



利用 RNA-Seq 技术分析草鱼生长性状 相关基因和 SNP 标记

孙 雪^{1,2}, 李胜杰^{1*}, 姜 鹏¹, 杜金星¹, 周家辉^{1,2}, 白俊杰¹

(1. 中国水产科学研究院珠江水产研究所, 农业农村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, 水产科学国家级实验教学示范中心, 上海 201306)

摘要: 为开发与草鱼生长性状相关基因及单核苷酸多态性 (single nucleotide polymorphism, SNP) 标记, 实验采用转录组测序技术 (RNA sequencing, RNA-Seq) 对草鱼快长组和慢长组的肝脏、肌肉和脑组织分别进行分析, 共获得 31 465 万条高质量短读序 (clean reads), 经组装得到 34 147 条拼接基因 (unigene), 平均长度为 1 060 bp, 其中有 30 751 条 unigene 获得注释。在肝脏、肌肉和脑组织中分别筛选到 1 013、552 和 372 个差异表达基因, 并从中检测到 4 580 个 SNP 标记。采用 SNaPshot 技术对其中 34 个 SNP 标记在 300 尾草鱼生长性状极端群体中进行多态性检测和验证, 30 个 SNP 标记分型成功, 准确率为 88.24%。进一步采用一般线性模型分析 30 个 SNP 标记与生长性状的相关性, 结果显示, unigene00810126-8014 标记 CC 基因型的体质量、体长、体高、头长和尾柄长性状均值显著高于 TT 基因型。unigene00810126-2903 标记 AA 基因型的体质量显著高于 GG 基因型。unigene00870394-525 标记 AA 基因型的体质量和体长性状均值显著高于 GG 基因型。unigene02938762-011628 标记的 TT 基因型与 CC 基因型在体质量、体长和头长性状上的差异均达到显著水平。这 4 个标记位于生长催乳素 α 基因 (*sla*)、早期生长反应蛋白-1 基因 (*egr-1*) 和肌球蛋白重链基因 (*myh*) 上。其他标记不同基因型个体间的生长性状均不存在显著差异。其中 8 个 SNP 标记在草鱼群体中的平均多态信息含量 (PIC)、平均期望杂合度 (H_e) 和平均观测杂合度 (H_o) 分别为 0.300、0.377 和 0.363, 表明草鱼选育群体遗传多样性较高。本研究获得与生长性状相关的 1 937 个差异表达基因和 4 个 SNP 标记, 为草鱼分子辅助育种研究提供基础资料。

关键词: 草鱼; 转录组测序; 差异表达基因; 单核苷酸多态性; 生长性状

中图分类号: Q 785; S 917.4

文献标志码: A

草鱼 (*Ctenopharyngodon idella*), 隶属于鲤形目 (Cypriniformes) 鲤科 (Cyprinidae) 草鱼属 (*Cte-*

nopharyngodon), 是我国淡水养殖鱼类中养殖年产量最高的养殖品种。草鱼繁殖亲本主要来源

收稿日期: 2019-09-25 修回日期: 2020-05-16

资助项目: 现代农业产业技术体系专项 (CARS-45-04); 中国水产科学研究院中央级公益性科研院所基本科研业务费专项 (2020TD34, 2019ZX-001); 广州市科技计划项目 (201906010004)

第一作者: 孙雪, 从事生物技术及动物遗传育种研究, E-mail: 1457753528@qq.com

通信作者: 李胜杰 (照片), E-mail: ssjli@163.com



于江河中直接捕捞的野生鱼, 未经过驯化选育。此外, 多数繁殖场不遵守科学的苗种生产规范, 选择个体小、性成熟早、亲缘关系近的草鱼作为亲本, 导致养殖草鱼出现一定程度的种质退化现象, 主要表现为生长速率减慢、抗逆性能下降等, 制约草鱼养殖产业的健康发展^[1]。因此, 亟需加强草鱼种质资源的挖掘利用, 尽快选育出生产性状优良的草鱼品种。

草鱼的性成熟年龄为 4~5 a, 育种周期较长。随着分子生物技术的快速发展, 分子标记辅助育种技术有望缩短草鱼育种周期和快速培育出草鱼优良品种的一个重要途径。分子标记辅助育种技术是通过与目的基因紧密连锁或共分离的分子标记, 对目的基因进行筛选, 从而不受环境条件的影响, 提高了选择的可靠性和育种效率, 在动植物育种中已有广泛应用^[2-3]。规模化开发与草鱼生长相关的功能基因以及挖掘与生长性状连锁的分子标记是开展草鱼分子标记辅助育种研究的基础性工作。单核苷酸多态性标记 (single nucleotide polymorphism, SNP) 是指单个核苷酸发生改变而引起的 DNA 序列多态性, 具有数量多、密度大、遗传稳定性高等优点而被广泛应用^[4]。由于全基因组测序费用高, 因此越来越多的研究者运用高通量低成本的转录组测序技术 (RNA-Seq) 筛选与生长性状相关的功能基因和 SNP 标记^[5-7]。Salem 等^[8]对虹鳟 (*Oncorhynchus mykiss*) 快长组和慢长组群体分别进行 RNA-Seq 分析, 鉴定出 22 个与生长性状相关的 SNP 标记。Ulloa 等^[9]采用 RNA-Seq 技术分析了斑马鱼 (*Danio rerio*) 快长组和慢长组群体的转录组, 获得了与生长性状相关的 124 个差异表达基因, 进一步对这些基因上的 164 个 SNP 标记与生长性状进行关联分析, 鉴定出 5 个与生长性状关联的 SNP 标记。Lv 等^[10]对三疣梭子蟹 (*Portunus trituberculatus*) 的大小蟹群体进行 RNA-Seq 分析, 筛选得到 117 个差异表达基因及 1 个与生长性状相关联的 SNP 标记。Li 等^[11]对大口黑鲈 (*Micropterus salmoides*) 快长组和慢长组群体的转录组进行高通量测序, 筛选到 4 330 个与生长性状相关的差异表达基因, 并对其中部分基因上的 SNP 标记与生长性状进行关联分析, 鉴定出 3 个与生长性状显著相关的 SNP 标记。本实验采用 RNA-Seq 技术分析快长组与慢长组草鱼群体中肌肉、肝脏和脑组织的基因转录表达谱, 筛选出与

生长性状相关的差异表达基因, 进一步对差异表达基因中的 SNP 标记与生长性状进行关联分析, 以期草鱼分子辅助育种研究提供基础资料, 加快草鱼优良品种的选育进程。

1 材料与方法

1.1 实验鱼及基因组 DNA 提取

用于 RNA-Seq 分析的实验鱼来自佛山市三水白金水产种苗有限公司青岐基地, 从同批次繁殖的草鱼中选取 16 月龄的极大个体 10 尾 [快长组, 平均体质量为 (2.46 ± 0.14) kg] 和极小个体 10 尾 [慢长组, 平均体质量为 (1.48 ± 0.10) kg], 从每尾鱼中剪取肌肉、肝脏和脑组织分别冻存于液氮中, 用于总 RNA 提取。

