DOI:10.15928/j.1674-3075.2018.06.013

基于RNA-seq的黄尾鲴肝脏转录组测序与分析

张燕萍[1],章海鑫[1],崔 璀[2],傅义龙[1],刘志放[2],范鸿潮[2]

(1.江西省水产科学研究所,江西 南昌 330000;
2.萍乡市水产科学研究所,江西 萍乡 337000)

摘要: 为了发掘黄尾鲴(*Xenocypris davidi*)功能基因特别是免疫相关基因,为其种质资源评价、群体遗传学多样性分析、基因连锁图谱构建、免疫相关功能基因地位以及分子标记辅助育种提供基础信息,采用Illumina HiSeq 2500高通量技术平台对黄尾鲴肝脏进行转录组测序。通过去除低质量的raw reads,共获得了50 702 046条clean reads,组装得到53 527条Unigenes。采用BLAST相似性比对方法,把这些序列比对NR、String、Swiss-sprot、KEGG和Pfam数据库,共有26 613条Unigenes得到了注释。有15 532条Unigenes获得了GO注释并分类到64个功能类别中,有7 737条Unigenes被归纳到COG的26个功能分类,共有14 642条获得了KEGG注释,共参与33条代谢通路中。根据免疫系统分类,共有1 299条Unigenes参与了16条代谢通路。从53 527个Unigenes中共找到98 826个单核苷酸多态性位点(SNP)(转换64 396个,颠换34 430个)和18 119个SSR位点。研究结果为黄尾鲴免疫学、基因组学、基因克隆以及分子标记辅助育种提供了重要依据。

关键词: 黄尾鲴;转录组;高通量测序;功能注释

中图分类号: Q503 **文献标志码:** A **文章编号:** 1674-3075(2018)06-0087-08

黄尾鲴(*Xenocypris davidi*)隶属于鲤科(Cyprinidae)、鲴亚科(Xenocyprininae)、鲴属(*Xenocypris*),为底层鱼类,通常生活在江河、湖泊的中下层,尤其喜栖息于多水草、软泥底质的水域底层,是一种中小型淡水鱼类(叶富良,1987)。由于其具有肉厚质实、味道鲜美、营养价值高、调控水质等特点,已成为当前水产养殖品种结构调整中首选的优良品种之一。近年来,由于黄尾鲴野生资源被过度捕捞,导致其野生资源量锐减,市场供不应求,促进了人工养殖的发展。随着黄尾鲴养殖规模的不断扩大,特别是池塘高密度养殖条件下,其病害频繁爆发,并且由于近交等因素引起种质衰退,导致经济损失巨大。

黄尾鲴为我国特有种类,国外未见有相关研究报道。从20世纪50年代开始移入池塘驯养,至60年代初人工繁殖获得成功,目前国内在人工繁殖、育苗、养殖方面已有相关的研究与报道(凌志勇,2002;冯晓宇等,2005;黄邦新,2006);但其免疫学、分子遗传学、基因序列和分子标记方面未见报道,导致有关黄尾鲴的免疫机制、防御机理的遗传信息、基因序列不足以及分子标记缺乏等。肝脏作为动物个体最大的代谢器官,其在个体生命活动中承担着极其重要的作用,参与了很多疾病的发生和消亡过程。因此,本研究利用Illumina测序技术对黄尾鲴肝脏进行了转录组测序分析,旨在发掘其功能基因特别是免疫相关基因,期望为今后黄尾鲴免疫系统和免疫防御机制提供资源;同时开发一批分子标记,为其种质资源评价、群体遗传学多样性分析、基因连锁图谱构建、免疫相关功能基因地位以及分子标记辅助育种提供基础信息。

1 材料与方法

1.1 试验材料

试验用黄尾鲴由江西省萍乡市水产科学研究所自繁后池塘养殖的成鱼。选择个体大、体表无损伤的黄尾鲴个体,体重为(204.2±28.9) g,在该所水族箱中充气暂养1周。尾部放血,取其肝脏(3尾混合),经液氮速冻后,用干冰运送至上海美吉生物有限公司进行转录组测序。

1.2 总RNA提取

总RNA的提取根据Trizol reagent试剂盒说明书操作。总RNA的浓度和纯度用紫外分光光度计进行检测,并用0.8%的琼脂糖凝胶电泳检测和A_{260}/A_{280}比率确定RNA的完整性,用带有Oligo(dT)的磁珠富集mRNA。

