水产学报, 2024, 48(11): 119604

JOURNAL OF FISHERIES OF CHINA

DOI: 10.11964/jfc.20220613562

Science Press

苏氏圆腹鲃全长转录组测序

宦　章[1]，李东宇[2]，李伟豪[1]，高　进[3]，
王中铎[1]，潘　志[2]，董忠典[1,4*]

(1. 广东海洋大学水产学院，南海水产经济动物增养殖广东普通高校重点实验室，广东 湛江　524088；
2. 北京市水产技术推广站，北京　100176；
3. 海南热带海洋学院，热带海洋生物资源利用与保护教育部重点实验室，海南 三亚　572022；
4. 广东海洋大学，广东省水产动物病害防控与健康养殖重点实验室，广东 湛江　524088)

摘要：为深入解析苏氏圆腹鲃的遗传信息和开展基因功能研究，本实验分别提取性成熟苏氏圆腹鲃脑、鳃、心脏、肝脏、脾脏、头肾、胃、肠、性腺和肌肉组织的总 RNA，等质量混合为一个样本后，利用 PacBio 高通量测序平台单分子实时测序技术对其进行了测序分析，获得了性成熟苏氏圆腹鲃各组织的全长转录组信息。结果显示，在苏氏圆腹鲃中共获得 1 487 336 条高质量 reads，平均长度和 N50 分别为 83 592 和 162 901 bp；校正后共获得 1 005 955 条循环一致序列 (circular consensus sequencing，CCS)，过滤后共鉴定出 667 973 条含有 polyA 结构的全长非嵌合序列 (full-length non-concatemer，FLNC) 序列，平均长度和 N50 分别为 2 057 和 2 359 bp。614 078 条 (91.93%) FLNC 用于基因和转录本的注释，鉴定到的 19 835 个已知基因对应 80 915 个转录本，9 348 个新基因对应 9 954 条新转录本，预测到 50 311 个开放阅读框、79 922 个可变剪接、18 个融合基因和 20 215 个选择性多聚腺苷酸化位点。新基因在 NR、GO、KEGG、KOG 和 SwissProt 数据库中分别有 3 912、2 385、2 167、81 和 1 520 个获得注释。另外，还预测到 4 624 个长链非编码 RNA，调控 32 283 个靶 mRNA。研究表明，通过全长转录组测序数据及功能注释分析，丰富了苏氏圆腹鲃的遗传资源信息。本研究可为进一步开展苏氏圆腹鲃生物学特性、基因功能研究提供基础。

关键词: 苏氏圆腹鲃；全长转录组；基因功能；可变剪切；长链非编码 RNA

中图分类号: Q 786；S 965.1　　**文献标志码:** A

苏氏圆腹鲃(*Pangasius sutchi*) 是东南亚地区重要的淡水经济鱼类，生长速率快，无鳞和肌间刺，肉质白嫩、营养丰富，抗病性强且价格低廉，深受养殖户和消费者青睐[1-2]。另外，苏氏圆腹鲃的鱼皮含有丰富的胶原蛋白，常被用来提取胶原蛋白制备生物医药材料[3-4]。1978 年苏氏圆腹鲃被从泰国引入中国，于 1997 年突破人工繁殖，随后在广东、广西和海南等地获得大力推广[5]。目前对苏氏圆腹鲃的研究主要集中在人工繁殖和养殖模式探索、营养与饲料开发，抗病性和鱼产品加工等方面[2, 6-12]。但是，苏氏圆腹鲃基础生物学，特别是分子生物学研究尚处于起步阶段[13-14]。

转录连接了承载遗传信息的基因与发挥生物功能的蛋白质，是基因结构和功能研究的基础和

收稿日期：2022-06-24　　**修回日期：**2023-01-03

资助项目：国家自然科学基金 (31201996)；广东海洋大学“南海学者计划”青年人才项目 (QNXZ201903，201807)；广东海洋大学博士启动项目

第一作者：宦章 (照片)，从事水产动物健康养殖与水生生态毒理研究，E-mail：2531184897@qq.com

通信作者：董忠典，从事鱼类分子育种及水生生态毒理研究，E-mail：zddong@gdou.edu.cn

起点。因此，在进行基因结构和功能的研究中，获得基因的完整转录本信息尤为重要，全长转录本测序技术在获取物种基因完整转录本上优势巨大[15]，美国太平洋生物技术公司 (Pacific Biosciences) 开发的基于单分子实时 (SMRT) 测序技术的全长转录组测序是一种直接对目标 RNA 进行测序的高通量测序技术。该技术集高通量、检测速率快和超长读数优势，避免了 PCR 冗余和碱基替换等常见 PCR 错误，大大提高了 RNA 测序的准确度，测序精准度可达到 99.9%[16]。全长转录组测序在可变剪接事件分析、基因差异表达水平分析、基因功能注释、新基因预测、SNP 位点筛选和选择性多聚腺苷酸化位点分析中发挥重要作用[17-19]。因此，本研究采用 SMRT 测序技术对性成熟苏氏圆腹鲀进行全长转录组测序，通过生物信息学方法进行了基因结构分析和功能注释等，旨在丰富和挖掘苏氏圆腹鲀的遗传资源，服务于苏氏圆腹鲀的基因功能、基因表达调控和进化关系等多方面基础与应用研究。

1　材料与方法

1.1　实验材料

实验所用的苏氏圆腹鲀由北京市水产技术推广站南繁育种中心 (海口) 提供，随机挑选体长约 55 cm、体重约 2.3 kg，体表完整无伤痕、活力好的苏氏圆腹鲀雌雄各 1 尾，解剖取脑、鳃、心脏、肝脏、脾脏、头肾、胃、肠道、性腺和肌肉组织，迅速放入液氮中保存，随后移至 −80 °C 冰箱保存，用于后续的 RNA 提取。本研究获得了广东海洋大学实验动物管理和使用伦理委员会批准 (审批号：20220103)，实验过程中操作人员严格遵守广东海洋大学实验动物管理和使用伦理委员会伦理规范，并按照广东海洋大学伦理委员会制定的规章制度执行。

1.2　RNA 提取、cDNA 文库构建与测序

参照 TRIzol 试剂 (Invitrogen，美国) 操作说明，提取雌雄苏氏圆腹鲀各组织的总 RNA，用 DNase Ⅰ [宝日医生物技术 (北京) 有限公司] 去除基因组 DNA 污染，通过琼脂糖凝胶电泳检测总 RNA 是否存在杂质或基因组 DNA 的污染，Nanodrop 2000 检测 OD 值 260/280 比值确定 RNA 浓度。Qubit 精确定量 RNA 浓度。Agilent 2100 精确检测总 RNA 的完整性，将符合测序质量要求的各组织总 RNA 等质量混合成一个总 RNA 样本池 (本研究目的是获得苏氏圆腹鲀的全长转录组信息，辅助基因组注释和基因功能研究，为尽可能获得物种的转录组信息，选择了混合雌雄个体不同组织总 RNA 的策略)。进行 cDNA 文库制备和全长转录组测序：使用 Clonetech SMARTer PCR cDNA Synthesis Kit 合成 mRNA 的全长 cDNA，经磁珠纯化后去除部分 1 kb 以下的 cDNA；对全长 cDNA 进行末端修复，连接 SMART 接头；使用核酸外切酶消化未连接接头的片段，再次使用磁珠进行纯化，获得测序 cDNA 文库；cDNA 文库经 Agilent 2100 检测合格后在 PacBio 测序平台上进行全长转录组测序，原始数据已提交至国家基因库生命大数据平台 (CNGBdb) (CNX0566612)。

1.3　全长转录组测序原始数据预处理

将 SMRT 技术测序获得的 polymerase read 去掉接头获得子序列 subread，同一个 polymerase read 得到的 subread 经过自我校正形成循环一致序列 (circular consensus sequencing，CCS)，根据是否含有嵌合体序列、5′和 3′端测序引物，将 CCS 分为 3 类：全长 (full-length sequence, FL) 序列、全长非嵌合 (full-length non-concatemer，FLNC) 序列和带 polyA 的 FLNC 序列。其中带有 polyA 的 FLNC 序列用于后续分析。

