婴鲤属的全长转录组学分析

叶昀[①] 张素蓉[①] 万柏祺[①] 张亚琪[①] 史文天[③]
杨慧[①*] 张万昌[②*]

① 江西农业大学生物科学与工程学院 南昌 330045；② 南昌大学生命科学学院 南昌 330031；
③ 德国图宾根大学 德国图宾根 72074

摘要：婴鲤属（*Paedocypris*）是世界上已知最小的鱼类，也是最小的脊椎动物之一。前期研究通过绘制 *P. carbunculus* 和 *P. micromegethes* 的基因组草图和比较基因组学分析表明，*Hox* 家族基因的丢失和内含子缩短/重复序列减少可能是分别导致婴鲤属体型小型化和基因组简化的潜在原因。然而，其体型小型化在转录表达层面上的分子机制尚未得到探究。本研究对 *P. carbunculus* 和 *P. micromegethes* 样本分别开展全长转录组测序，获得两种婴鲤的全长转录组文库。经校正去冗余后，分别获得 38 651 和 23 165 条校正后共识序列，平均长度分别为 2 413 bp 和 2 460 bp，这些序列分别有 35 883（92.84%）和 20 381（87.98%）条比对到了参考基因组上。对这两个物种全长转录组数据进行分类及特征分析，分别获得 19 352 和 11 139 条全长转录本。本研究在 *P. carbunculus* 和 *P. micromegethes* 的全长转录本数据中分别检测到 9 404 和 6 068 个可变剪切事件，以 A5 型可变剪接事件为主。还观察到大多数基因（分别占 74.55% 和 84.75%）只有一个可变多聚腺苷酸化位点。对转录因子进行注释和分类显示，在婴鲤属中数量最多的转录因子家族为 zf-C2H2，其次是 ZBTB。还在这两个婴鲤属物种中分别预测到 1 382 和 1 649 条长链非编码 RNA。通过将婴鲤与鲤（*Cyprinus carpio*）、鲢（*Hypophthalmichthys molitrix*）、大西洋鲑（*Salmo salar*）和银大麻哈鱼（*Oncorhynchus kisutch*）的全长转录本数据进行比较，发现婴鲤 *Hox* 家族表达的基因数量远小于其他 4 种鱼，这表明在基因表达层面上，*Hox* 家族基因的缺失可能也是从功能上导致婴鲤属体型小型化的原因之一。研究还发现婴鲤属表达更多与重要发育通路相关的基因，这些与发育功能相关的基因相比转录组中其他基因拥有更少的基因融合事件。本研究报道了婴鲤属两个物种的全长转录组全景图谱及相关分析结果，为进一步解析婴鲤体型小型化的分子机制提供了新的见解。

关键词：婴鲤；小型化；全长转录组；*Hox* 家族基因

中图分类号：Q953 **文献标识码：**A **文章编号：**0250-3263（2025）02-224-15

Full-Length Transcriptomic Analysis of the Genus *Paedocypris*

YE Yun[①] ZHANG Su-Rong[①] WAN Bo-Qi[①] ZHANG Ya-Qi[①]
SHI Wen-Tian[③] YANG Hui[①*] ZHANG Wan-Chang[②*]

① *School of Bioscience and Bioengineering, Jiangxi Agricultural University, Nanchang* 330045;
② *School of Life Sciences, Nanchang University, Nanchang* 330031, *China*; ③ *University of Tübingen, Tübingen* 72074, *Germany*

基金项目 江西省杰出青年基金项目（No. 20232ACB215003），江西省自然科学基金面上项目（No. 20224BAB205013）；

* 通讯作者，E-mail: 389846652@qq.com，wanchangzhang@ncu.edu.cn；

第一作者介绍 叶昀，男，硕士研究生；研究方向：生物信息学；E-mail: yeyun@stu.jxau.edu.cn。

收稿日期：2024-03-06，修回日期：2024-10-06 DOI: 10.13859/j.cjz.202524047 CSTR: 32109.14.cjz.24047