收稿日期: 2017-01-04

基金项目: 江西省科技计划项目"黄尾密鲴良种选育及健康养殖技术研究与推广"(20141BBF60036)。

作者简介: 张燕萍,1979年生,女,博士,助理研究员,主要从事鱼类育种、渔业生态环境和渔业资源调查研究。E-mail:zhangyanpingxie@163.com

1.3 文库构建与测序

采用 Illumina TruseqTM RNA sample prep Kit 方法构建黄尾鰤肝脏文库。制备好的文库用 Illumina Hi Seq 2500 进行测序。采用 Illumina 双末端测序(Paired-end, PE)方法进行高通量测序,原始测序结果去除制备文库时产生的接头序列、两端低质量序列和低度复杂序列。利用 Trinity (http://trinityrnaseq.sourceforge.net)软件对样品数据进行组装,获得 RNA-seq 高质量测序数据后,将所有测序读段通过从头组装生成重叠群(contig)和单一序列(singleton)。

1.4 功能注释、分类和代谢途径

将拼接组装后的序列,使用 Blast X 分别与 NR、String、Swissprot、KEGG、Pfam 数据库进行比对获得相应的注释信息。

1.5 SSR 和 SNP 位点搜索

利用 MISA 软件对黄尾鰤转录组中筛选得到 Unigene 进行简单重复序列(simple sequence repeats,SSR)位点分析,搜索标准为:单、二、三、四、五、六核苷酸基序(motif),至少重复次数分别为 10、6、4、4、4、4,对查找的 SSR 类型进行特征分析。利用 Samtools(http://samtools.sourceforge.net/)和 VarScanv. 2. 2. 7 (http://varscan. sourceforge.net/)软件寻找候选 SNP。

2 结果与分析

2.1 测序结果与数据组装

采用 Illumina HiSeq 2500 高通量测序技术对黄尾鰤肝脏转录组进行测序,总计产出 54 460 398 个reads 片段,去除低质量和含有接头的 reads 后,得到 50 702 046 条 clean reads,共计 4 942 900 890 个核苷酸,GC 含量平均值为 47.57%,Q20 为 98.82%,Q30 为 93.53%。数据表明,转录组测序数据量和质量都比较高,可为后续的数据组装提供很好的原始数据。

利用 Trinity 软件对所得的 reads 片段进行组装得到 53 527 条 Unigenes,平均长度为 946 bp,N50 为 1 925 bp(表 1)。其中,Unigenes 长度为 1～400 bp 的有 25 345 条,占总体的 47.35%,长度为 400～600 bp 的有 7170 条,占 13.40%,长度＞1000 bp的占 27.52%;转录本长度为 1～400 bp 的有 31 063 条,占总体的 36.92%;长度为 400～600 bp的有 10492 条,占 12.47%,长度＞1 000 bp 占 33.25%。

表 1 黄尾鰤转录组 Transcript 和 Unigene 数据组装质量统计

Tab.1 Data assembly for Transcript and Unigene in the transcriptome of *X. davidi*

长度范围/bp	Transcript		Unigene	
	数量	占比/%	数量	占比/%
1～400	31063	36.92	25345	47.35
401～600	10492	12.47	7170	13.40
601～800	6080	7.23	3689	6.89
801～1000	4675	5.56	2591	4.84
1001～1200	3842	4.57	2025	3.78
……	…	…	…	…

2.2 Unigene 功能注释、分类与代谢途径

2.2.1 *序列注释与相似性* 为了预测黄尾鰤 Unigene 功能,将获得的 53 527 个 Unigenes 分别与 NR、Swissprot、KEGG、Pfam、String 等公共数据库进行比对,进行 Unigenes 的序列相似性分析(表 2)。在 NR 数据库中注释成功的 Unigenes 的数量最多(49.59%),其后依次为 Swissprot(39.57%)、Pfam (30. 09%)、KEGG (27. 35%)、String (25.50%)。对该 5 组数据库进行拓扑分析,其结果如图 1,共有 7 989 条 Unigenes 在 5 个数据库中同时标注成功,占总 Unigenes 数的 14.92%;并且在以上 5 个数据库中至少 1 个数据库注释成功的 Unigenes 有 26 613 条,占总 Unigenes 数的 49.72%。