1.4　FLNC 序列校正、基因组比对分析和全长率评估

通过多次测序对 FLNC 序列进行校正，随后借助 LoRDEC[20] 软件用本实验室获得的苏氏圆腹鲀2 代转录组测序数据 (NCBI：PRJNA902092) 对 FLNC 序列进一步校正。用 GMAP[21] 软件将校正后的 FLNC 序列比对到苏氏圆腹鲀参考基因组 (NCBI Genome ID: 310915) 上。根据每条 FLNC 序列的比对位置进行基因和转录本 (isoform) 的鉴定，以参考基因注释的多外显子转录本为参考，筛选测序结果中与其方向一致、overlap 大于 20% 的 isoform，与对应的基因组注释转录本进行剪切位点比较。依据 isoforms 序列 5′至 3′方向第一个剪接位点的对应关系，如果注释转录本的第一个剪切供体位点 (即第一个内含子的 5′端剪接位点) 被测序得到的 isoform 的剪切供体位点包括，即认为该测序 isoform 为全长 isoform，对应的 FLNC

视为全长 FLNC。

1.5 新基因及新 isoform 注释和功能预测

利用测序得到的基因和 isoform 与参考基因组比对，可以确定检测到的已知注释基因和 isoform，以及鉴定到的新基因和新 isoform。相对参考基因组注释基因，测序得到的基因与已注释基因没有 overlap 或 overlap 小于 20% 的，或与已注释基因 overlap 大于 20%，但基因方向不一致的判定为新基因。测序得到的 isoform 与参考基因组注释基因转录本比较，如果测序得到的 isoform 存在 1 个或多个新的剪切位点，或参考基因组注释基因转录本与测序得到的 isoform 不同时为单外显子，则认为是一个新的 isoform。用 Diamond[22] 软件将苏氏圆腹鲀上述新转录比对到 NR、GO[23]、COG/KOG[24]、KEGG[25-27] 和 Swiss-Prot[28] 数据库上进行功能注释 (E-value≤1×10^{-5})。

1.6 长链非编码 RNA 分析

将新基因的 isoform、已知基因的新 isoform 序列与 NR、KEGG、KOG 和 Swiss-Prot 数据库比对，过滤掉可能的编码序列。用 CNCI、CPC2、CPAT 和 PLEK 软件对剩余序列的编码潜能进行评估，其交集作为最终的长链非编码 RNA (long non-coding RNA, lncRNA) 预测的结果。

1.7 新 isoform 开放阅读框预测

过滤掉已知基因新 isoform、新基因 isoform 中被预测为 lncRNA 的部分，使用 TransDecoder[29] 软件预测新 isoform 编码区序列及其对应的氨基酸序列，获得开放阅读框 (ORF) 。

1.8 可变剪接、融合基因和选择性多聚腺苷酸化位点分析

基因经转录形成的 mRNA 前体通过不同的剪接方式产生不同的 mRNA 剪接异构体，这个过程称为可变剪接。本实验使用 Astalavista 3.2[30] 软件对测序得到的 isoform 可变剪接事件进行分类和统计。融合基因是指将两个或多个基因的编码区首尾相连，置于同一套调控序列控制之下构成的嵌合基因。导致两个基因发生融合的机制包括基因结构变异、转座或者基因转录后的反式剪接等。mRNA polyA 位点的改变是一类重要的 RNA 转录后调控修饰，通过检测 FLNC 序列在基因组上 3′ 末端序列的比对位置，可以直接得到可能的 mRNA 多聚腺苷酸化位点。本研究基于已发表的全长转录组选择性多聚腺苷酸化位点检测方法[31]，使用 Tapis 软件 (默认参数) 对每个表达基因的 polyA 位点个数以及每个位点对应的转录本个数进行检测。

2 结果

2.1 SMRT 测序数据统计

本研究共获得了 1 487 336 条 polymerase reads (124.33 Gb)，去掉接头序列后得到 58 868 917 条 subreads (119.78 Gb)，subreads 平均长度和 N50 分别为 2 035 和 2 386 bp (表 1)。校正后共计获得 1 005 955 条 CCS 序列，过滤后共鉴定出 667 973 条含有 polyA 的 FLNC 序列用于后续分析，平均长度和 N50 分别为 2 057 和 2 359 bp (图 1，表 1)。

表 1 苏氏圆腹鲀全长转录组序列类型统计

Tab. 1 Sequence type statistics of full-length transcriptome of *P. sutchi*

类型 type	数量/条 total no.	最小长度/bp min length	平均长度/bp average length	最大长度/bp max length	N50/bp
polymerase reads	1 487 336	51	83 592	344 074	162 901
subreads	58 868 917	51	2 035	246 491	2 386
CCS	1 005 955	61	2 271	14 729	2 591
FLNC (含Ploy A) FLNC (includes ployA)	667 973	50	2 057	8 558	2 359

2.2 FLNC 校正和参考基因组比对

使用苏氏圆腹鲀二代转录组数据对上述 667 973 个 FLNC 序列进行校正，共计 665 832 个 FLNC 被校正。228 281 个在校正后序列质量优于校正前 (表 2)。用 GMAP 将 FLNC 与苏氏圆腹鲀参考基因组进行比对，最终获得 614 078 条 (91.93%) FLNC 可用于基因和 isoform 注释 (表 3)。

2.3 基因结构、新基因及新 isoform 注释

根据每条 FLNC 序列在参考基因组上比对位置，共鉴定到 29 183 个基因和 90 869 条 isoform，进一步与苏氏圆腹鲀的参考基因组注释的基因和转录本进行比对，结果显示，苏氏圆腹鲀全长转录组测序鉴定到 19 835 个已获得注释的基因，对应 80 915 个 isoform，另外鉴定到 9 348 个新基因和 9 954 个新 isoform。

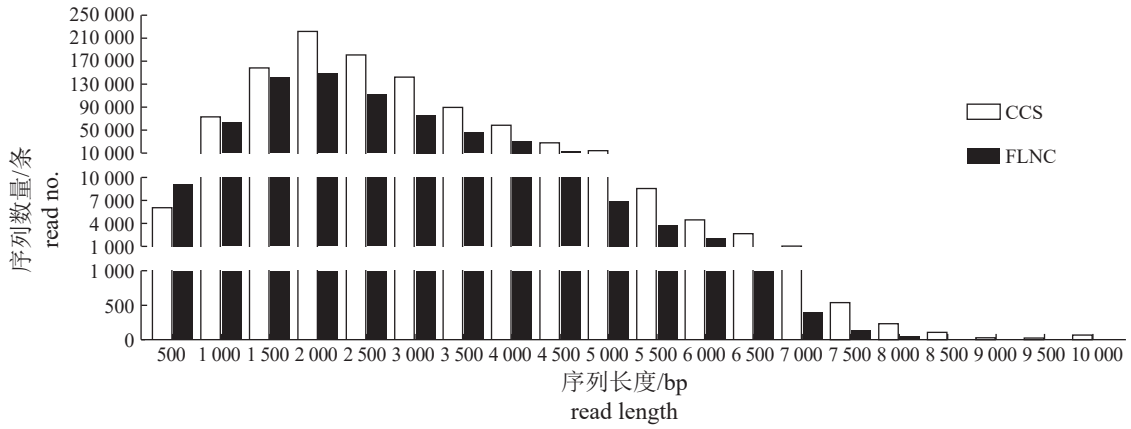


图 1　苏氏圆腹鲃全长转录组 CCS 和 FLNC 序列长度分布

横坐标为序列长度范围，数值 500 表示序列长度为 1~500 bp，1 000 表示长度为 501~1 000 bp，后续数值保持相同规律；纵坐标表示对应长度的序列数量。CCS 表示循环一致序列，FLNC 表示全长非嵌合序列。

Fig. 1　Distribution of CCS and FLNC sequence length in the full-length transcriptome of striped *P. sutchi*

The X-axis indicates the length range of the sequence, the value of 500 indicates the length of the sequence in the range of 1-500 bp, the value of 1 000 indicates the length in the range of 501-1 000 bp, and the subsequent values maintain the same rule; Y-axis values indicate the number of sequences. CCS represents circular consensus sequencing and FLNC represents full-length non-concatemer sequences.