用于关联分析的草鱼群体来自于佛山市南海百容水产良种有限公司清远基地养殖的草鱼选育品系, 随机选取同批次繁殖、且养殖在同一个池塘中的 16 月龄草鱼 300 尾, 分别测量体质量、体长、体高、头长、尾柄长和尾柄高, 同时剪取腹鳍样本置于无水乙醇中保存备用。按照海洋动物组织基因组 DNA 提取试剂盒 [天根生化科技 (北京) 有限公司] 说明书提取鳍条样品基因组 DNA, 用琼脂糖 (Sigma 公司, 美国) 凝胶电泳和多功能酶标仪 (BioTek 公司, 美国) 测定 DNA 质量和浓度, 在 -20°C 条件下保存备用。

1.2 RNA 提取、文库构建与测序

肌肉、肝脏和脑组织总 RNA 的提取按照 TRIzol 试剂盒 (TaKaRa 公司, 北京) 说明书进行, 并用 2% 琼脂糖凝胶电泳和多功能酶标仪分别检测其质量和浓度。构建 RNA 文库之前, 通过 DNase I 去除基因组 DNA 污染。将每组个体各个组织的等量 RNA 混合一起, 共组成 6 个测序样品。用带有 Oligo(dT) 的磁珠富集 mRNA, 然后加入 fragmentation buffer 使其片断化为短片段, 再以短片段 mRNA 为模板, 用六碱基随机引物合成 cDNA 第一链, 并加入缓冲溶液、dNTPs、RNase H 和 DNA polymerase I 合成 cDNA 第二链, 经 QiaQuick PCR 试剂盒 (Qiagen 公司, 美国) 纯化、加测序接头、琼脂糖凝胶电泳回收目的片段并进行 PCR 扩增, 从而完成整个文库的制备工作。构建好的文库用 Illumina HiSeq™ 2000 进行测序。

1.3 差异表达基因筛选

将 RNA-Seq 得到的原始序列 (raw reads) 用 Cufflinks(2.2.1) 软件^[12] 组装得到拼接基因 (unigene), 通过 BLASTX 方法将 unigene 分别比对到 Nr、KEGG 和 Swiss-Prot 数据库中, 使用 Blast2GO 软件^[13] 得到 unigene 的 GO 注释信息, 之后用 WEGO 软件^[14] 执行 GO 功能分类统计。采用 edgeR 软件^[15] 对样本间基因表达量进行差异分析, 用 FPKM (fragments per kilobase of exon per million fragments mapped) 法^[16] 对基因表达量进行计算, 通过 FDR (false discovery rate) 控制 P -value 的阈值, 将 $FDR < 0.05$ 且 $|\log_2(\text{Fold Change})| > 1$ 作为筛选条件, 筛选不同组间差异表达基因。

1.4 SNP 筛选与验证

用 GATK 软件^[17] 在差异表达基因中进行 SNP 标记的筛选, 条件限于突变位点至少在一个样本中、reads 丰度不小于 3、QD 值=1、FS(Fisher-Strand)>30。将关联分析群体中 30 尾极大个体和 30 尾极小个体作为 SNP 标记验证群体, 从差异表达基因中挑选 34 个潜在的 SNP 标记进行 SNaPshot 分型验证。进一步分析极端群体中与生长性状关联的位点在其余个体中的基因型, 最后在群体中进行标记与生长性状之间的关联分析。

SNaPshot 分型检测委托上海捷瑞生物工程有限公司完成。首先是根据 SNP 标记上下游的序列设计引物, 长度在 200~500 bp, 用于扩增含有 SNP 标记的目的片段, 通过多重 PCR 对模板进行扩增, 反应程序采用 Touch-down 方法: 95 °C 预变性 3 min; 94 °C 变性 15 s, 60 °C 退火 15 s, 72 °C 延伸 30 s, 11 个循环 (每个循环的退火温度降 0.5 °C); 随后进入下一个循环程序: 94 °C 变性 15 s, 54 °C 退火 15 s, 72 °C 延伸 30 s, 24 个循环; 72 °C 终延伸 3 min。将扩增到的目的片段

用 *Exo I* 和 FastAP 酶进行纯化, 去除反应产物中剩余的引物和 dNTP。采用 SNaPshot 试剂盒 (ABI 公司, 美国) 对纯化后的产物进行延伸反应, 然后在 ABI3730 全自动测序仪上进行测序。在 SNP 位点的上游或下游位置进行延伸引物的设计。

1.5 草鱼群体的遗传多样性分析

根据 SNP 标记分型结果统计各基因型频率, 以 PopGene 32 (Version 1.31)^[18] 和 PIC-Calculator 0.6 软件^[19] 计算 Hardy-Weinberg 平衡 (HWE)、多态信息含量 (PIC) 及有效等位基因数 (N_e) 等遗传多样性指标。

1.6 SNP 标记与生长性状的关联分析

采用 SPSS 19.0 软件 (一般线性模型中的多元方差分析) 进行基因型与生长性状的相关性分析。以 SNP 标记的不同基因型为固定因子, 体质量、体长、体高、头长、尾柄长和尾柄高性状值为因变量, 对不同标记基因型之间的数量性状值进行差异显著性检验^[20]。

2 结果

2.1 转录组测序数据、基因注释及差异表达基因筛选

采用 RNA-Seq 技术对草鱼快长组和慢长组群体的肝脏、肌肉和脑组织的转录组进行分析, 总共获得 31 953 万条 raw reads, 过滤后得到 31 465 万条高质量短读序 (clean reads)。各组织的 Q20 平均值为 97.9 %, 说明测序的准确度较高。组装后共获得了 34 147 条 unigene, 平均长度为 1 060 bp, 其中有 30 751 个基因获得注释, 占比 90.05% (表 1)。对草鱼快长组和慢长组群体的肝脏、肌肉和脑组织的转录组中的差异表达基因进行比较分析, 分别检测到 1 013、552 和 372 个差异表达基因。进一步对筛选到的差异表达基因进行

表 1 草鱼肝脏、肌肉、脑组织转录组文库测序数据

Tab. 1 Transcriptome library sequencing data of liver, muscle and brain tissues in *C. idella*

测序产量 sequencing production			组装 assemblys			注释 annotation				
原始短读序/bp raw reads	高质量短读序/bp clean reads	Q20/ %	拼接基因数量/个 all unigene no.	长度/bp length	平均长度/bp average length	GO	KEGG	Nr	Swiss-Prot	注释基因 数量/条 total annotated gene
319 533 882	314 657 346	97.9%	34 147	36 218 500	1 060	16 616 (48.66%)	8 075 (23.65%)	30 715 (89.95%)	24 805 (72.64%)	30 751 (90.05%)