以 NR 数据库为例进行黄尾鰤 Unigenes 的序列相似性分析,26 406 条 Unigenes 在 NR 数据库中可以找到相似序列。在大于 4%相似序列匹配的近缘物种中,斑马鱼所占比例最高(79.41%),随后依次是墨西哥脂鲤(6.15%)、虹鳟(1.70%)、罗非鱼(1.10%)(图 2)。E 值在 0 区间内的 Unigenes 数量最多(18 178 个,占总体的 68.84%),匹配结果最高,E 值介于 0～10^{-30} 之间的 Unigenes 有3 517 个(13.32%),而 E 值介于 10^{-10} 到 10^{-5} 之间的 Unigenes 数量最少(842 个,3.19%);匹配序列相似度

表 2 黄尾鰤基因注释成功率

Tab.2 Success rate of gene annotation in *X. davidi*

数据库	Unigenes 数	比例/%
Nr	26 543	49.59
Swissprot	21 179	39.57
KEGG	14 642	27.35
Pfam	16 104	30.09
String	13 561	25.50
至少 1 个数据库注释成功	26 613	49.72
上述数据库中均注释成功	7 989	14.92
总 Unigenes	53 527	100

达到 80％以上的 Unigenes 数量最多，有22 290 个(84.41％)，相似度 40％～80％的 Unigenes 有 4 112 个(15.57％)，相似度低于 40％的 Unigenes 数量最少，仅有 4 个(0.02％)(图 3)。

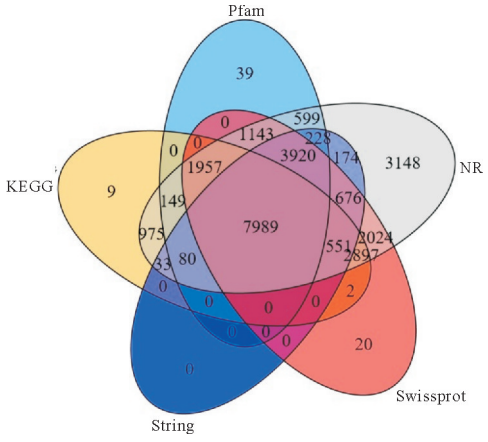


图 1　注释信息统计 Venn 图

Fig.1　Annotation information displayed on a Venn diagram

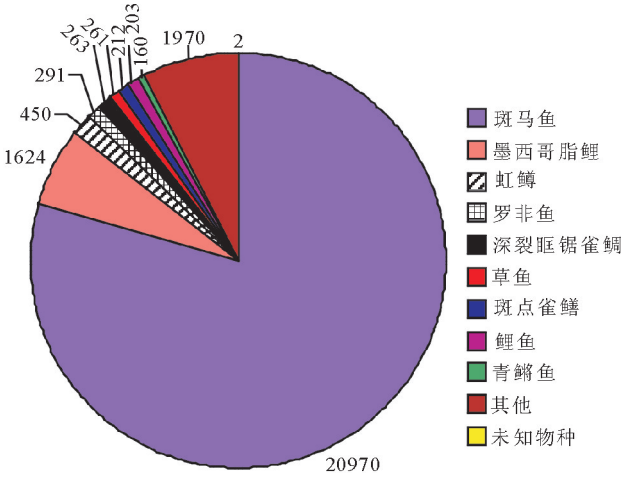
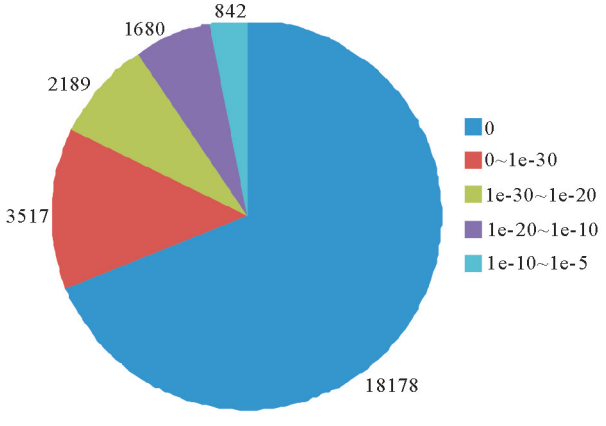