表 2　序列纠错前后全局 PID 统计

Tab. 2　Global PID statistics before and after sequencing error correction

组别 groups	FLNC数/条 FLNCs no.	校正前PID平均值/% mean entire PID of precorr	校正后PID 平均值/% mean entire PID of postcorr
仅在校正前识别　only precorr mapped	27	40.03	—
仅在校正后识别　only postcorr mapped	508	—	82.44
校正前PID大于校正后PID　pre PID > post PID	91 840	97.73	95.49
校正前PID等于校正后PID　pre PID = post PID	345 684	98.45	98.45
校正前PID小于校正后PID　pre PID < post PID	227 773	97.86	98.44
总数　total	665 832	98.07	98.02

注：“—”表示未计算。

Notes: “—” indicates no value.

表 3　参考基因组比对结果分类统计表

Tab. 3　Statistical table of classification of reference genome comparison results

比对类型 feature	校正前每种比对类型FLNC序列数/条(占比/%) no. and proportion of alignment precorr FLNC sequences	校正后每种比对类型FLNC序列数/条(占比/%) no. and proportion of alignment postcorr FLNC sequences	合并校正前后每种比对类型FLNC序列数/条(占比/%) no. and proportion of alignment merge FLNC sequences
在设定阈值范围内比对不上参考基因组 FLNC that unmapped the reference genome within the threshold range	2 649 (0.40)	2 168 (0.32)	2 141 (0.32)
最高PID的比对位置有多个 FLNC that the highest PID has multiple mapped locations	347 (0.05)	355 (0.05)	322 (0.05)
最高PID的比对位置只有一个，且满足局部PID<97%，或者全局PID<95% FLNC that the highest PID only one mapped location, and the party PID <97% or entire PID <95%	56 586 (8.47)	67 665 (10.13)	51 432 (7.70)
最高PID的比对位置只有一个，且满足局部PID>97%，同时全局PID>95% FLNC that the highest PID only one mapped location, and the party PID >97% and entire PID >95%	608 391 (91.08)	597 785 (89.49)	614 078 (91.93)

2.4　新基因功能注释

将鉴定到的 9 348 个新基因比对到 NR、GO、KEGG、COG/KOG 和 Swiss-Prot 数据库上，分别有 3 912 (占新基因总数 9 348 的 41.85%)、2 385 (25.51%)、2 167 (23.18%)、81 (0.87%) 和 1 520 (16.26%) (图 2)，另有 5 395 个新基因未获得注释。对注释到的新基因进行 GO、KEGG 和 KOG 富集分析，在 GO 结果中新基因注释到生物学过程、细胞成分和分子功能三大类别，其中细胞过程 (1 309)、细胞结构体 (1 351)、结合 (1 524) 分别是上述 3 种类型中新基因被注释最多的条目 (图 2)。

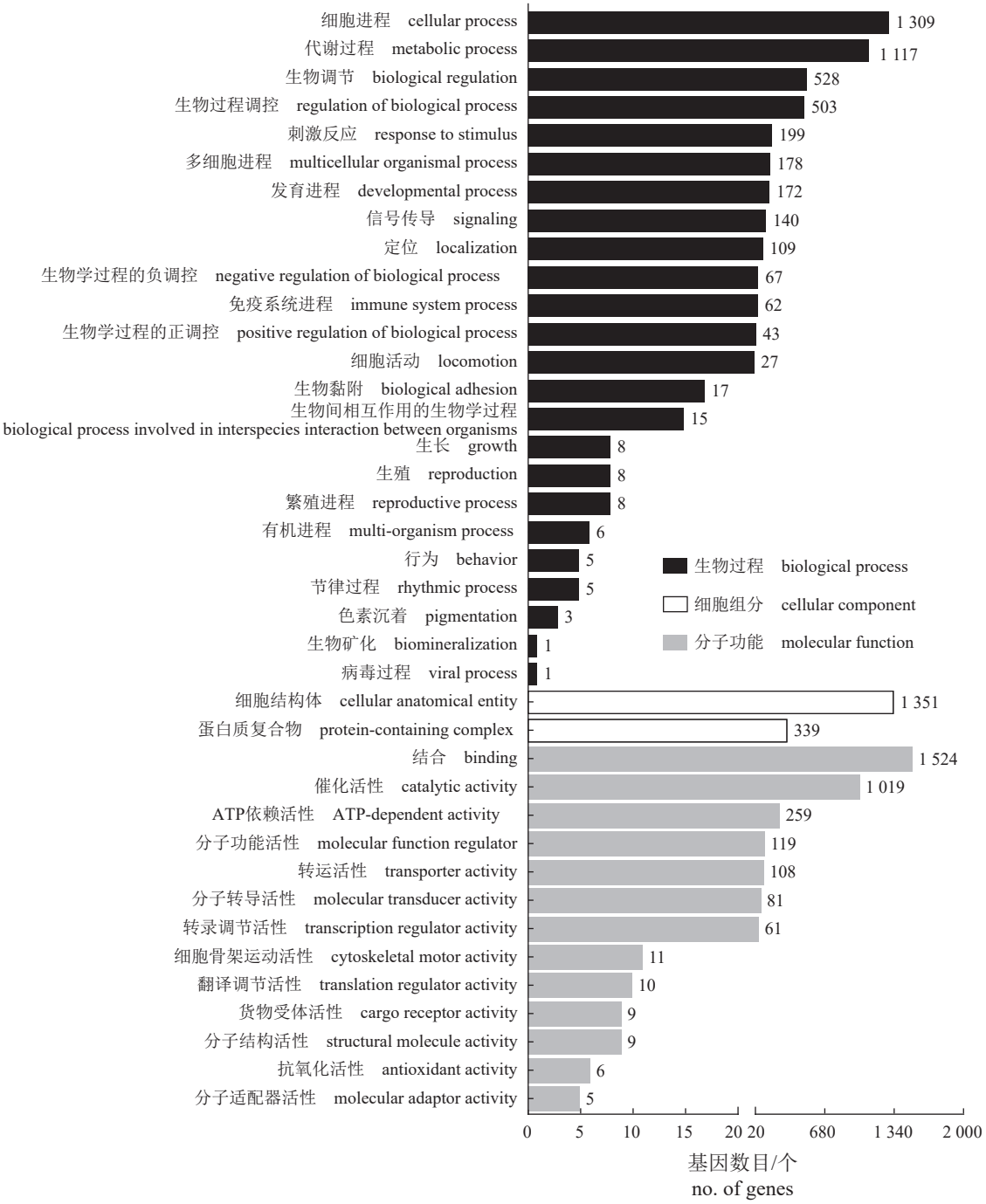


图 2　苏氏圆腹鲀新基因 GO 功能注释

纵坐标表示 GO 条目名称，横坐标表示条目对应的基因数目。

Fig. 2　Gene ontology (GO) assignment of new genes of *P. sutchi*

The y-axis represents the GO term, and the x-axis shows the number of new genes with them.

KEGG 注释包含细胞过程、环境信息处理、遗传信息处理、代谢和有机系统 5 大类。其中各类中包含新基因最多的通路依次为真核细胞群落 (106)，信号转导 (276)，折叠、分类和降解 (79)，氨基酸代谢 (63) 和内分泌系统 (197) (图 3)。新基因在 KOG 数据库共被注释到 19 个功能类别的簇。包含新基因最多的是主要功能预测 (14)，随后是信号转导机制 (13) 和转录 (10)，而最少的是辅酶转运和代谢 (1) (图 4)。

2.5　基因结构分析

lncRNA 鉴定及靶基因的功能富集　本研究中苏氏圆腹鲎全长转录组中共有 4 624 个新 isoform 被 CPC2、CNCI、PLEK 和 CPAT 共同预测为 lncRNA (图 5-a)，包含内含子 (intronic) lncRNA (1 915，41.41%)，正义 (sense) lncRNA (1 368，29.58%)、基因间区 (intergenic) lncRNA (848，18.34%) 和反义 (antisense) lncRNA (493，10.66%)。lncRNA 和 mRNA 的长度分布密度结果表明，测序长度低于 1 800 bp 的片段多为 lncRNA，最大的 lncRNA 长度不超过 7 295 bp (图 5-b)。