Abstract: [Objectives] The genus *Paedocypris* comprises the smallest known fish species and one of the smallest vertebrates, with adult sizes ranging between 8–10 mm. Previous studies have indicated that the loss of *Hox* gene family members and the reduction of intron length and repetitive sequences might be potential causes for the miniaturization of the *Paedocypris* species and the simplification of their genomes, as demonstrated through draft genome assembly and comparative genomic analysis of *P. carbunculus* and *P. micromegethes*. The objectives of this study include characterizing the full-length transcriptome of the two species, *P. carbunculus* and *P. micromegethes*, identifying potential regulatory changes that could contribute to their miniaturized stature, with a particularly focus on *Hox* family and developmental gene expression patterns and gene structure. **[Methods]** All samples were purchased from an aquarium market in Guangzhou, comprising seven *P. carbunculus* and eleven *P. micromegethes* live specimens (Fig. 1). RNA was extracted from whole fish, and the purity, concentration, and integrity were tested. RNA from the seven *P. carbunculus* and the eleven *P. micromegethes* was mixed respectively and used for the construction of a full-length transcriptome library. The sequencing was performed on the PacBio Sequel platform. Quality control of full-length transcriptome sequencing data from the two Paedocypris species were performed. Then, we described the data quality, mapping statistics, annotation statistics and transcript classification of the full-length transcriptome data. By employing different software and pipelines, we annotated the gene structure, transcription factors, alternative splicing sites, long non-coding RNAs (LncRNAs), and gene fusions. To characterize the expression situation of *Hox* family genes in *Paedocypris*, we downloaded 81 zebrafish *Hox* family gene coding sequences belonging to 48 different *Hox* genes and align to the full-length transcriptome sequences of *Paedocypris* and other four species using Blastn with a threshold e-value < 10^{-20}. To explore the expression of genes in important developmental pathways in the *Paedocypris* genus, we selected seven key GO developmental pathways encompassing 1 581 non-redundant zebrafish genes and align with the full-length transcriptome data of *Paedocypris* and the other four species for analysis. To explore whether there are significant structural differences between the developmental-related genes mapped in the *Paedocypris* genus and other genes on the *Paedocypris* genome, based on the GO annotation information of the *Paedocypris* reference genome, we conducted significance tests (*t*-test) on the developmental-related genes mapped in the *Paedocypris* genus against all annotated genes on the *Paedocypris* reference genome for alternative splicing, polyadenylation, exon count, and fusion genes. **[Results]** In this study, we conducted full-length transcriptome sequencing on whole-body samples of *P. carbunculus* and *P. micromegethes* (Tables 1–4), obtaining the first full-length transcriptome library for *Paedocypris*. After correction and deduplication, 38 651 and 23 165 corrected consensus sequences were obtained, with average lengths of 2 413 bp and 2 460 bp, respectively. Among these sequences, 35 883 (92.84%) and 20 381 (87.98%) were mapped to the *Paedocypris* reference genome (Table 2). The classification and characterization analyses of the full-length transcriptome data for these two species resulted in 19 352 and 11 139 full-length transcripts, respectively (Tables 4 and 5). In the full-length transcriptome data of *P. carbunculus* and *P. micromegethes*, 9 404 and 6 068 alternative splicing events were identified, predominantly of the A5 type (Table 6). We also observed that the majority of genes (74.55% and 84.75%, respectively) had only one alternative polyadenylation site. Annotation and classification of transcription factors revealed that the most abundant transcription factor family in *Paedocypris* was zf-C2H2, followed by ZBTB. Additionally, we predicted 1 382 and 1 649 LncRNAs in the two *Paedocypris* species, respectively (Fig. 2). Comparative analysis of the full-length

transcriptome data between *Paedocypris* and carp, silver carp, salmon, and rainbow trout revealed that the number of expressed genes in the *Hox* gene family in *Paedocypris* was significantly lower than that in the other four fish species (Fig. 3). We also identified that *Paedocypris* has the most expressed developmental genes belonging to the seven GO terms than the four fish species (Fig. 3). We found that those developmental-related genes have fewer gene fusion events than other genes in the full-length transcriptome of *Paedocyrpis* (Fig. 4). **[Conclusion]** We provide a full-length transcriptome landscape of the genus *Paedocypris* by performing full-length transcriptome sequencing. Data quality, mapping statistics, annotation statistics, and transcript classification of the full-length transcriptome data of *Paedocypris* are described. The analysis highlights that the loss of *Hox* genes at the expression level may be one of the functional reasons for the miniaturization of *Paedocypris*. This study provides new insights into the molecular mechanisms underlying the miniaturization of *Paedocypris*.