注: Q20 为转录组测序质量控制值
Notes: Q20 is the quality control value of transcriptome sequencing

GO 功能性显著富集分析, 参与生物过程、分子功能和细胞组分相关联的基因个数所占的比率分别为 48.26%、20.51% 和 31.23%(图 1)。

2.2 SNP 标记筛选、有效性检验及其遗传多样性分析

在草鱼快长组和慢长组群体的肝脏、肌肉和脑组织的转录组差异表达基因中分别筛选到 2 540、1 619 和 421 个 SNP 标记。为进一步验证这些 SNP 标记的有效性, 在差异表达基因中挑选出 34 个 SNP 标记在草鱼生长性状极端群体中进行分型检测。结果显示, 30 个 SNP 标记可成

功分型, 准确率为 88.24%(表 2), 并进行基因型检测(部分结果展示)(图 2, 图 3)。

对其中的 8 个 SNP 标记的遗传参数进行统计分析, 各标记的遗传多态信息含量 (PI) 介于 0.185~0.374, 结果显示, 除 unigene02722463-1898、unigene04620686-435 为低度多态位点 ($PI < 0.25$) 以外, 其他均为中度多态位点 ($0.25 < PI < 0.5$)。8 个标记的平均有效等位基因数 (N_e)、平均期望杂合度 (H_e) 和平均观测杂合度 (H_o) 分别为 1.634、0.377 和 0.363, 表明草鱼人工选育群体具有较丰富的遗传多样性。卡方分析显示, 8 个 SNPs 标

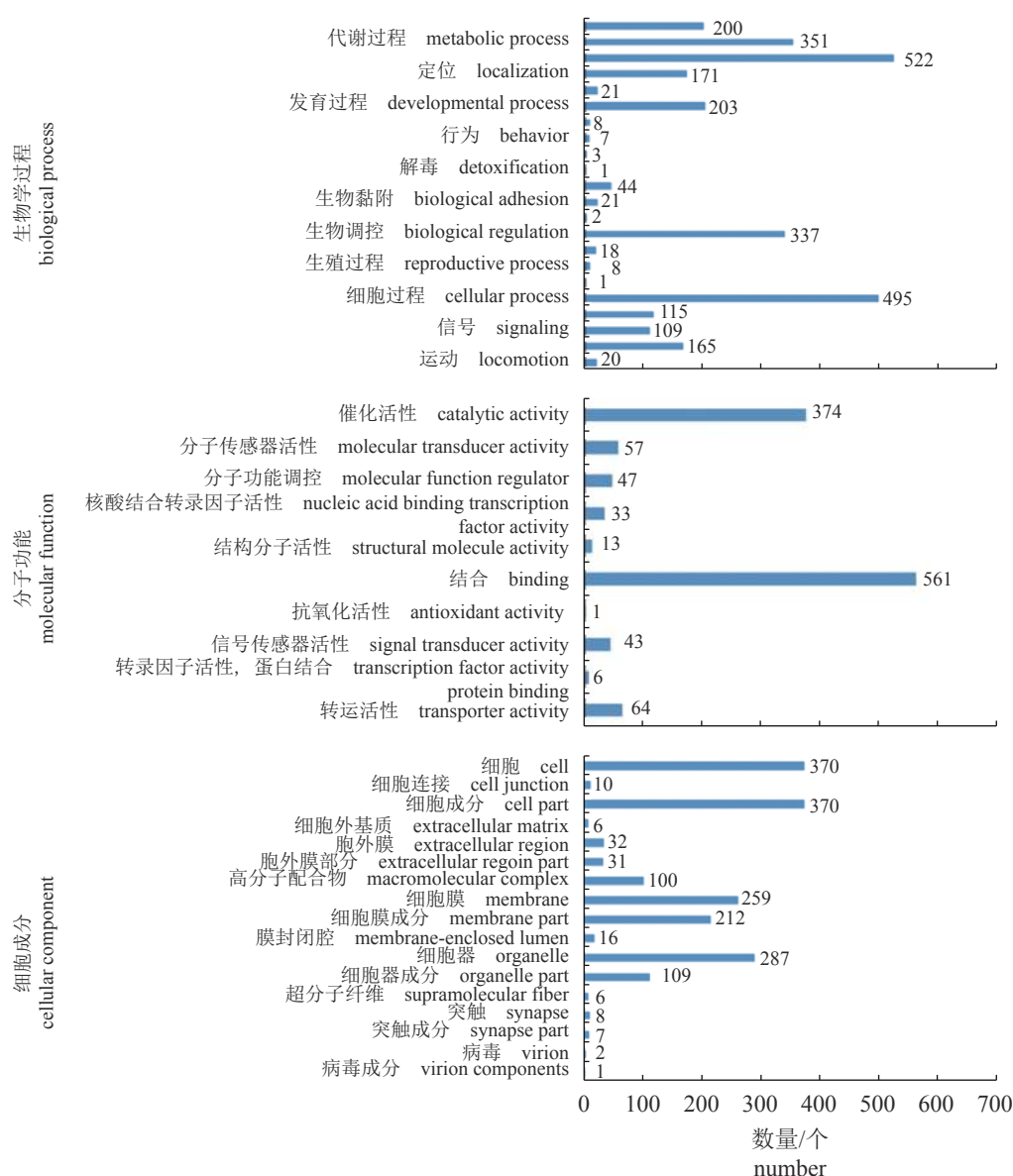


图 1 差异表达基因 GO 功能分类

Fig. 1 GO functional classification of differentially expressed genes

表 2 差异表达基因中 SNP 位置及其基因注释

Tab. 2 SNP positions and their gene annotations in differentially expressed genes

编号 ID number	SNP位置 SNP position	基因名称 gene name
unigene00810126	8 014、2 903、568	生长催乳素 α 基因 somatolactin alpha, sla
unigene01116379	1 526	早期生长反应蛋白2b基因 early growth response protein 2b
unigene00870394	79、525、602	早期生长反应蛋白1b基因 early growth response protein 1b, egr-1b
unigene00340124	1 068、818、657	生长催乳素 β 基因 somatolactin beta
unigene02938762	18 537、16 046、15 610、11 751、11 628	肌球蛋白重链X3基因 myosin heavy chain-like isoform X, myh X3
unigene02722463	423、1 898	促甲状腺激素释放激素受体基因 thyrotropin-releasing hormone receptor b
unigene00105489	775、13	CREB结合蛋白X3亚型基因 CREB-binding protein isoform X3
unigene07584666	3 997、3 909	转录因子基因Adf-1 transcription factor Adf-1
unigene04468671	2 533、2 692	成纤维细胞生长因子19基因 fibroblast growth factor 19
unigene01787612	1 723、285	类胰岛素生长因子结合蛋白1基因 insulin-like growth factors binding protein 1
unigene03521602	3 168、2 790、2 358、1 350	胰岛素受体底物2基因 insulin receptor substrate 2
unigene01028640	242、795	类胰岛素生长因子结合蛋白复合物酸不稳定亚单位基因 insulin-like growth factor-binding protein complex acid labile subunit
unigene04620686	435、821、833	增强子结合蛋白基因 enhancer binding protein