进一步对苏氏圆腹鲎的 LncRNA 调控的 32 283 个靶 mRNA 进行 GO、KEGG 富集分析。GO 结果表明，生物学过程中，靶 mRNA 主要富集在细胞过程 (12 084)、生物调节 (6 391 条)、生物过程调节 (5 903) 和代谢过程 (5 758)，而与解毒 (4) 相关的最少；细胞组分中，靶 mRNA 主要富集在细胞结构体 (18 034) 和蛋白质复合物 (2 541) 中；分子功能中，富集到结合相关的靶 mRNA 数量最多 (12 772)，其次是催化活性 (8 193)、分子载体活性富集的靶 mRNA 数量最少 (39) (图 6)。KEGG 结果表明，靶 mRNA 在细胞过程中，主要富集在运输与代谢 (1 437) 相关通路；环境信息处理中，主要富集在信号传导 (4 165) 相关的通路中；遗传信息处理中，主要富集在折叠、分类和降解

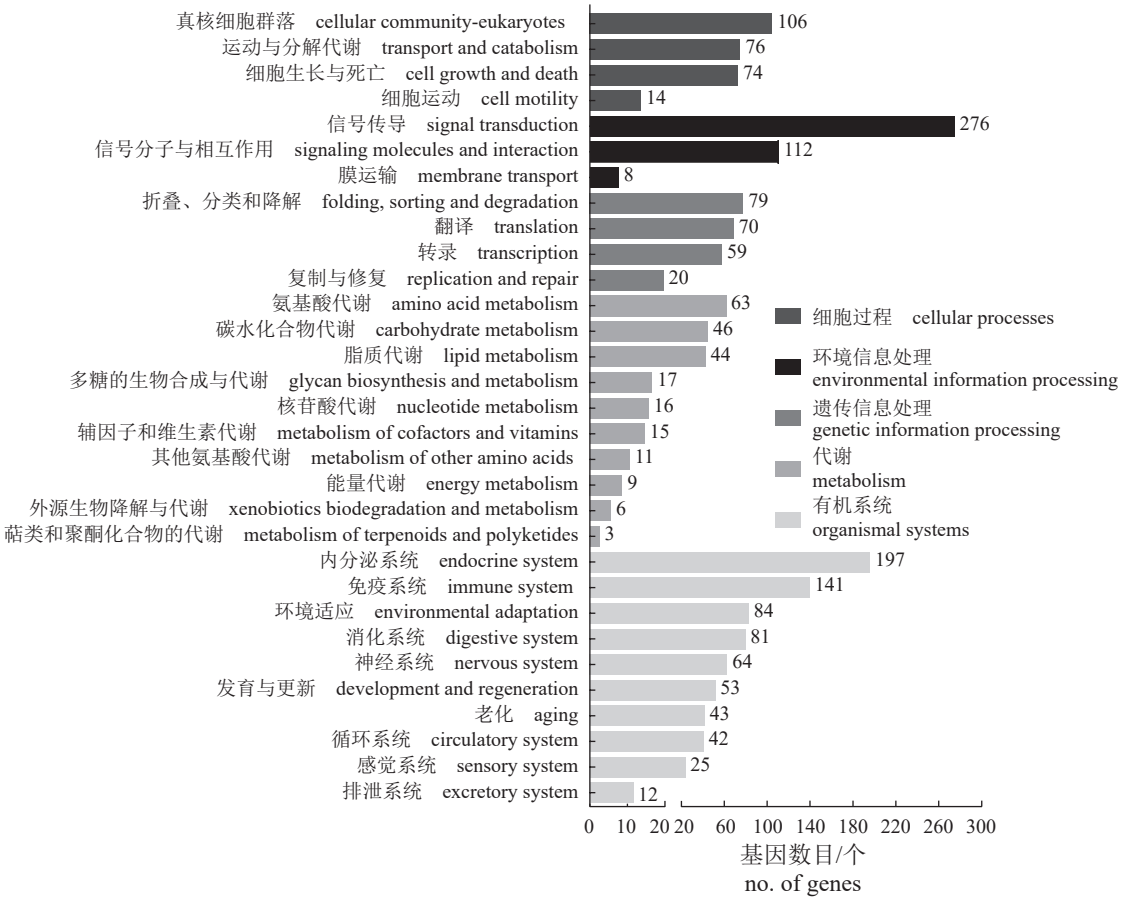


图 3　苏氏圆腹鲎新基因 KEGG 功能注释

纵坐标表示 KEGG 通路名称，横坐标表示对应的基因数目。

Fig. 3　KEGG assignment of new genes of *P. sutchi*

The y-axis represents the KEGG pathway, and the x-axis shows the number of new genes with them.

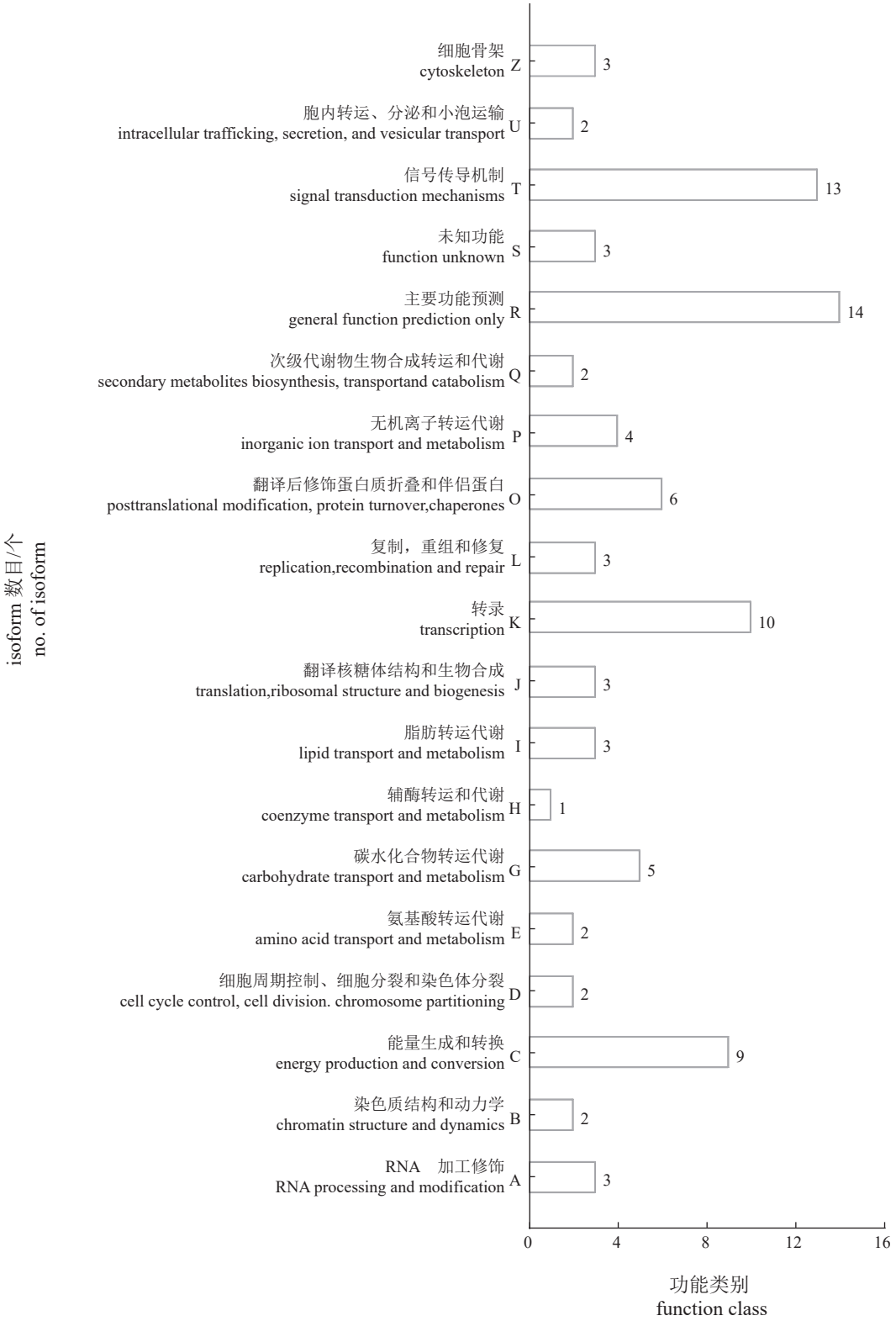
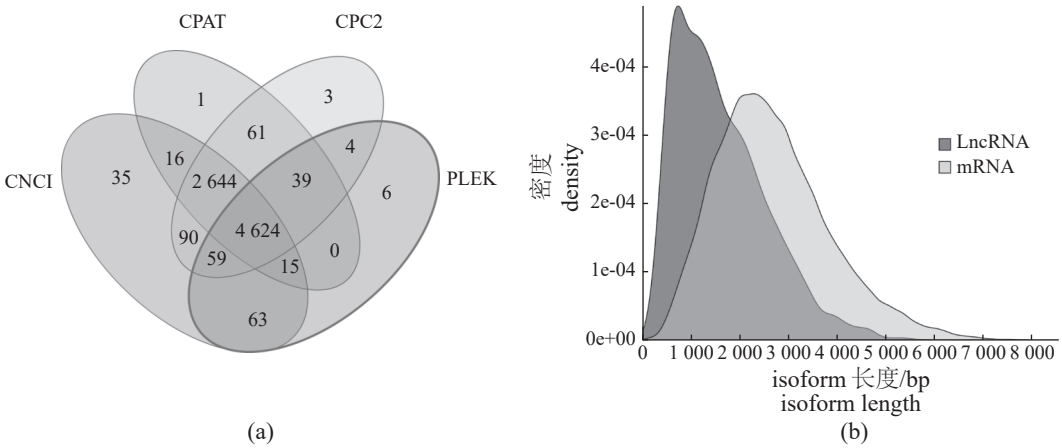