Key words: *Paedocypris*; Miniaturization; Full-length transcriptome; *Hox* gene family

物种体型小型化是指在进化过程中物种体型减小的一种现象，该现象影响着物种的生存能力、食性、繁殖、能量利用和生态位，因此引起了科研人员的广泛关注（Hanken et al. 1993）。尽管体型小型化可能导致物种更容易受到捕食和其他自然灾害的影响，削弱其生存和繁殖能力，但自然界中仍存在众多体型小型化的例子，如阿马乌童蛙（*Paedophryne amauensis*）、纳米变色龙（*Brookesia nana*）和柄翅小蜂（*Tinkerbella nana*）（Rittmeyer et al. 2012，Polilov 2017，Glaw et al. 2021）。然而，当前对于物种体型小型化的进化机制尚不清楚，特别对于演化中如何选择促使物种体型小型化的问题上存在诸多疑问。

在遗传上，体型大小是一个由多基因在分子水平上共同决定的复杂性状（Bouwman et al. 2018）。对物种体型大小的研究不仅在农业动物领域具有重要意义，还有助于理解进化过程和生物多样性的形成（Cooper et al. 2010，Boegheim et al. 2017，Bouwman et al. 2018）。当前，研究人员已发现许多与体型大小相关的基因，如会导致人类侏儒症的 *fgfr*3 基因（Harada et al. 2007）；影响犬类体型差异的 *fgf*4 和 *pou*1*f*1 基因（Parker et al. 2009，Kyöstilä et al. 2021）；导致日本棕色牛和提洛尔灰牛生长受阻的 *evc*2 基因（Galdzicka et al. 2002，Murgiano et al. 2014）；以及参与骨骼发育，可能导致荷斯坦牛和家猪侏儒症的 *col*2*a*1 和 *col*10*a*1 基因（Agerholm et al. 2016，Boegheim et al. 2017）。鉴别调控物种体型大小的关键基因及分子基础具有重要研究价值，并且在农业动物生产上具有广泛的实际意义。

婴鲤属（*Paedocypris*）隶属于鲤形目（Cypriniformes）鲤科（Cyprinidae）鲃亚科（Danioninae）（Mayden et al. 2010，Britz et al. 2014），主要分布于印度尼西亚婆罗洲、邦加和苏门答腊等岛屿上，栖息于缓流黑水泥炭沼泽中（Kottelat et al. 2006，Britz 2008）。它们以极小的体型著称，被认为是已知最小的脊椎动物之一，最小成年个体长度仅为 7.9 mm（Kottelat et al. 2006）。婴鲤属目前已知包括 3 个物种：*P. carbunculus*、*P. micromegethes* 和 *P. progenetica*（Kottelat et al. 2006，Britz 2008）。这些物种的成年个体保留着部分幼体特征，如狭窄的头部、大眼睛和部分骨骼的缺失（Kottelat et al. 2006，Britz et al. 2009）。婴鲤属与斑马鱼（*Danio rerio*）同属鲤科鱼类，基于模式生物斑马鱼更为完善的研究数据信息来对邻近物种中的极端表型开展比较基因组学分析，将有利于认识体型表型变化的分子基础。此前研究发现 *P. carbunculus* 具有 15 对染色体（2n = 30），基因组大小仅为 0.43 Gb，远小于

斑马鱼基因组（1.5 Gb）（Liu et al. 2012，Malmstrøm et al. 2018），这可能归因于内含子缩短和重复序列减少（Malmstrøm et al. 2018）。*Hox* 家族是一类专门调控生物形体的基因，它们在鱼类的生长发育及躯干形成中发挥着重要作用（Sordino et al. 1995，Santini et al. 2005）。Malmstrøm 等（2018）发现，在婴鲤属基因组中丢失了多个 *Hox* 家族基因，这可能是导致其微小体型和发育截断（developmentally truncated）的原因之一。但目前在转录表达水平上探究婴鲤体型小型化的分子机制及 *Hox* 家族基因的转录表达情况还未见相关报道。此外，栖息地破坏、气候变化和人类活动等因素导致婴鲤等微小化物种种群数量下降和分布范围缩小，为维护生态系统的稳定，对这类物种开展深入研究显得至关重要（Kottelat et al. 2006）。因此，本研究采用全长转录组测序策略从表达谱层面开展全面分析，进一步解析婴鲤体型小型化的分子机制，也为深入探究物种微小化提供新的视角。