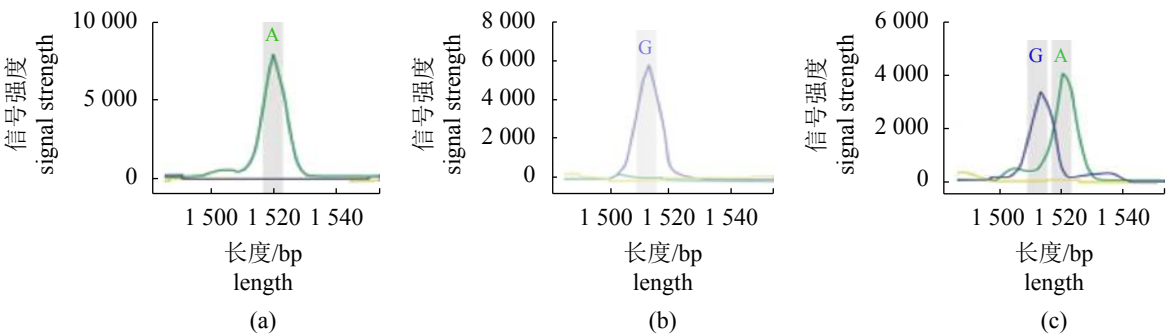


图 2 Unigene00810126-2903 标记基因型检测图

(a) AA 基因型; (b) GG 基因型; (c) AG 基因型

Fig. 2 Unigene00810126-2903 marker genotype detection map

(a) AA genotype; (b) GG genotype; (c) AG genotype

记的各个基因型频率均符合 Hardy-Weinberg 定律 ($P>0.05$) (表 3)。

2.3 SNP 标记与生长性状的关联分析

在极端群体中将 30 个 SNP 标记与生长性状进行关联分析, 筛选出 8 个与草鱼生长性状存在潜在性相关的标记, 然后进一步扩大群体进行验证和分析, 8 个 SNP 标记中仅有 4 个 SNP 标记与草鱼生长性状存在显著相关性, 分别是 uni-

gene00810126-8014、unigene00810126-2903、unigene00870394-525 和 unigene02938762-11628, 位于生长催乳素 α 基因 (*sla*)、早期生长反应蛋白-1 基因 (*egr-1*)和肌球蛋白重链基因 (*myh*) 上。unigene00810126-8014 标记的 CC 和 CT 基因型个体的体质量、体长、体高、头长和尾柄长的性状均值都显著高于 TT 基因型个体 ($P<0.05$), CC 基因型个体的体质量、体高和尾柄高性状均值

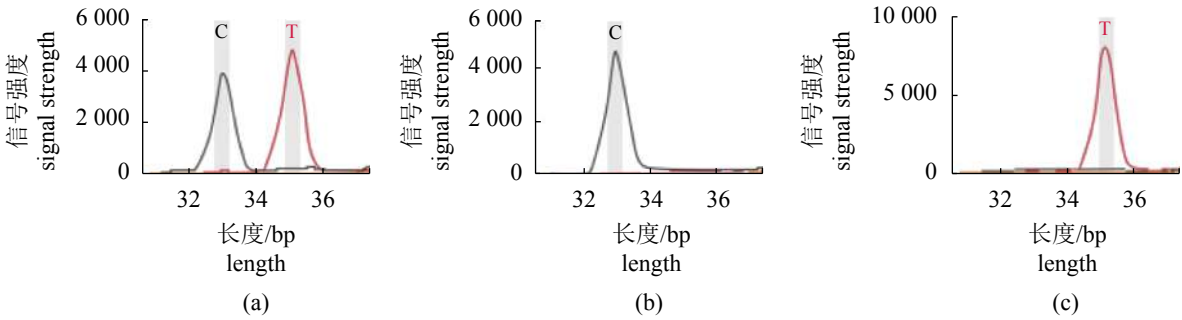


图3 Unigene00810126-8014 标记基因型检测图

(a) CT 基因型; (b) CC 基因型; (c) TT 基因型

Fig. 3 Unigene00810126-8014 marker genotype detection map

(a) CT genotype; (b) CC genotype; (c) TT genotype

表3 8个SNPs位点在草鱼群体中的遗传多态性

Tab. 3 Genetic polymorphisms of 8 SNPs in the *C. idella* population

SNP位点 SNP locus	碱基突变类型 mutation type	有效等位基因数/个 N_e	期望杂合度 H_e	观测杂合度 H_o	多态信息含量 PIC	Hardy-Weinberg平衡 (P -value)
unigene00810126-8014	C/T	1.995	0.503	0.417	0.374	0.180
unigene00810126-2903	A/G	1.704	0.417	0.450	0.328	0.539
unigene00810126-568	C/T	1.621	0.386	0.417	0.310	0.539
Unigene00870394-525	A/G	1.471	0.323	0.300	0.269	0.579
unigene02938762-11628	C/T	1.910	0.481	0.383	0.363	0.114
unigene02722463-1898	T/G	1.260	0.208	0.200	0.185	0.762
unigene04468671-2533	G/T	1.744	0.430	0.450	0.335	0.717
unigene04620686-435	A/T	1.363	0.269	0.283	0.231	0.667
平均值 mean		1.634	0.377	0.363	0.300	0.445

都比 CT 基因型个体的均值高 ($P>0.05$)。unigene 00810126-2903 标记的 AA 基因型个体的体质量、体长、体高、头长和尾柄长性状均值都比其他 2 个基因型的均值高, 且 AA 基因型个体的体质量显著高于 GG 基因型个体 ($P<0.05$)。unigene00870394-525 标记的 AA 基因型个体的体质量、体长、体高、头长、尾柄长和尾柄高的性状均值都比其他 2 个基因型的均值高, 且 AA 基因型个体的体质量和体长性状均值都显著高于 GG 基因型个体 ($P<0.05$)。unigene02938762-011628 标记的 TT 基因型个体在体质量、体长、体高、头长和尾柄长的性状均值都比其他 2 个基因型个体的均值高, 且 TT 基因型与 CC 基因型个体在体质量、体长和头长性状上的差异达到显著水平 ($P<0.05$)

(表 4)。

3 讨论

3.1 利用 RNA-Seq 数据进行 SNP 标记开发

利用 RNA-seq 测序技术在草鱼肝脏、肌肉和脑组织转录组数据库中的差异表达基因上获得 4 580 个 SNP, 平均每 3.73 kb 序列中出现一个 SNP (1/3 730 bp), 明显低于大口黑鲈 (1/2 000 bp)^[21]、大黄鱼 (*Larimichthys crocea*)(1/506 bp)^[22]、半滑舌鳎 (*Cynoglossu semilaevis*)(1/491 bp)^[23]、大西洋鲑 (*Salmo salar*)(1/614 bp) 中 SNPs 的分布密度^[24], SNPs 在不同物种间分布密度差异较大, 可能与测序质量和过滤标准等有关。从差异表达基因的 SNP 标记中挑选出 34 个进行多态性验证, 发现其中