图 4 苏氏圆腹鲚新基因 KOG 富集结果

纵坐标表示 KOG 功能类别的名称，横坐标表示对应的基因数目 (柱子右侧数字)。

Fig. 4 KOG enrichment of new genes of *P. sutchi*

The y-axis represents the KOG categories, and the x-axis shows the number of new genes with them (the number on the right side of the column).



(a)　　(b)

图 5　苏氏圆腹鲀lncRNA 数目预测韦恩图 (a)，lncRNA 与 mRNA 长度密度分布 (b)

(b) 图中，横坐标表示 lncRNA 与 mRNA 的长度，纵坐标表示 lncRNA 与 mRNA 的密度。

Fig. 5　Venn diagram for predicting lncRNA number (a), and the length density distribution of lncRNA and mRNA (b) of *P. sutchi*

The x-axis represents the length of lncRNA and mRNA, and the y- axis represents the density of lncRNA and mRNA in Fig. b.

(643) 相关通路中；代谢中，主要富集在碳水化合物代谢 (584) 相关通路；有机系统中，主要富集在免疫系统 (2 135) 相关通路 (图 7)。

新 isoform 分析　　本研究中苏氏圆腹鲀全长转录组测序中鉴定到已知基因新 isoform 和新基因 isoform 共计 63 248 个，其中 50 311 个 isoform 经 TransDecoder 预测具有 ORF。用 Astalavista 软件对测序得到的 isoform 可变剪接事件进行分类和统计，共获得 79 922 个可变剪接事件，包括 5 937 (7.43%) 个外显子跳跃，4 992 (6.25%) 个可变受体，2 995 (3.75%) 个可变供体，11 738 (14.69%) 个内含子保留，371 (0.46%) 个外显子互斥和 53 889 (67.43%) 个其他类型。共鉴定到 18 个融合基因，其中 17 个为不同染色体的片段融合，1 个为相同染色体的片段融合。多聚腺苷酸化位点是一种常见的基因表达调控机制，本研究苏氏圆腹鲀全长转录组测序结果中分别有 2 915、1 288、493 和 209 个基因至少含有 2、3、4 和 5 个支持的 polyA 位点，共有 11 392 个 polyA 位点。

3　讨论

目前，转录组测序技术已广泛应用于生物、农业和医学等研究中，在功能基因的序列组成和转录表达模式研究中意义重大[15]。传统的转录组测序研究主要利用二代测序技术，在物种功能基因差异表达分析、新基因挖掘、寻找单核苷酸多态性及应用、基因功能注释等研究中发挥了重要作用[32-34]。但是基于二代测序技术的转录组测序长度 (100~150 bp) 远低于真核生物 RNA 的实际长度，序列组装拼接质量和基因注释效果不佳[35-37]。美国太平洋生物技术公司开发了第三代测序平台 (PacBio Sequel 测序仪)，具有测序通量高、周期短、单位数据成本低等特点，深受研究人员青睐。PacBio SMRT 测序技术不需要 PCR 扩增、组装，就可直接获得全长转录序列，是转录组从头测序的首选[38-39]。N50 是衡量转录组测序质量的标准之一，本研究对苏氏圆腹鲀全长转录组测序分析，共获得 58 868 917 条子序列，序列矫正过滤后共鉴定了 667 973 条含 polyA 的 FLNC，N50 为 2 359 bp，高于青鳉 (*Oryzias melastigma*，N50 为 2 162 bp)[34]、鲇 (*Pterygoplichthys anisitsi*，N50 为 1 571 bp)[40] 和花鲈 (*Lateolabrax maculatus*，N50 为 1 536 bp)[41] 等二代转录组组装效果，说明本研究全长转录组测序数据质量较好。

研究表明，汇总物种不同组织样本进行全长转录组测序可以获得该物种较为完整的转录本信息[38, 42-44]。本研究分别提取了性成熟雌、雄苏氏圆腹鲀的脑、鳃、心脏、肝脏、脾脏、头肾、胃、肠道、性腺和肌肉总 RNA，等质量混合后构建了 cDNA 文库并进行全长转录组测序，测序结果结合二代测序数据校正，获得了雌雄苏氏圆腹鲀各组织中较为完整的转录表达图谱。本研究鉴定到了 63 644 个已知新基因的新 isoform和 9 348 个新基因，并成功对其中 4 053 (43.36%) 个新基因完成了功能注释，对苏氏圆腹鲀参考基