1　材料与方法

1.1　样本收集

本研究所使用样本均为广州市水族市场采购的商业样品，共 7 条 *P. carbunculus* 和 11 条 *P. micromegethes* 活体（图 1）。两种鱼的原产地分别为印度尼西亚婆罗洲的帕朗卡拉亚（Palangkaraya）和马来西亚的古晋（Kuching）。解剖去除这些鱼的肠道内容物后立即保存于液氮中。

1.2　核酸提取及建库测序

由于婴鲤属的体型较小，单个个体重量小于 5 mg，为提取足够量的核酸用于建库测序，采用德国 Qiagen 公司的 QIAsymphony RNA 提取试剂盒对整鱼进行 RNA 提取。通过琼脂糖凝胶电泳和超微量紫外-可见光分光光度计（Nanodrop 2000）对提取的 RNA 进行初步检测，使用荧光定量仪（Qubit）和微流控分析系统（Agilent 2100）对 RNA 纯度、浓度和完整性进行检测。满足建库要求后将 7 条 *P. carbunculus* 和 11 条 *P. micromegethes* 的 RNA 分别混合，使用日本 Takara 公司的 Clontech SMARTer PCR cDNA 合成试剂盒构建全长转录组文库，使用 PacBio Sequel 平台开展全长转录组测序，文库构建和测序工作均委托北京诺禾致源科技股份有限公司完成。

1.3　下机数据预处理

本研究使用 SMRTlink v7.0（--minLength 50，--maxLength 15000，--minPasses 1）对婴鲤属两个物种的全长转录组测序下机数据进行质控过滤，去掉接头和低质量 reads 得到 subreads，再进行自我纠错，得到环形一致序列（circular consensus sequence，CCS）。接着使用 pbclassify.py 脚本通过检测 CCS 是否包含建库时 PCR 用的 5′端引物、3′端引物和 mRNA 的多聚腺苷酸化尾，分类找出全长非嵌合（full-length non-chimera）序列和非全长非嵌合（non-full-length non-chimeric）序列。最后将同一转录本的全长非嵌合序列使用 hierarchical n*log(n)算法聚类和校正，获得校正后共识序列。

使用 GMAP 软件将 *P. carbunculus* 和 *P. micromegethes* 校正后共识序列分别比对到此前已构建的参考基因组上（Wu et al. 2005，Malmstrøm et al. 2018），参数为“--no-chimeras --cross-species --expand-offsets 1 -B 5 -K 50000 -f samse -n 1”，得到 Unmapped、Multiple mapped、Uniquely mapped、Reads map to '+'、Reads map to '-'类别，根据比对结果统计评估 reads 质量。

1.4　基因结构分析及功能注释

使用 TAPIS（V1.2.1）流程对 *P. carbunculus* 和 *P. micromegethes* 的全长转录本数据进行分类及特征分析，将 GMAP 输出的 bam 和 gff 文件，比对到基因组未注释区域的 reads 被认为是新基因位点，比对到已知基因区域不同外显子的 reads 被认为是新转录本。然后开展可变剪接、多聚腺苷酸化、新基因和新转录本的鉴

定（Wu et al. 2005，Abdel-Ghany et al. 2016）。为获得全面的注释信息，将未比对到参考基因组上的转录本和新基因及新转录本在以下七个数据库进行功能注释，即非冗余蛋白质数据库（non-redundant protein database，NR）（Pruitt et al. 2005）、核苷酸序列数据库（nucleotide sequences，NT）（Pruitt et al. 2005）、Pfam 蛋白质家族数据库（nucleotide sequences，Pfam）（Bateman et al. 2004）、蛋白质同源群集数据库（cluster of orthologous groups of proteins，KOG/COG）（Tatusov et al. 2003）、SwissProt 蛋白质序列数据库（SwissProt protein sequence database，SwissProt）（Bairoch et al. 1991）、京都基因与基因组百科全书（Kyoto Encyclopedia of Genes and Genomes，KEGG）（Kanehisa 2002）和基因本体论（Gene Ontology，GO）（Ashburner et al. 2000）。