图 2　物种分类

Fig.2　Species distribution



2.2.2　Unigene 的 GO 分类　黄尾鰤共有 15 532 条 Unigene 在 GO 数据库中 3 大类 64 个功能中找到对应(图 4)。可以看出，33 339 个 GO 条目归属于细胞组分，18 964 个 GO 条目归属于分子功能，61 999个 GO 条目归属于生物学过程，这一分类结果显示了黄尾鰤肝脏基因表达谱的总体情况。其中生物学进程组中较大的是细胞进程(9 960 个)；细胞成分组中较大的是细胞组分(6 683 个)和细胞(6 683 个)；分子功能组中较大的是结合活性(8 478 个)和催化活性(5 810 个)。

2.2.3　Unigene 的 KEGG 通路注释　KEGG 是系统分析基因产物在细胞中的代谢途径以及基因产物功能的数据库。根据 KEGG 数据库的注释信息能进一步得到 Unigene 的 pathway 注释。结合 KEGG 数据库，对黄尾鰤的 14 642 条 Unigene 可能参与或涉及的代谢途径进行了统计分析。结果表明，黄尾鰤 Unigene 归属于 A(代谢)、B(遗传信息处理)、C(环境信息处理)、D(细胞过程)和 E(有机系统)五大类；其代谢途径主要包括信号转导、免疫系统、运输与代谢、能量代谢、脂类物质代谢、氨基酸代谢、蛋白折叠、转录与翻译等 33 类代谢通路(图 5)。其中注释数量最多的通路有癌症通路 (661 个)、P13K-Akt 信号通路(640 个)、粘合斑信号通路(541 个)、MAPK 信号通路(477 个)等；另外，根据免疫系统分类，共有 1 299 条 Unigenes 参与了 16 条代谢通路，包括趋化因子信号通路 (Chemokine signaling pathway)，白细胞经上皮移行机制(Leukocyte transendothelial migration)，T 细胞受体信号途径 (Tcell receptor signaling pathway) 和 Toll 样受体信号途径 (Toll-like receptor signaling pathway)(图 6)。在这些免疫系统代谢通路中，参与到趋化因子信号通路的 Unigenes 数量最多。

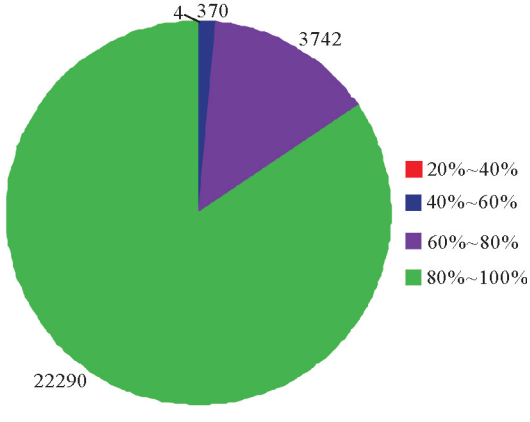


图 3　比对结果 e-value 分布及相似度分布

Fig.3　E-value and similarity distributions

图4 黄尾鲷 Unigene 的 GO 分类

Fig.4 GO functional categories of *X. davidi* Unigenes

2.2.4 Unigene 的 COG 注释结果 在黄尾鲷 Unigene 数据库的 COG 功能注释中，有 7 737 条 Unigene 被归纳到 COG 的 26 个功能类别中（图 7；A：RNA 加工与修饰，B：染色体结构与活力，C：能量生成和转换，D：细胞周期控制、细胞分裂、染色体区分，E：氨基酸运输和代谢，F：核苷酸运输和代谢，G：碳水化合物运输和代谢，H：辅酶运输和代谢，I：脂质运输和代谢，J：翻译、核糖体结构和代谢，K：转录，L：复制、重组和修复，M：细胞膜生物合成，N：细胞运动，O：翻译后修饰，蛋白质折叠和分子伴侣，P：矿物质运输和代谢，Q：次生代谢物合成、运输和代谢，R：一般功能基因，S：未知功能，T：信号传导机制，U：细胞内转运、分泌和小泡运输，V：防卫机制，W：胞外结构，Y：核结构，Z：细胞构架）。其中参与