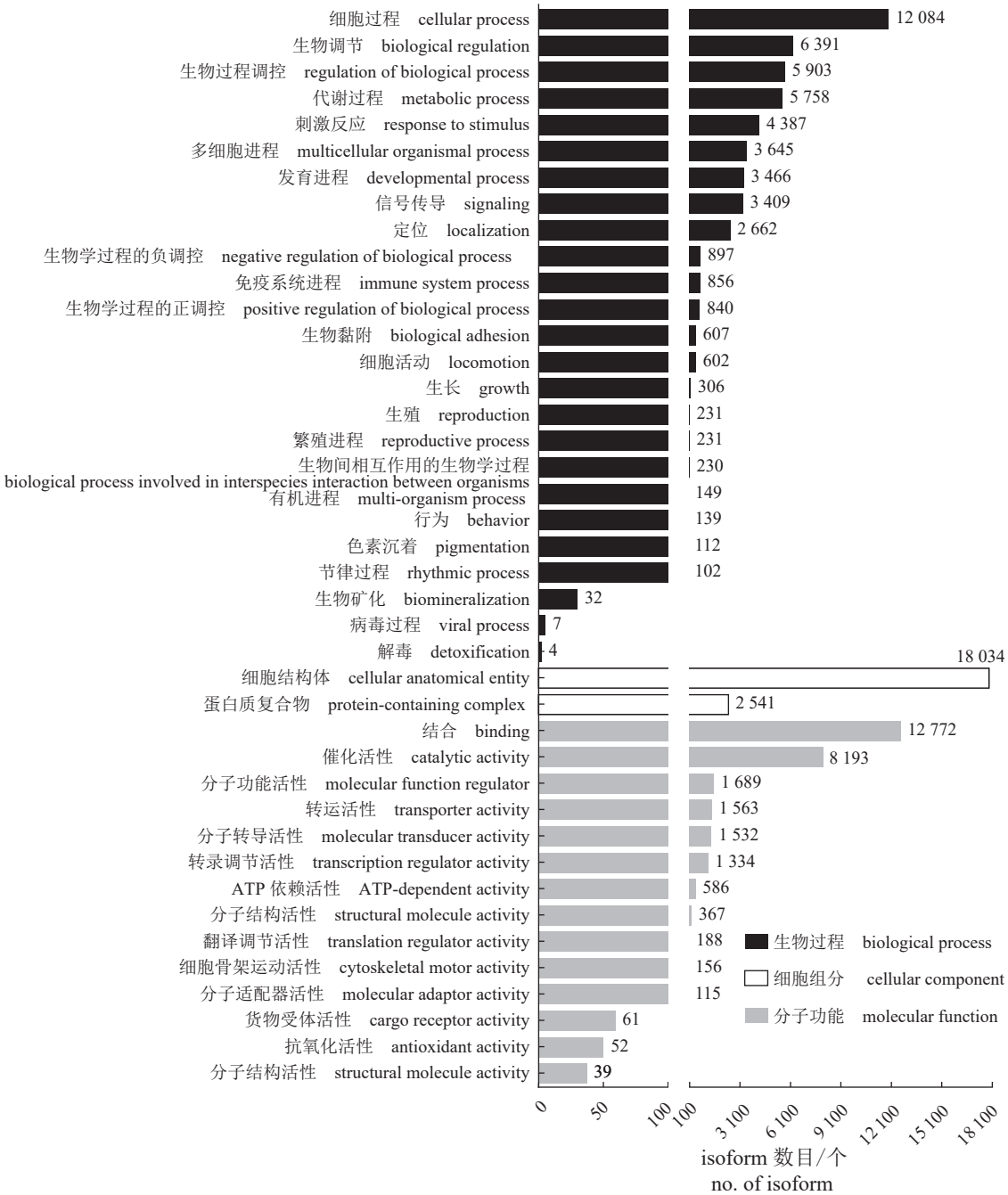


图 6　苏氏圆腹鲀LncRNA 调控的 mRNA GO 功能注释

纵坐标表示 GO 条目名称，横坐标表示条目对应的 mRNA 数目。

Fig. 6　Gene ontology (GO) assignment of lncRNA -regulated mRNA of *P. sutchi*

The y-axis represents the GO term, and the x-axis shows the number of mRNAs with them.

因组的组装注释起到了矫正和补充的作用，为苏氏圆腹鲀基因结构及功能的研究提供了基础。

生物体内不仅存在大量的编码蛋白的 mRNA，也含有大量的不编码蛋白的 lncRNA，lncRNA 在不同的组织和细胞中具有高度的特异性[45-46]。lncRNA 可以通过顺式或反式作用调控靶基因的转录，从而参与调控生物体的生长、发育和免疫平衡等作用[47-48]。lncRNAs 可以通过全长转录组测序进行鉴定[49]，本研究在性成熟苏氏圆腹鲀全长转录组测序结果中共鉴定出 4 624 个 lncRNA，调控 32 283 个靶 mRNA，其中 2 135 个富集在免疫相关通路中。从养殖情况来看，苏氏圆腹鲀的发病率普遍较低，可能说明该鱼有较强的免疫力。因此，在今后的工作中，可以将鉴定得到的 lncRNA

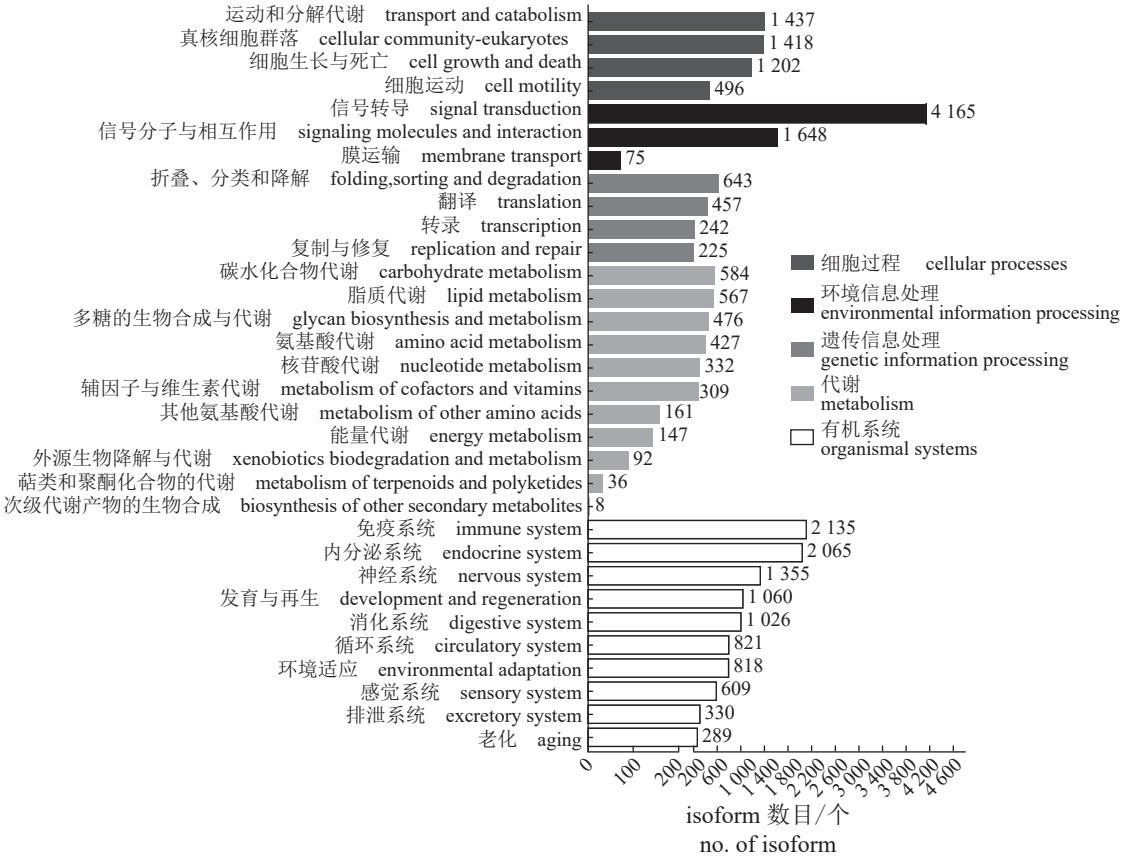


图 7　苏氏圆腹鲀LncRNA 调控的 mRNA KEGG 功能注释

纵坐标表示 KEGG 通路名称，横坐标表示对应的 mRNA 数目。

Fig. 7　KEGG assignment of lncRNA -regulated mRNA of *P. sutchi*

The y-axis represents the KEGG pathway, and the x-axis shows the number of mRNAs with them.

在免疫学方面做进一步的研究，深入探究 lncRNA 种类及丰度和鱼类免疫的关联。

4　结论

本研究使用 PacBio SMRT 测序技术对性成熟苏氏圆腹鲀各组织的全长转录组进行了测序分析。共鉴定了 667 973 条含 poiyA 的 FLNC，获得 63 644 个已知基因的新 isoform 和 9 348 个新基因，预测到 50 311 个 ORF、79 922 个可变剪切事件和 18 个融合基因；另外，还鉴定出 4 624 个 lncRNA。通过对苏氏圆腹鲀全长转录组的注释分析，初步阐明了苏氏圆腹鲀的基因参与的生物过程、所在的代谢途径或信号通路等，为深入研究苏氏圆腹鲀基因的功能奠定了基础。

（作者声明本文无实际或潜在的利益冲突）

参考文献 (References):

[1] Sriket P, La-Ongnual T. Quality changes and discoloration of basa (*Pangasius Bocourti*) fillet during frozen storage[J]. Journal of Chemistry, 2018, 2018: 5159080.

[2] Rahmah S, Nasrah U, Lim L S, *et al*. Aquaculture wastewater-raised azolla as partial alternative dietary protein for pangasius catfish[J]. Environmental Research, 2022, 208: 112718.

[3] 沈校, 龚明, 邹峥嵘, 等. 巴沙鱼皮胶原蛋白的提取、组成及变性温度研究 [J]. 中国海洋药物, 2017, 36(4): 41-46.
Shen X, Gong M, Zou Z R, *et al*. Study on extraction, composition and denaturation temperature of Collagen from *Pangasisus haniltoa* skin[J]. Chinese Journal of Marine Drugs, 2017, 36(4): 41-46 (in Chinese).