表 4 4 个 SNP 标记不同基因型与生长性状的相关性分析

Tab. 4 Correlation analysis between different genotype and growth traits of 4 SNP markers

SNP位点 SNP locus	基因型 genotype	样本数 number	体质量/kg body mass	体长/cm body length	体高/cm body height	头长/cm head length	尾柄长/cm caudal peduncle length	尾柄高/cm caudal peduncle height
unigene00810126-8014	CC	134	2.49±0.03 ^a	55.43±0.34 ^a	13.78±0.09 ^a	10.67±0.08 ^{ab}	8.20±0.07 ^{ab}	6.81±0.09
	CT	112	2.42±0.03 ^a	55.47±0.37 ^a	13.75±0.10 ^a	10.81±0.08 ^a	8.32±0.08 ^a	6.73±0.10
	TT	54	2.24±0.05 ^b	54.11±0.53 ^b	13.32±0.14 ^b	10.42±0.13 ^b	8.01±0.11 ^b	6.95±0.14
unigene00810126-2903	AA	168	2.45±0.03 ^a	55.52±0.31	13.77±0.08	10.79±0.07 ^a	8.33±0.06 ^a	6.72±0.08
	AG	112	2.40±0.03 ^{ab}	54.86±0.37	13.55±0.10	10.55±0.88 ^{ab}	8.08±0.08 ^{ab}	6.91±0.09
	GG	20	2.28±0.08 ^b	54.55±0.88	13.69±0.24	10.45±0.21 ^b	7.99±0.18 ^b	6.92±0.23
unigene00870394-525	AA	15	2.59±0.09 ^a	57.11±1.01 ^a	14.16±0.27	10.99±0.24	8.44±0.21	6.84±0.26
	AG	97	2.48±0.04 ^{ab}	55.41±0.39 ^{ab}	13.69±0.11	10.64±0.09	8.25±0.08	6.83±0.11
	GG	188	2.38±0.03 ^b	54.95±0.29 ^b	13.64±0.08	10.67±0.07	8.18±0.06	6.79±0.08
unigene02938762-11628	CC	65	2.33±0.34 ^b	54.16±3.72 ^b	13.51±1.16	10.43±0.89 ^b	8.04±0.75	6.70±1.00
	CT	135	2.44±0.36 ^a	55.10±3.61 ^a	13.66±1.03	10.62±0.89 ^a	8.24±0.86	6.85±1.17
	TT	100	2.46±0.36 ^a	56.03±4.34 ^a	13.83±1.04	10.92±0.98 ^a	8.29±0.78	6.82±0.86

注: 表中上标为平均值的差异显著性(最小显著性差异法, LSD), 同一列数值中, 不同小写字母代表差异显著($P<0.05$)
Notes: Values with different superscript letters within a column indicate significant difference ($P<0.05$)

30 个标记具有多态性, 准确率达 88.24%。先前有研究对大西洋鳕 (*Gadus morhua*) 全基因组进行测序并开发和验证 SNP, 其准确率为 74.6%^[25]; 采用 Illumina HiSeq™ 2000 测序平台对大口黑鲈进行转录组测序, 对数据库中存在潜在的 SNP 进行挖掘和验证, 其准确率为 70.0%^[20]; 运用 Solexa 高通量测序技术对大菱鲆 (*Scophthalmus maximus*) 转录组中 SNP 进行发掘, 经分析验证其准确率为 46.7%^[26], 上述结果中 SNP 标记的准确率均低于本研究中 SNP 标记的准确率, 可能与不同的测序方法、质量及深度有关。

3.2 SNP 标记的遗传多样性分析

杂合度反映了遗传一致程度, 而多态信息含量是衡量多态性的指标。杂合度越高, 群体遗传一致程度越低, 遗传多样性越高; 多态信息含量越大, 则该位点的杂合子比例越大, 遗传信息就越丰富^[20]。本研究中 8 个 SNP 标记在草鱼群体中的 PIC 为 0.185~0.374, 平均值为 0.300。除 unigene02722463-1898、unigene04620686-435 为低度多态 ($PIC<0.25$) 以外, 其他均为中度多态 ($0.25<PIC<0.5$), 未发现高度多态性标记。本研究结果与张猛^[27] 利用 SNP 标记分析人工选育草鱼群体的遗传多样性结果相似, 所选择的标记

均没有表现出高度多态性。Yu 等^[28] 采用 SNP 标记分析长江和珠江 6 个草鱼群体的遗传多样性, 所有群体的平均 H_o 和 H_e 分别为 0.605 6 和 0.451 0, 均明显高于本研究中草鱼群体的平均 H_o 和 H_e 值, 表明草鱼人工选育群体遗传多样性低于野生群体, 推测可能是人工定向选育造成草鱼群体遗传多样性的降低^[27]。

3.3 生长相关基因中 SNP 标记对草鱼生长性状的影响

生长催乳素 α (Sl α) 是生长激素 (Gh)/催乳素 (Prl) 家族的成员之一, 在生长发育和代谢等生物过程中发挥重要作用^[29]。已有研究发现, 长期饥饿导致莫桑比克罗非鱼 (*Oreochromis mossambicus*) 肝脏中 *sla* 的 mRNA 水平显著升高, 复投喂后, *sla* 的表达水平下降至正常水平^[30]。在金头鲷 (*Sparus aurata*)^[31] 和虹鳟^[32] 中长期饥饿均导致血浆中 Sl α 的蛋白水平显著升高, 复投喂后, Sl α 的蛋白水平下降至正常水平, 在金头鲷中血浆 Gh 的表达水平变化趋势与 Sl α 相反, Sl α 与 Gh 的功能作用相反^[33]。在罗非鱼肝细胞中添加重组的同源蛋白 SL, 可引起胰岛素生长因子 I 基因 (*igf I*) 和胰岛素生长因子 II 基因 (*igf II*) 的表达水平升高, 采用 SL 抗血清免疫中和罗非鱼肝细胞

中表达的内源性 SL 蛋白, 导致 *igf I* 和 *igf II* 的表达水平受到抑制^[34]。然而, 在金头鲷幼鱼中腹腔注射一定剂量的 *sla* 并未引起 *igf I* 表达水平的改变, 对于 *sl* 和 *igf* 之间的相互作用还存在一定的争议^[35]。本研究发现, 慢长组草鱼脑组织中 *sla* 表达水平比快长组显著上调, 而快长组草鱼脑组织中 *gh* 的相对表达水平是慢长组的 2 倍, 说明 *sla* 与 *gh* 之间可能存在一定的互补, *sla* 对生长起一定的负调控作用^[31], 且该基因中的 unigene00810126-8014 标记和 unigene00810126-2903 标记与生长性状显著相关, 说明 *sla* 是影响草鱼生长发育的一个关键基因。