到一般功能的 Unigenes 数目最多，共 1 855 条，占 23.78%；其次是复制、重组和修复以及信号传导机制相关的 Unigenes 较多，分别有 748 条和 737 条，各占 9.67%和 9.52%。参与翻译、核糖体结构和氨基酸运输和代谢、脂质运输和代谢等的 Unigenes 较为普遍，然而细胞核结构、细胞外基质结构、细胞膜生物合成、细胞运动以及 RNA 加工和修饰等所占的比例都很少，都低于 1%；另外，还发现有 35 条 Unigenes 参与到免疫防御机制中，其出现可能与黄尾鲷的免疫防御有关，需进一步进行研究。

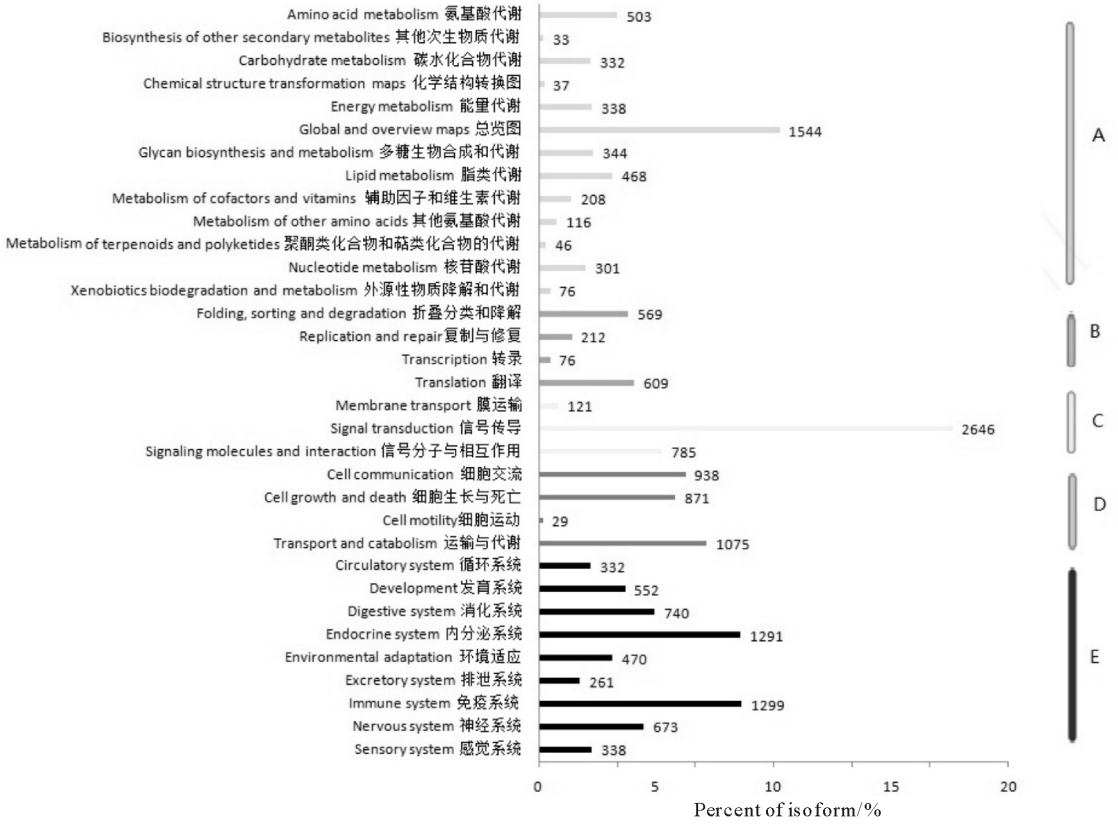


图 5　KEGG 注释统计

Fig.5　KEGG annotation statistics

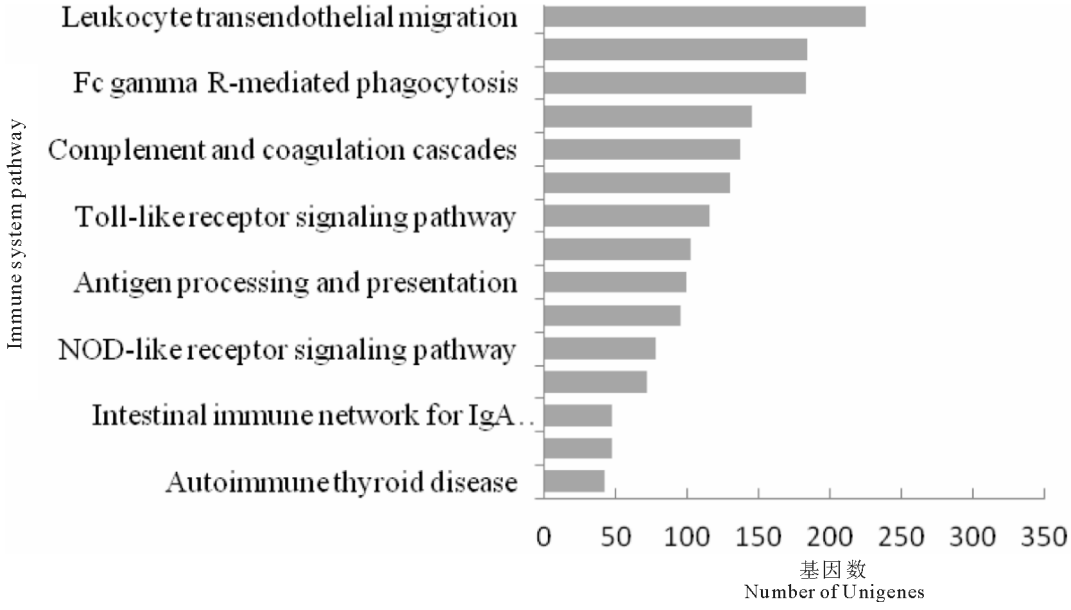


图 6　参与到免疫系统代谢通路的 Unigenes 分布

Fig.6　Classification of Unigenes based on the immune system pathway

2.2.5　SNP和SSR统计分析　本次试验表明，利用Samtools和VarScan软件检测到了98 826个预测的SNP位点，其中包括64 396个转换和34 430个颠换。不同转换形式(A/G和C/T)的个数基本相似，但不同形式的颠换(A/T、A/C、G/T和G/C)的个数存在一定的差异(表3)。