1.5 鉴别转录因子、可变剪切位点及长链非编码 RNA（long non-coding RNAs，LncRNA）

由于婴鲤属未被动物转录因子数据库—animalTFDB 2.0（Zhang et al. 2015）收录，因此根据转录因子家族的 pfam 文件，利用 hmmsearch 对婴鲤全长转录组数据开展转录因子从头预测（Bateman et al. 2004）。使用 SUPPA 对全长转录组数据的可变剪接事件进行鉴定（Alamancos et al. 2015），共识别出 7 种可变剪接类型，外显子跳跃型（skipped exon，SE）、外显子互斥型（mutually exclusive exon，MX）、5′端可变剪接型（alternative 5′ splice site，A5）、3′端可变剪接型（alternative 3′ splice site，A3）、内含子滞留型（retained intron，RI）、起始外显子可变剪接型（alternative first exon，AF）以及终止外显子可变剪接型（alternative last exon，AL）。利用 PLEK（v1.2）、CNCI（V2）（Sun et al. 2013，Li et al. 2014）和 CPC（v0.9）软件（Kong et al. 2007）将预测得到的转录本序列与 Pfam-A 和 Pfam-B 数据库开展 hmmscan 同源搜索，经数据库比对后得到 LncRNA 序列（Bateman et al. 2004）。

1.6 融合基因鉴别

为鉴定 *P. carbunculus* 和 *P. micromegethes* 中的融合基因事件，使用 GMAP 将这两个物种的校正后共识序列分别比对到各自参考基因组上（Malmstrøm et al. 2018），参数为--max-intronlength-ends 50000；-f 4；-z sense_force；-n 0（Wu et al. 2005）。根据比对结果筛选出符合以下条件的基因：1）一条全长转录本 map 到参考基因组的两个或更多的基因位点；2）每个基因位点必须比对上这条转录本的 10%区域；3）这条转录本比对到参考基因组上的 coverage 必须达到 99%以上；4）每个比对的基因位点在参考基因组上必须相隔 100 kb 以上。同时符合上述 4 个条件的基因被认定为发生了基因融合事件。

1.7 *Hox* 家族基因和发育相关基因鉴别与分析

前期研究表明，在婴鲤属基因组中存在多个 *Hox* 家族基因缺失，这可能是导致婴鲤属体型小型化的原因之一（Malmstrøm et al. 2018）。为进一步确定 *Hox* 家族基因在婴鲤属中的转录表达情况，从 Ensembl（https://www.ensembl.org）中下载了 81 条斑马鱼的 *Hox* 家族基因编码序列（coding sequence，cds）作为参照，这些序列属于 *Hox* 家族的 48 个不同基因。然后，从 NCBI（https://www.ncbi.nlm.nih.gov/）下载了鲤（*Cyprinus carpio*，NCBI 登录号为 PRJNA752470）、鲢（*Hypophthalmichthys molitrix*，NCBI 登录号为 PRJNA705843）、银大麻哈鱼（*Oncorhynchus kisutch*，NCBI 登录号为 PRJNA352719）和大西洋鲑（*Salmo salar*，NCBI 登录号为 PRJNA680991）4 个物种的全长转录组数据，进行比较分析。使用 Blastn 将上述 81 条斑马鱼 *Hox* 家族基因的 cds 序列分别比对到婴鲤和上述 4 个物种的全长转录组序列上，将阈值（e-value）小于 10^{-20} 的全长转录本分别鉴定为婴鲤和 4 个物种中的 *Hox* 家族基因（McGinnis et al. 2004）。与此同时，为探究婴鲤属中重要发育相关功能通路中的基因表达情况，本研究参照 Malmstrøm 等（2018）开展的

婴鲤全基因组发育基因分析，选取了下列与发育相关的关键 GO 功能通路：GO: 0009948（前后轴形成）、GO: 0009950（背腹轴形成）、GO: 0040007（生长）、GO: 0007517（肌肉器官发育）、GO: 0007399（神经系统发育）、GO: 0001501（骨骼系统发育）和 GO: 0007379（节段特化），共包含来自斑马鱼的 1 581 个非重复基因（Ashburner et al. 2000，Malmstrøm et al. 2018）。从 Ensembl 中下载斑马鱼上述 1 581 个基因的 cds 序列，然后比对婴鲤和上述 4 个物种的全长转录组数据进行分析。为了探究在婴鲤属中，上述比对到的发育相关基因与婴鲤属基因组上其他基因是否存在基因结构上的显著差异，依据参考基因组的 GO 注释信息，对上述比对到婴鲤属中的发育相关基因与婴鲤参考基因组上所有注释基因的可变剪切、多聚腺苷酸化、外显子数量和融合基因进行显著性检验（*t*-test）。本研究使用 Python 的 matplotlib 包绘制相关图片（Hunter 2007），使用 pandas 和 scipy 包开展数据统计分析（McKinney 2010，Virtanen et al. 2020）。