早期生长反应蛋白-1(Egr-1)属于转录因子锌指蛋白家族成员, 是正向调控 *igf II*、成纤维细胞生长因子基因(*fgf*)和表皮生长因子受体基因(*egfr*)的重要表达调控因子^[36-37]。Egr-1 作为一种激活因子, 通过刺激胰岛素样生长因子受体 I 基因(*igf- I r*)表达激活 PI3K+/AKT 信号通路, 进而促进生长发育^[38]。Egr-1 也可以通过抑制转化生长因子 β II 基因(*tgfb II*)型受体表达, 从而限制 *tgfb* 的生长抑制作用, 发挥促进机体生长效应^[39]。本研究结果显示, 草鱼快长组脑组织中 *egr-1* 的表达水平比慢长组显著上调, 受其调控的 *igf II*、*fgf* 和 *egfr* 的表达水平变化趋势与 *egr-1* 一致, 且 *egr-1* 上的 unigene00870394-525 标记与体质量和体长性状显著相关, 表明 *egr-1* 是影响草鱼生长性状的一个重要功能基因。

肌球蛋白重链(*myh*)是肌肉细胞的重要组成部分, 是一种多功能马达蛋白, 为肌肉增长和细胞分裂等提供动力^[40]。*myh* 是影响鱼类生长的重要功能基因之一。陈之航等^[41]研究了 *myh* 在翘嘴鲌(*Siniperca chuatsi*)早期发育过程中的表达水平, 发现其表达量高低与鲌生长快慢密切相关。Overturf 等^[42]利用虹鳟肌肉组织中 *myh* 的表达量监测其生长速度情况, 结果发现, 摄食量大且生长快的个体中 *myh* 的表达水平最高。*myh* 在大口黑鲈快长组和慢长组肌肉组织中的表达水平同样存在显著差异^[11]。本研究结果显示, *myh* 在草鱼快长组中的表达水平显著高于慢长组, 这与在鲌、虹鳟和大口黑鲈中的研究结果一致, 表明 *myh* 在草鱼生长发育过程中发挥重要作用。本研究进一步筛选该基因中的 SNP 位点并与生长性状进行关联分析, 结果发现该基因存在 5 个 SNP 标记, 其中 unigene02938762-11628 标记

与体质量和全长性状显著相关, 这与鲌和大口黑鲈 *myh* 中均存在与生长性状显著相关的 SNP 标记的研究结果一致^[41, 43]。进一步说明 *myh* 是调控生长性状的重要基因, 提示其可作为草鱼分子标记辅助育种研究中的目标基因并应用于选育工作中。

本研究在 *sla*、*egr-1* 和 *myh* 3 个基因中筛选到了 4 个与草鱼生长性状显著相关的 SNP 标记, 其中 unigene00810126-8014 和 unigene02938762-11628 标记位于 3'UTR, 不具有编码蛋白质的功能, unigene00810126-2903 和 unigene00870394-525 标记为同义突变, 这 4 个标记均不会引起氨基酸序列的改变, 但功能基因的 3'非编码区突变和编码区的同义突变都可能影响到基因的 mRNA 翻译等而使蛋白质的功能发生变化^[44], 进而影响到草鱼的生长性状。

本研究采用 RNA-Seq 技术对草鱼快长组与慢长组群体肌肉、肝脏和脑组织的转录组进行比较分析, 分别筛选到 1 013、552 和 372 个差异表达基因, 从这些差异表达基因中共检测得到 4 580 个 SNP 标记, 通过关联分析方法发现差异表达基因 *sla*、*egr-1* 和 *myh* 中存在 4 个与生长性状显著相关的 SNP 标记, 可应用于草鱼分子标记辅助育种研究。

参考文献 (References):

- [1] 帅方敏, 李新辉, 黄艳飞, 等. 珠江水系四大家鱼资源现状及空间分布特征研究[J]. 水生生物学报, 2017, 4(6): 172-180.
Shuai F M, Li X H, Huang Y F, et al. Resource status and spatial distribution characteristics of four major Chinese carps in the Pearl River[J]. Acta Hydrobiologica Sinica, 2017, 4(6): 172-180(in Chinese).
- [2] 黄晨, 李思易. 分子标记辅助选择技术的应用研究[J]. 生命科学仪器, 2008, 6(12): 47-50.
Huang C, Li S Y. Molecular marker-assisted selection technology research[J]. Life Science Instruments, 2008, 6(12): 47-50(in Chinese).
- [3] 刘志文, 傅廷栋, 刘雪平, 等. 作物分子标记辅助选择的研究进展、影响因素及其发展策略[J]. 植物学通报, 2005, 22(S1): 82-90.
Liu Z W, Fu T D, Liu X P, et al. Advances, influence factors and development strategies in crop molecular-
中国水产学会主办 sponsored by China Society of Fisheries

