annotated. By exploring the merged largemouth bass EST library, over 25 469 and 8 547 putative SSRs and SNPs were identified. Comparison of the northern subspecies and Florida subspecies sequence was made by randomly selecting 75 contigs from the EST database. Results showed that EST sequences of the two subspecies homology were 99.2%.

Key words: largemouth bass; northern subspecies; florida subspecies; roche 454 sequencing; ESTs