根据SNP所在ORF的位置对SNP位点进行分类，在所有预测的SNP位点中，能够被注释的SNP位点总计有42 128个，其中位于编码区的有28 075个，位于5′和3′UTR区域的SNP有14 055个，由于NCBI中未见黄尾鲷的基因数据，因此有56 698个SNP位点没有被注释。

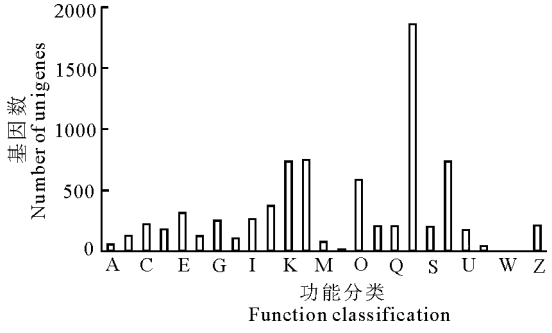


图7　COG分类统计结果

Fig.7　Results of COG classification

表3　SNP统计

Tab.3　SNP statistics

类型		数目	频率/kb	占比/%
转换	C/T	32303	0.31	32.69
	A/G	32093	0.30	32.47
颠换	A/T	11809	0.11	11.95
	A/C	8598	0.08	8.70
	T/G	8538	0.08	8.64
	C/G	5485	0.05	5.55
	总计	98826	0.95	

利用MISA软件在黄尾鲷53 527条Unigenes的13 220条中共检测到18 119个SSR位点，总体上SSR出现频率为33.85%。Unigenes序列中各种类型的SSR出现的频率不同(表4)，单碱基、二碱基和三碱基重复类型占优势，分别占总SSR的56.40%、19.51%和21.30%，其他3种重复类型所占比例相对较少，四核苷酸重复占2.4%，五核苷酸重复占0.31%，六核苷酸重复占0.07%。所有SSR中以≥13次重复数目的SSR最多，占24.14%，10次重复数目的SSR占19.24%，4次重复数目的SSR占15.14%(表4)。另外，还发现单核苷酸酸重复基元中，A/含量最高；二核苷酸重复基元中，AC/GT含量最高，其次是AG/CT。在所有的重复基元类型中CG/CG发生频率最低。