2　结果

2.1　婴鲤属两个物种的全长转录组数据统计

基于三代 PacBio 测序技术，构建了婴鲤属两个物种 *P. carbunculus* 和 *P. micromegethes* 整鱼的全长转录组文库，分别获得 20.29 Gb 和 15.08 Gb 的 subreads 数据，过滤后分别得到 502 996 条和 390 292 条 CCS 序列，平均读长为 2 629 bp 和 2 675 bp。随后，选择包含 3′和 5′引物的完整序列，获得 371 939 条和 236 787 条全长非嵌合序列，N50 为 2 874 bp 和 4 185 bp（表 1）。此外，采用 GMAP 将校正后共识序列分别比对到 *P. carbunculus* 和 *P. micromegethes* 参考基因组草图（Malmstrøm et al. 2018），分别获得 35 883 和 20 381 个全长转录组本序列，比对率为 92.84%和 87.98%（表 2）。使用 NR、NT、Pfam、KOG/COG、SwissProt、KEGG 和 GO 数据库对未映射到参考基因组的序列进行注释，分别获得 2 595 个和 2 464 个转录本信息（表 3）。GO 数据库在 *P. carbunculus* 和 *P. micromegethes* 中分别注释到 1 666 个和 1 526 个转录本，并将它们按照生物过程（biological process）、细胞组分（cellular component）和分子功能（molecular function）进行功能富集。*P. carbunculus* 和 *P. micromegethes* 的转录本分析中，生物过程类别主要富集在细胞过程（分别为 53.30%和 48.69%）、代谢过程（分别为 43.04%和 39.78%）和单一生物体过程（分别为 32.41%和 30.14%）。细胞组分类别富集于细胞（分别为 28.93%和 30.28%），其次是细胞器（分别为 19.51%和 24.90%）。在分子功能类别中，结合亚类是最常见的注释类型（分别为 65.67%和 69.26%），其次是催化活性（分别为 41.36%和 41.02%）。此外根据 GMAP 比对结果，在 *P. carbunculus* 和 *P. micromegethes* 中分别发现 16 009 和 8 608 个新转录本，以及 2 339 和 2 224 个新基因的转录本（表 4，5）。

表 1　婴鲤属两个物种全长转录组测序数据质量

Table 1　Quality of full-length transcriptome sequencing data of two *Paedocypris* species

指标 Index	*Paedocypris carbunculus*	*P. micromegethes*
子读段数量 Number of subreads	10 900 499	9 962 155
子读段总长度 Total length of subreads (Gb)	20.29	15.08
环形一致序列数量 Number of circular consensus sequence	502 996	390 292
环形一致序列平均长度 Mean length of circular consensus sequence (bp)	2 629	2 675
全长非嵌合序列数量 Number of full-length non-chimeric reads	371 939	236 787
全长非嵌合序列 N50 N50 of full-length non-chimeric reads (bp)	2 874	4 185
校正后共识序列数量 Number of polished consensus sequence	38 651	23 165
校正后共识序列 N50 N50 of polished consensus sequence (bp)	2 863	4 290

表 2 全长转录组校正后共识序列比对结果

Table 2 Full-length transcriptome polished consensus sequence GMAP alignment results

指标 Index	*Paedocypris carbunculus*	*P. micromegethes*
比对到的读段数量（占所有序列百分比） Number of aligned reads (percentage of all consensus sequences)	35 883 (92.84%)	20 381 (87.98%)
未比对到的读段数量（占所有序列百分比） Number of unaligned reads (percentage of all sequences)	2 768 (7.16%)	2 784 (12.02%)
比对到多处的读段数量（占所有序列百分比） Number of reads aligned to multiple locations (percentage of all sequences)	604 (1.56%)	417 (1.80%)
比对到一处的读段数量（占所有序列百分比） Number of reads aligned to a single location (percentage of all sequences)	35 279 (91.28%)	19 964 (86.18%)