[4] Nguyen T T, Luong T H, Vo Van T, *et al*. Investigation of extraction and characterization of collagen from the skin of striped catfish (*Pangasianodon hypophthalmus*) [C]//Biomedical Engineering Society in Vietnam. Proceedings of the 6th International Conference on the Development of Biomedical Engineering in Vietnam. Ho Chi Minh: Springer, 2017: 789-792.

[5] 魏于生. 巴沙鱼人工繁殖获得成功 [J]. 淡水渔业,

1998, 28(2): 36.
Wei Y S. A success in the artificial propagation of *Pangasius hanigtoa*[J]. Freshwater Fisheries, 1998, 28(2): 36 (in Chinese).

[6] 申玉春, 黄木珍, 苏秋生. 苏氏圆腹鲃胚胎发育的初步观察 [J]. 广东海洋大学学报, 2008, 28(4): 41-44.
Shen Y C, Huang M Z, Su Q S. A preliminary observation on the embryonic development of *Pangasius sutchi*[J]. Journal of Guangdong Ocean University, 2008, 28(4): 41-44 (in Chinese).

[7] 刘修英, 王岩, 王建华. 利用豆粕、菜粕和棉粕替代饲料中鱼粉对苏氏圆腹鲃摄食、生长和饲料利用的影响 [J]. 水产学报, 2009, 33(3): 479-487.
Liu X Y, Wang Y, Wang J H. Effects of replacing fish meal with soybean meal, rapeseed meal or cottonseed meal on feeding, growth and feed utilization of sutchi catfish (*Pangasius sutchi*)[J]. Journal of Fisheries of China, 2009, 33(3): 479-487 (in Chinese).

[8] 刘修英, 刘叶子, 朱崇梅, 等. 苏氏圆腹肌肉化学成分分析与评价 [J]. 食品工业科技, 2017, 38(2): 349-353.
Liu X Y, Liu Y Z, Zhu C M, *et al*. Analysis and evaluation of chemical components of muscles of *Pangasius hypophthalmus*[J]. Science and Technology of Food Industry, 2017, 38(2): 349-353 (in Chinese).

[9] Ali Z, Hossain A, Mazid A. Effect of mixed feeding schedules with varying dietary protein levels on the growth of sutchi catfish, *Pangasius hypophthalmus* (Sauvage) with silver carp, *Hypophthalmichthys molitrix* (Valenciennes) in ponds[J]. Aquaculture Research, 2005, 36(7): 627-634.

[10] Van Doan H, Doolgindachbaporn S, Suksri A. Effects of low molecular weight agar and *Lactobacillus plantarum* on growth performance, immunity, and disease resistance of basa fish (*Pangasius bocourti*, Sauvage 1880)[J]. Fish & Shellfish Immunology, 2014, 41(2): 340-345.

[11] Meidong R, Khotchanalekha K, Doolgindachbaporn S, *et al*. Evaluation of probiotic *Bacillus aerius* B81e isolated from healthy hybrid catfish on growth, disease resistance and innate immunity of Pla-mong *Pangasius bocourti*[J]. Fish & Shellfish Immunology, 2018, 73: 1-10.

[12] Guimarães C F M, Márcico E T, Monteiro M L G, *et al*. The chemical quality of frozen Vietnamese *Pangasius hypophthalmus* fillets[J]. Food Science & Nutrition, 2016, 4(3): 398-408.

[13] Thanh N M, Jung H, Lyons R E, *et al*. A transcriptomic analysis of striped catfish (*Pangasianodon hypophthalmus*) in response to salinity adaptation: De novo assembly, gene annotation and marker discovery[J]. Comparative Biochemistry and Physiology-Part D: Genomics and Proteomics, 2014, 10: 52-63.

[14] Gao Z J, You X X, Zhang X H, *et al*. A chromosome-level genome assembly of the striped catfish (*Pangasianodon hypophthalmus*)[J]. Genomics, 2021, 113(5): 3349-3356.

[15] 崔凯, 吴伟伟, 刁其玉. 转录组测序技术的研究和应用进展 [J]. 生物技术通报, 2019, 35(7): 1-9.
Cui K, Wu W W, Diao Q Y. Application and research progress on transcriptomics[J]. Biotechnology Bulletin, 2019, 35(7): 1-9 (in Chinese).

[16] Liao Y C, Lin S H, Lin H H. Completing bacterial genome assemblies: strategy and performance comparisons[J]. Scientific Reports, 2015, 5: 8747.

[17] Deng A P, Li J P, Yao Z B, *et al*. SMRT sequencing of the full-length transcriptome of the *Coelomactra antiquata*[J]. Frontiers in Genetics, 2021, 12: 741243.

[18] Huang Y Y, Zhang L L, Huang S Y, *et al*. Full-length transcriptome sequencing of *Heliocidaris crassispina* using PacBio single-molecule real-time sequencing[J]. Fish & Shellfish Immunology, 2022, 120: 507-514.

[19] Yang Y J, Chen M Y, Wang Y X, *et al*. A novel full-length transcriptome resource for sea cucumber *Apostichopus japonicus* using Pacbio SMRT sequencing[J]. Frontiers in Marine Science, 2022, 9: 834255.

[20] Salmela L, Rivals E. LoRDEC: accurate and efficient long read error correction[J]. Bioinformatics, 2014, 30(24): 3506-3514.

[21] Wu T D, Watanabe C K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences[J]. Bioinformatics, 2005, 21(9): 1859-1875.

[22] Buchfink B, Xie C, Huson D H. Fast and sensitive protein alignment using DIAMOND[J]. Nature Methods, 2015, 12(1): 59-60.

[23] Ashburner M, Ball C A, Blake J A, *et al*. Gene ontology: tool for the unification of biology[J]. Nature Genetics, 2000, 25(1): 25-29.

[24] Tatusov R L, Fedorova N D, Jackson J D, *et al*. The COG database: an updated version includes eukaryotes[J]. BMC Bioinformatics, 2003, 4: 41.

[25] Kanehisa M. The KEGG database [M]. Chichester: Novartis foundation symposium, 2002, 247: 91-103.

[26] Kanehisa M, Goto S, Kawashima S, *et al*. The KEGG resource for deciphering the genome[J]. Nucleic Acids Research, 2004, 32(S1): D277-D280.

[27] Xie C, Mao X Z, Huang J J, *et al*. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases[J]. Nucleic Acids Research, 2011, 39(S2): W316-W322.

[28] Gasteiger E, Jung E, Bairoch A. SWISS-PROT: connecting biomolecular knowledge via a protein database[J]. Current Issues in Molecular Biology, 2001, 3(3): 47-55.

[29] Khor Y P, Wan S Y, Tan C P, *et al*. Potential of using basa catfish oil as a promising alternative deep-frying medium: a thermo-oxidative stability study[J]. Food Research International, 2021, 141: 109897.

[30] Florea L, Song L, Salzberg S L. Thousands of exon skipping events differentiate among splicing patterns in sixteen human tissues[J]. F1000Research, 2013, 2: 188.

[31] Abdel-Ghany S E, Hamilton M, Jacobi J L, *et al*. A survey of the sorghum transcriptome using single-molecule long reads[J]. Nature Communications, 2016, 7: 11706.

[32] 林明德, 陈刚, 马骞, 等. 杂交石斑鱼和母本褐点石斑鱼转录组测序及差异表达基因分析 [J]. 广东海洋大学学报, 2019, 39(3): 15-23.
Lin M D, Chen G, Ma Q, *et al*. Differentially expressed genes analysis of hybrid grouper and female tiger grouper based on transcriptome sequencing[J]. Journal of Guangdong Ocean University, 2019, 39(3): 15-23 (in Chinese).

[33] Shen M H, Jiang Z Y, Zhang K, *et al*. Transcriptome analysis of grass carp (*Ctenopharyngodon idella*) and Holland's spinibarbel (*Spinibarbus hollandi*) infected with *Ichthyophthirius multifiliis*[J]. Fish & Shellfish Immunology, 2022, 121: 305-315.