- assisted selection[J]. Chinese Bulletin of Botany, 2005, 22(S1): 82-90(in Chinese).
- [4] 张晓萌, 马普, 王洪迪, 等. SNPs在水产动物中的研究进展[J]. 生物技术通报, 2013(8): 7-11.
- Zhang X M, Ma P, Wang H D, *et al.* Progresses of SNPs studies in aquaculture animals[J]. Biotechnology Bulletin, 2013(8): 7-11(in Chinese).
- [5] Mastrochirico-Filho V A, Hata M E, Sato L S, *et al.* SNP discovery from liver transcriptome in the fish *Piaractus mesopotamicus*[J]. Conservation Genetics Resources, 2016, 8(2): 109-114.
- [6] 程方圆, 陶紫玉, 李晨虹. 应用单核苷酸多态性(SNP)标记鉴定短颌鲚、湖鲚和刀鲚[J]. 上海海洋大学学报, 2019, 28(1): 10-19.
- Cheng F Y, Tao Z Y, Li C H. Species identification of *Coilia brachygnathus*, *C. nasus* and *C. nasus taihuensis* with SNP markers[J]. Journal of Shanghai Ocean University, 2019, 28(1): 10-19(in Chinese).
- [7] Sudhagar A, Kumar G, El-Matbouli M. Transcriptome analysis based on RNA-Seq in understanding pathogenic mechanisms of diseases and the immune system of fish: a comprehensive review[J]. International Journal of Molecular Sciences, 2018, 19(1): 245.
- [8] Salem M, Vallejo R L, Leeds T D, *et al.* RNA-Seq identifies SNP markers for growth traits in rainbow trout[J]. PLoS One, 2012, 7(5): e36264.
- [9] Ulloa P E, Rincón G, Islas-Trejo A, *et al.* RNA sequencing to study gene expression and SNP variations associated with growth in zebrafish fed a plant protein-based diet[J]. Marine Biotechnology, 2015, 17(3): 353-363.
- [10] Lv J J, Zhang D N, Gao B Q, *et al.* Transcriptome and massARRAY analysis for identification of transcripts and SNPs for growth traits of the swimming crab *Portunus trituberculatus*[J]. Gene, 2015, 566(2): 229-235.
- [11] Li S J, Lu H, Bai J J, *et al.* Transcriptome assembly and Identification of genes and SNPs associated with growth traits in largemouth bass (*Micropterus salmoides*)[J]. Genetica, 2017, 145(2): 175-187.
- [12] Trapnell C, Roberts A, Goff L, *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks[J]. Nature Protocols, 2012, 7(3): 562-578.
- [13] Conesa A, Götz S, García-Gómez J M, *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research[J]. Bioinformatics, 2005, 21(18): 3674-3676.
- [14] Ye J, Fang L, Zheng H K, *et al.* WEGO: a web tool for plotting GO annotations[J]. Nucleic Acids Research, 2006, 34(S2): W293-W297.
- [15] Robinson M D, McCarthy D J, Smyth G K. EdgeR: a bioconductor package for differential expression analysis of digital gene expression data[J]. Bioinformatics, 2010, 26(1): 139-140.
- [16] Mortazavi A, Williams B A, McCue K, *et al.* Mapping and quantifying mammalian transcriptomes by RNA-Seq[J]. Nature Methods, 2008, 5(7): 621-628.
- [17] McCormick R F, Truong S K, Mullet J E. RIG: recalibration and interrelation of genomic sequence data with the GATK[J]. G3: Genes, Genomes, Genetics, 2015, 5(4): 655-665.
- [18] Yeh F C, YANG R C, Boyle T B J, *et al.* PopGene32, Microsoft Windows-based freeware for population genetic analysis. version 1.32[M]. Edmonton, Canada: University of Alberta, 2000.
- [19] 季晓芬, 段辛斌, 刘绍平, 等. 基于微卫星评估草鱼放流亲本对野生群体遗传多样性的影响[J]. 水产学报, 2018, 42(1): 10-17.
- Ji X F, Duan X B, Liu S P, *et al.* Genetic effect of released brood grass carp (*Ctenopharyngodon idella*) on wild population in the Yangtze River inferred from microsatellite markers[J]. Journal of Fisheries of China, 2018, 42(1): 10-17(in Chinese).
- [20] 全迎春, 马冬梅, 白俊杰, 等. 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, 2016, 40(6): 1128-1134.
- Quan Y C, Ma D M, Bai J J, *et al.* SNPs identification in RNA-Seq data of largemouth bass (*Micropterus salmoides*) fed on formulated feed and association analysis with growth trait[J]. Acta Hydrobiologica Sinica, 2016, 40(6): 1128-1134(in Chinese).
- [21] 李胜杰, 白俊杰, 赵莹, 等. 大口黑鲈EST-SNP标记开发及其与生长性状的相关性分析[J]. 海洋渔业, 2018, 40(1): 38-46.
- Li S J, Bai J J, Zhao L, *et al.* Development of EST-SNPs in largemouth bass (*Micropterus salmoides*) and analysis of their correlation with growth traits[J]. Marine

- [Fisheries](#), 2018, 40(1): 38-46(in Chinese).
- [22] 王盼盼. 大黄鱼基因组和转录组 SNP 的挖掘与应用 [D]. 厦门: 集美大学, 2016.
- Wang P P. Development and application of SNP in large yellow croaker (*Larimichthys crocea*) based on genome and transcriptome[D]. Xiamen: Jimei University, 2016 (in Chinese).
- [23] 庄伟. 半滑舌鳎 (*Cynoglossus semilaevis*) 基因组 SNP 标记的开发与检测 [D]. 青岛: 中国海洋大学, 2010.
- Zhuang W. Isolation and characterization of SNPs from half-smooth tongue sole (*Cynoglossus semilaevis*)[D]. Qingdao: Ocean University of China, 2010 (in Chinese).
- [24] Leu J H, Chen S H, Wang Y B, *et al.* A review of the major penaeid shrimp EST studies and the construction of a shrimp transcriptome database based on the ESTs from four penaeid shrimp[J]. [Marine Biotechnology](#), 2011, 13(4): 608-621.
- [25] Hubert S, Higgins B, Borza T, *et al.* Development of a SNP resource and a genetic linkage map for Atlantic cod (*Gadus morhua*)[J]. [BMC Genomics](#), 2010, 11: 191.
- [26] 王婷, 黄智慧, 马爱军, 等. 基于转录组数据的大菱鲂 (*Scophthalmus maximus*)SNP标记开发及多态性分析[J]. 海洋与湖沼, 2014, 45(6): 1300-1307.
- Wang T, Huang Z H, Ma A J, *et al.* Development and polymorphic analysis of SNP markers in *Scophthalmus maximus* based on transcriptome database[J]. *Oceanologia et Limnologia Sinica*, 2014, 45(6): 1300-1307(in Chinese).
- [27] 张猛. 草鱼 SNP 标记开发及与生长性状关联分析 [D]. 上海: 上海海洋大学, 2016.
- Zhang M. Development of SNP markers and their association with growth traits in grass carp (*Ctenopharyngodon idella*)[D]. Shanghai: Shanghai Ocean University, 2016 (in Chinese).
- [28] Yu L Y, Bai J J, Cao T T, *et al.* Genetic variability and relationships among six grass carp *Ctenopharyngodon idella* populations in China estimated using EST-SNP markers[J]. [Fisheries Science](#), 2014, 80(3): 475-481.
- [29] Zhu Y, Stiller J W, Shaner M P, *et al.* Cloning of somatolactin alpha and beta cDNAs in zebrafish and phylogenetic analysis of two distinct somatolactin subtypes in fish[J]. [Journal of Endocrinology](#), 2004, 182(3): 509-518.
- [30] Pierce A L, Fox B K, Davis L K, *et al.* Prolactin receptor, growth hormone receptor, and putative somatolactin receptor in Mozambique tilapia: tissue specific expression and differential regulation by salinity and fasting[J]. [General and Comparative Endocrinology](#), 2007, 154(1-3): 31-40.
- [31] Company R, Astola A, Pendón C, *et al.* Somatotrophic regulation of fish growth and adiposity: growth hormone (GH) and somatolactin (SL) relationship[J]. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology*, 2001, 130(4): 435-445.
- [32] Smal J, Rand-Weaver M, Sumpter J P, *et al.* Somatolactin and growth hormone are differentially correlated to various metabolic parameters in trout[J]. [Netherlands Journal of Zoology](#), 1994, 45(1-2): 129-131.
- [33] Mingarro M, de Celis S V R, Astola A, *et al.* Endocrine mediators of seasonal growth in gilthead sea bream (*Sparus aurata*): the growth hormone and somatolactin paradigm[J]. [General and Comparative Endocrinology](#), 2002, 128(2): 102-111.
- [34] Peng J P, Lian A J, Jiang Q. Production and purification of recombinant somatolactin and its effects on insulin-like growth factors gene expression in tilapia hepatocytes[J]. *Fisheries and Aquaculture Journal*, 2016, 7(2): 1000166.
- [35] de Celis V R S, Gómez V R P, Caldach-Giner V R J A, *et al.* Expression and characterization of European sea bass (*Dicentrarchus labrax*) somatolactin: assessment of *in vivo* metabolic effects[J]. [Marine Biotechnology](#), 2003, 5(1): 92-101.
- [36] Liu C T, Calogero A, Ragona G, *et al.* EGR-1, the reluctant suppression factor: EGR-1 is known to function in the regulation of growth, differentiation, and also has significant tumor suppressor activity and a mechanism involving the induction of TGF-beta1 is postulated to account for this suppressor activity[J]. [Critical Reviews™ in Oncogenesis](#), 1996, 7(1-2): 101-126.
- [37] Min I M, Pietramaggiori G, Kim F S, *et al.* The transcription factor EGR-1 controls both the proliferation and localization of hematopoietic stem cells[J]. [Cell Stem Cell](#), 2008, 2(4): 380-391.
- [38] 马旻. EGR-1 诱导 IGF-1R 表达促进前列腺癌细胞增生 [D]. 合肥: 中国科学技术大学, 2011.