表4　黄尾鲷EST-SSR重复类型及数量

Tab.4　Repeat type and number of EST-SSR in *X. davidi*

重复基元长度	重复次数										总计	占比/%
	4	5	6	7	8	9	10	11	12	≥13		
单核苷酸	-	-	-	-	-	-	3071	1763	1012	4374	10220	56.40
二核苷酸	-	-	1228	709	505	560	416	114	3	0	3535	19.51
三核苷酸	2373	822	444	213	8	0	0	0	0	0	3860	21.30
四核苷酸	302	126	7	0	0	0	0	0	0	0	435	2.40
五核苷酸	55	1	0	0	0	0	0	0	0	0	56	0.31
六核苷酸	13	0	0	0	0	0	0	0	0	0	13	0.07
总　计	2743	949	1679	922	513	560	3487	1877	1015	4374	18119	100
占比/%	15.14	5.24	9.27	5.09	2.83	3.09	19.24	10.36	5.60	24.14		

3　讨论

为了获得黄尾鲷的转录组信息，本研究对其进行了高通量转录组测序，通过序列的预处理Trinity软件组装、拼接、转录本功能注释等步骤，得到了数据量为5.46G的转录组信息。大量的黄尾鲷转录本信息，可为其今后的疫学、基因组学、基因克隆以及分子标记辅助育种提供重要依据(车荣波，2015)。

3.1　转录组质量

从转录组质量来看，本研究利用Trinity软件对所得的reads片段进行组装得到53 527条Unigenes，长度为201～16 678 bp，平均长度为946 bp，长度＞1 000 bp Unigenes所占比例达33.25%。其中N50为1 925 bp，N50值越大说明组装得到的长片段越多，组装效果越好(车荣波，2015)。碱基Q30为93.53%，当Q30值在80%以上就认为测序质量非常可靠。因此，本文构建的转录组数据库为后续黄尾鲷基因克隆及功能验证提供了基础数据。

3.2　数据库注释

从数据库注释上看，黄尾鲷在NR、Swissprot、KEGG、Pfam、String等公共数据库中均得到注释，对深入了解基因功能有重要帮助。但从数量上看，

各类数据库中可注释的基因占比例较少，在 25.5%～49.59%，50.28%的 Unigene 可能是由于序列片段过短而、非编码序列或者是特有的新基因等未与公共数据库中的序列匹配上(赵刚等，2016)，说明国际公共数据库中收录的黄尾鰤数据较少。但各数据注释结果可以帮助了解黄尾鰤更多的生物学信息，了解基因的分子功能、免疫机制、所参与的生物学过程、所处的代谢途径和信号通路等。

3.3　分子标记

本研究挖掘出 18 119 个 SSR 和 98 826 个 SNP 位点信息，SNP 和 SSR 是利用转录组数据开发最多的两类标记(Weeks et al，2001；Hinomoto et al，2005)。目前，已经有很多研究者通过转录组数据库发掘分子标记(Zhai et al，2013；袁文成，2015；Li et al，2016；全迎春等，2016；龚诗琦等，2016)，与传统开发分子标记的方法相比，具有成本低、时间短、信息量等优点。本研究将为黄尾鰤后续的多态性检测、群体遗传多样性分析以及分子鉴定等方面打下基础，同时还可用于鰤亚科鱼类遗传多样性分析、品种鉴定、遗传图谱构建、重要性状辅助选择等研究，有助于促进鰤亚科鱼类分子生物学的发展。

本研究首次采用 Illumina HiSeq 2500 高通量测序技术对黄尾鰤转录组进行从头测序和无参考基因组的情况下进行从头组装。除了对黄尾鰤转录组的基本生物学信息学进行了分析，还得到了大量新功能基因和分子标记。尤其是那些免疫相关基因可以为黄尾鰤的免疫系统和疾病防御提供大量的资源。另外，在本研究中获得分子标记可以用于今后黄尾鰤遗传多样性分析、种质鉴定、基因连锁图谱构建以及基因功能定位分析等。

参考文献

车荣波，2015. 基于转录组数据的鱼分子标记筛选及基因差异表达分析[D]. 舟山：浙江海洋大学.