[34] Lai K P, Li J W, Wang S Y, *et al*. Tissue-specific transcriptome assemblies of the marine medaka *Oryzias melastigma* and comparative analysis with the freshwater medaka *Oryzias latipes*[J]. BMC Genomics, 2015, 16(1): 135.

[35] Li X D, Wu J M, Xiao X P, *et al*. Characterization and complexity of transcriptome in *Gymnocypris przewalskii* using single-molecule long-read sequencing and RNA-seq[J]. DNA Research, 2021, 28(3): dsab005.

[36] Yi J K, Yang S, Wang S, *et al*. Identification of candidate chemosensory receptors in the antennal transcriptome of the large black chafer *Holotrichia parallela* Motschulsky (Coleoptera: *Scarabaeidae*)[J]. Comparative Biochemistry and Physiology-Part D: Genomics and Proteomics, 2018, 28: 63-71.

[37] 高胜寒, 禹海英, 吴双阳, 等. 复杂基因组测序技术研究进展 [J]. 遗传, 2018, 40(11): 944-963.
Gao S H, Yu H Y, Wu S Y, *et al*. Advances of sequencing and assembling technologies for complex genomes[J]. Hereditas, 2018, 40(11): 944-963 (in Chinese).

[38] Luo W, Wu Q, Wang T Z, *et al*. Full-length transcriptome analysis of *Misgurnus anguillicaudatus*[J]. Marine Genomics, 2020, 54: 100785.

[39] Han X, He B, Xin Y, *et al*. Full-length sequencing of *Ginkgo biloba* L. reveals the synthesis of terpenoids during seed development[J]. Industrial Crops and Products, 2021, 170: 113714.

[40] Parente T E, Moreira D A, Magalhães M G P, *et al*. The liver transcriptome of suckermouth armoured catfish (*Pterygoplichthys anisitsi*, Loricariidae): identification of expansions in defensome gene families[J]. Marine Pollution Bulletin, 2017, 115(1-2): 352-361.

[41] Zhang X Y, Wen H S, Wang H L, *et al*. RNA-Seq analysis of salinity stress-responsive transcriptome in the liver of spotted sea bass (*Lateolabrax maculatus*)[J]. PLoS One, 2017, 12(3): e0173238.

[42] Amparyup P, Charoensapsri W, Samaluka N, *et al*. Transcriptome analysis identifies immune-related genes and antimicrobial peptides in Siamese fighting fish (*Betta splendens*)[J]. Fish & Shellfish Immunology, 2020, 99: 403-413.

[43] Li C, Zhong S H, Jin J J, *et al*. Full-length transcriptome data for the white cloud mountain minnow (*Tanichthys albonubes*) from a wild population based on isoform sequencing[J]. Frontiers in Marine Science, 2022, 9: 831148.

[44] Ge H L, Zhang H Y, Yang L J, *et al*. Full-length transcriptome sequencing from the longest-lived freshwater bony fish of the world: bigmouth buffalo (*Ictiobus cyprinellus*)[J]. Frontiers in Marine Science, 2021, 8: 736188.

[45] Kaushik K, Leonard V E, Kv S, *et al*. Dynamic expression of long non-coding RNAs (lncRNAs) in adult zebrafish[J]. PLoS One, 2013, 8(12): e83616.

[46] Jiang L H, Liu W, Zhu A Y, *et al*. Transcriptome analysis demonstrate widespread differential expression of long noncoding RNAs involve in *Larimichthys crocea* immune response[J]. Fish & Shellfish Immunology, 2016, 51: 1-8.

[47] Liu N K, Xu Y Z, Li Q, *et al*. A lncRNA fine-tunes salicylic acid biosynthesis to balance plant immunity and growth[J]. Cell Host & Microbe, 2022, 30(8): 1124-1138. e8.

[48] Yang N, Wang B B, Yu Z X, *et al*. Characterization of a novel lncRNA (*SETD3-OT*) in turbot (*Scophthalmus maximus* L.)[J]. Fish & Shellfish Immunology, 2020, 102: 145-151.

[49] Kapranov P, Cheng J, Dike S, *et al*. RNA maps reveal new RNA classes and a possible function for pervasive transcription[J]. Science, 2007, 316(5830): 1484-1488.

Full-length transcriptome analysis of *Pangasius sutchi*

HUAN Zhang [1], LI Dongyu [2], LI Weihao [1], GAO Jin [3],
WANG Zhongduo [1], PAN Zhi [2], DONG Zhongdian [1,4*]

(1. *Key Laboratory of Aquaculture in South China Sea for Aquatic Economic Animal of Guangdong Higher Education Institutes*, *Fisheries College*, *Guangdong Ocean University*, *Zhanjiang* 524088, *China*;
2. *Aquatic Technology Promotion Station of Beijing*, *Beijing* 100176, *China*;
3. *Key Laboratory of Utilization and Conservation for Tropical Marine Bioresources*, *Ministry of Education*, *Hainan Tropical Ocean University*, *Sanya* 572022;
4. *Guangdong Provincial Key Laboratory of Aquatic Animal Disease Control and Healthy Culture*, *Guangdong Ocean University*, *Zhanjiang* 524088, *China*)

Abstract: *Pangasius sutchi*, a significant freshwater economic fish in Southeast Asia, is characterized by rapid growth, ease of cultivation, rich nutritional content, and the absence of small intermuscular bones. First introduced to China from Thailand in 1978, *P. sutchi* achieved a breakthrough in artificial breeding in 1997 and has since been extensively promoted in Guangdong, Guangxi, and Hainan provinces. Current research on *P. sutchi* primarily focused on breeding models, nutritional feed development, disease control, and fish product processing technology, with less emphasis on basic biology, particularly molecular biology. This study sequenced the full-length transcriptome from brain, gills, heart, liver, spleen, head kidney, stomach, intestines, gonads, and muscles of sexually matured *P. sutchi* using Single Molecule Real-Time (SMRT) sequencing on the PacBio Sequel platform to elucidate the genetic basis and support molecular biology research. A total of 1 487 336 high-quality reads were obtained, averaging 83 592 bp in length with an N50 of 162 901 bp. After self-correction, 1 005 955 CCS (Circular Consensus Sequence) were derived, and following filtration, 667 973 polyA-containing FLNC (full-length non-concatenated) were identified, averaging 2 057 bp in length with an N50 of 2 359 bp. For gene and transcript annotation, 614 078 (91.93%) FLNC were used, identifying 19 835 known genes and 9 348 novel genes. In addition, 50 311 ORF (open reading frame), 79 922 alternative splicing, 18 fusion genes, and 20 215 alternative polyadenylation sites were predicted. Of the 9 348 novel genes, 3 912, 2 385, 2 167, 81 and 1 520 were annotated in NR (non-redundant protein sequences), GO (gene ontology), KEGG (Kyoto encyclopedia of genes and genomes), KOG (eukaryotic orthologous groups) and SwissProt databases, respectively. GO enrichment analysis revealed that 1 309, 1 351, and 1 524 new genes were enriched in cellular process, cellular anatomical entities, and binding terms, respectively. KEGG enrichment indicated that the new genes were primarily enriched in cellular processes such as eukaryotes (106), signal transduction (276), folding, sorting and degradation (79), amino acid metabolism (63), and endocrine system (197). 4 624 lncRNA were obtained in *P. sutchi*, regulating 32 283 target mRNA. GO enrichment results showed that target mRNA were mainly enriched in cellular processes (12 084), cellular anatomical entity (18 034) and binding (12 772). KEGG analysis indicated that the target mRNA were predominantly enriched in the transport and catabolism pathway (1 437), signal transduction (4 165) pathway, folding, sorting and degradation (643) pathway, carbohydrate metabolism (584) pathway and immune system (2 135). In this study, the full-length transcriptome sequencing data analysis and functional annotation have enriched the genetic resources of *P. sutchi* and provided a basis for further research on the biological characteristics and gene function of *P. sutchi*.

Key words: *Pangasius sutchi*; full-length transcriptome; gene function; alternative splicing; long non-coding RNA

Corresponding author: DONG Zhongdian. E-mail: zddong@gdou.edu.cn

Funding projects: National Natural Science Foundation of China (31201996); Guangdong Ocean University Nanhai Scholars Program Young Talent Project (QNXZ201903, 201807); Start-up Project from GDOU PhD