- Ma Y. EGR-1 promotes cell proliferation and tumor forming through up-regulation of *IGF-1R* in prostate cancer cells[D]. Hefei: University of Science and Technology of China, 2011 (in Chinese).
- [39] Du B H, Fu C Z, Kent K C, *et al.* Elevated Egr-1 in human atherosclerotic cells transcriptionally represses the transforming growth factor- β type II receptor[J]. *The Journal of Biological Chemistry*, 2000, 275(50): 39039-39047.
- [40] Yu I M, Planelles-Herrero V J, Sourigues Y, *et al.* Myosin 7 and its adaptors link cadherins to actin[J]. *Nature Communications*, 2017, 8: 15864.
- [41] 陈之航, 董浚键, 孙成飞, 等. 基于转录组测序对翘嘴鲌(*Siniperca chuatsi*)2种肌球蛋白重链基因的克隆与分析[J]. *渔业科学进展*, 2017, 38(3): 51-61.
- Chen Z H, Dong J J, Sun C F, *et al.* cDNA cloning and analyses of two myosin heavy chain isoforms of mandarin fish (*Siniperca chuatsi*) based on transcriptome sequencing[J]. *Progress in Fishery Sciences*, 2017, 38(3): 51-61(in Chinese).
- [42] Overturf K, Hardy R W. Myosin expression levels in trout muscle: a new method for monitoring specific growth rates for rainbow trout *Oncorhynchus mykiss* (Walbaum) on varied planes of nutrition[J]. *Aquaculture Research*, 2001, 32(4): 315-322.
- [43] 李胜杰, 姜鹏, 樊佳佳, 等. 大口黑鲈肌球蛋白重链基因SNPs的筛选及与生长性状的关联[J]. *水产学报*, 2018, 42(3): 305-313.
- Li S J, Jiang P, Fan J J, *et al.* SNPs detection of *MYH* gene and its association with growth traits in largemouth bass (*Micropterus salmoides*)[J]. *Journal of Fisheries of China*, 2018, 42(3): 305-313(in Chinese).
- [44] Pesole G, Mignone F, Gissi C, *et al.* Structural and functional features of eukaryotic mRNA untranslated regions[J]. *Gene*, 2001, 276(1-2): 73-81.

RNA sequencing to identify genes and SNP markers associated with growth traits in *Ctenopharyngodon idella*

SUN Xue^{1,2}, LI Shengjie^{1*}, JIANG Peng¹, DU Jinxing¹, ZHOU Jiahui^{1,2}, BAI Junjie¹

(1. Key Laboratory of Aquatic Genomics, Ministry of Agriculture, Key Laboratory of Tropical & Subtropical Fishery Resource Application & Cultivation of Ministry of Agriculture and Rural Affairs, Pearl River Fisheries Research Institute of Chinese Academy of Fishery Sciences, Guangzhou 510380, China;
2. National Demonstration Center for Experimental Fisheries Science Education, Shanghai Ocean University, Shanghai 201306, China)

Abstract: In order to identify genes associated with growth traits and SNP (Single nucleotide polymorphism) markers in *Ctenopharyngodon idella*, RNA-Seq (RNA Sequencing technology) was used to analyze the transcriptomes of liver, muscle and brain tissues of fast-growing and slow-growing *C. idella* groups. A total of 314.65 million clean reads were gained for the fast-growing and slow-growing groups, and 34 147 unigenes were assembled, with an average length of 1 060 bp, of which 30 751 genes were annotated. 1 013, 552, and 372 significant differentially expressed genes were screened in liver, muscle, and brain tissues, respectively, of which 4 580 SNP markers were detected. Then, the polymorphisms of 34 SNP markers was detected and validated with SNaPshot technique in the extreme populations of growth traits of *C. idella* (n=300), and among which, 30 SNPs (83.24%) were successfully genotyped. A general linear model was used to analyze the correlation between 30 SNPs and growth traits. The results showed that at the position of Unigene00810126-8014, the body weight, body length, body height, head length and caudal peduncle length in individuals with CC genotype had significantly higher values than TT genotype. At the position of Unigene00810126-2903, the body weight of AA genotype was significantly higher than those with GG genotype. At the position of Unigene00870394-525, individuals with the AA genotype had significantly higher values in body weight and body height than those with GG genotype. At the position of Unigene02938762-011628, TT and CC genotypes had significant differences in body mass, body length and head length traits. Besides, other SNPs loci were not significantly correlated with growth traits. The four selected markers were located at Somatolactin alpha (*sla*), Early growth response protein-1 (*egr-1*) and Myosin heavy chain (*myh*) genes. The average polymorphism information content (*PIC*), average expected heterozygosity (H_e) and average observed heterozygosity (H_o) of 8 EST-SNP markers in *C. idella* population were 0.300, 0.377 and 0.363, respectively, indicating that genetic diversity of *C. idella* breeding population had relatively high level. In conclusion, 1 937 differentially expressed genes and four SNP markers related to growth traits were obtained, which can be used as candidate genes and for molecular marker-assisted selection of *C. idella*, and can also be used in subsequent breeding work.

Key words: *Ctenopharyngodon idella*; transcriptome sequencing; differentially expressed genes; single nucleotide polymorphism; growth traits

Corresponding author: LI Shengjie. E-mail: ssjli@163.com

Funding projects: China Agriculture Research System (CARS-45-04); Central Public-interest Scientific Institution Basal Research Fund, CAFS (2020TD34, 2019ZX-001); Science and Technology Planning Project of Guangdong Province (201906010004)