冯晓宇，杨仲景，李行先，等，2005. 黄尾密鰤人工繁殖及鱼苗培育[J]. 杭州农业科技，(1)：19－21.

龚诗琦，王志勇，肖世俊，等，2016. 黄姑鱼转录组 SSR 的开发与验证[J]. 集美大学学报(自然科学版)，21(4)：241－246.

黄邦星，2006. 黄尾密鰤生物学特性及养殖技术[J]. 水产养殖，27(3)：32－34.

凌志勇，2002. 黄尾密鰤人工孵化与养殖技术[J]. 河南水产，(1)：23.

全迎春，马冬梅，白俊杰，等，2016. 大口黑鲈转录组 SNPs 筛选及其与生长的关联分析[J]. 水生生物学报，40(6)：1128－1134.

叶富良，1987. 东江黄尾密鰤的生物学及渔业利用[J]. 淡水渔业，(4)：9－11.

袁文成，2015. 基于转录组测序的翘嘴鲌微卫星标记的开发及 MCH class Ⅰ基因的克隆表达[D]. 苏州：苏州大学.

赵 刚，龚全，刘亚，等，2016. 基于 Illumina 高通量测序的岩原鲤转录组分析[J]. 西南农业学报，29(7)：1743－1749.

Hinomoto N，Maeda T，2005. Isolation of microsatellite markers in *Neoseiulus womersleyi* Schicha (Acari：Phytoseiidae) [J]. Journal of the Acarological Society of Japan，14(1)：25－30.

Li C Y，Chiang T Y，Chiang Y C，et al，2016. Cross-species，amplifiable EST-SSR markers for amentotaxus species obtained by next-generation sequencing [J]. Molecules，21(1)：67－76.

Weeks A R，Marec F，Breeuwer J A，2001. A mite species that consists entirely of haploid females[J]. Science，292：2479－2482.

Zhai L，Liu L，Zhu X，et al，2013. Development，characterization and application of novel expressed sequence tag-simple sequence repeat (EST-SSR) markers in radish (*Raphanus sativus* L.)[J]. African Journal of Biotechnology，12(9)：921－935.

(责任编辑　万月华)

Transcriptome Analysis of *Xenocypris davidi* Bleeker Based on RNA Sequencing

ZHANG Yan-ping[1], ZHANG Hai-xin[1], CUI Cui[2], FU Yi-long[1], LIU Zhi-fang[2], FAN Hong-chao[2]

(1.Fisheries Research Institute of Jiangxi Province, Nanchang 330000, P.R.China;
2.Fisheries Research Institute of Pingxiang, Pingxiang 337000, P.R.China)

Abstract: *Xenocypris davidi* Bleeker is a species considered excellent for aquaculture because of its high nutritional quality, delicious taste and good adaptability. While there is a high market demand for *X. davidi*, wild populations have rapidly declined because of over fishing and led to rapid development of artificial culturing. However, disease outbreaks in intensive aquaculture ponds and loss of genetic diversity are problematic. In this research, the total RNA of *X.davidi* Bleeker was extracted and a genetic library was established for the species. The transcriptome of *X. davidi* was developed and functional genes, particularly those related to the immune system, were identified to lay a solid foundation for molecular biology research. The *X. davidi* transcriptome was sequenced *de novo* on the Illumina platform. After data cleaning and testing, 50 702 046 high-quality reads were obtained and 53 527 unigenes were assembled. After searching against NR, String, Swissprot, KEGG and Pfam databases, 26 613 unigenes were successfully annotated: 15 532 unigenes were classified into 64 functional categories under three GO ontologies; 7 737 unigenes were assigned to COG, grouped into 26 functional categories; and 14 642 unigenes with significant matches in the database were assigned to 33 KEGG pathways. According to an immune system classification, 1 299 unigenes are involved in 16 metabolic pathways. A total of 98 826 SNPs (64 396 transitions and 34 430 transversions) and 18 119 SSRs were detected in the 53 527 unigenes. This study provides a foundation for future genetic research and molecular marker-assisted breeding of *X. davidi*.

Key wards: *Xenocypris davidi* Bleeker; transcriptome; high-throughput sequencing; function annotation