表 3 未比对上参考基因组的全长转录组序列注释信息

Table 3 Annotation information for full-length transcriptome sequences that unaligned to the reference

指标 Index	*Paedocypris carbunculus*	*P. micromegethes*
未比对到的读段数量 Number of unaligned reads	2 768	2 784
NR 注释到的读段数量（注释到的读段数量占比） Number of reads annotated by NR (percentage of annotated reads)	2 356 (85.12%)	2 219 (79.71%)
SwissProt 注释到的读段数量（注释到的读段数量占比） Number of reads annotated by SwissProt (percentage of annotated reads)	2 226 (80.42%)	2 141 (76.90%)
KEGG 注释到的读段数量（注释到的读段数量占比） Number of reads annotated by KEGG (percentage of annotated reads)	2 312 (83.53%)	2 164 (77.73%)
KOG 注释到的读段数量（注释到的读段数量占比） Number of reads annotated by KOG (percentage of annotated reads)	1 843 (66.58%)	1 742 (62.57%)
GO 注释到的读段数量（注释到的读段数量占比） Number of reads annotated by GO (percentage of annotated reads)	1 666 (60.19%)	1 526 (54.81%)
NT 注释到的读段数量（注释到的读段数量占比） Number of reads annotated by NT (percentage of annotated reads)	2 334 (84.32%)	2 224 (79.89%)
PFAM 注释到的读段数量（注释到的读段数量占比） Number of reads annotated by FPAM (percentage of annotated reads)	1 666 (60.19%)	1 526 (54.81%)
在所有数据库中都注释到的读段数量（注释到的读段数量占比） Number of reads annotated by all databases (percentage of annotated reads)	1 396 (50.43%)	1 235 (44.36%)
至少在一个数据库中注释到读段数量（注释到的读段数量占比） Number of reads annotated by at least one database (percentage of annotated reads)	2 595 (93.75%)	2 464 (88.51%)

表 4 全长转录本分类特征

Table 4 Classification characteristics of full-length transcript

指标 Index	*Paedocypris carbunculus*	*P. micromegethes*
转录本数 Number of transcripts	19 352	11 139
已知的转录本数 Number of known transcripts	914	307
已知基因的新转录本数 Number of novel transcripts from known genes	16 099	8 608
新基因的转录本数 Number of transcripts from novel genes	2 339	2 224
转录本平均长度 Average length of transcripts	2 134	1 886
转录本的 N50 N50 of transcripts	2 744	3 606

表 5　新基因数据库注释

Table 5　Novel gene database annotation

指标 Index	*Paedocypris carbunculus*	*P. micromegethes*
新转录本数量 Number of novel transcripts	16 099	8 608
新基因数量 Number of novel gene	1 823	1 883
NR 注释到的读段数量（注释到的读段数量占比） Number of reads annotated by NR (percentage of annotated reads)	810 (44.43%)	480 (25.49%)
SwissProt 注释到的读段数量（注释到的读段数量占比） Number of reads annotated by SwissProt (percentage of annotated reads)	642 (35.22%)	385 (20.45%)
KEGG 注释到的读段数量（注释到的读段数量占比） Number of reads annotated by KEGG (percentage of annotated reads)	755 (41.42%)	445 (23.63%)
KOG 注释到的读段数量（注释到的读段数量占比） Number of reads annotated by KOG (percentage of annotated reads)	460 (25.23%)	273 (14.50%)
GO 注释到的读段数量（注释到的读段数量占比） Number of reads annotated by GO (percentage of annotated reads)	535 (29.35%)	311 (16.52%)
NT 注释到的读段数量（注释到的读段数量占比） Number of reads annotated by NT (percentage of annotated reads)	927 (50.85%)	691 (36.70%)
PFAM 注释到的读段数量（注释到的读段数量占比） Number of reads annotated by FPAM (percentage of annotated reads)	535 (29.35%)	311 (16.52%)
在所有数据库中都注释到的读段数量（注释到的读段数量占比） Number of reads annotated by all databases (percentage of annotated reads)	243 (13.33%)	137 (7.28%)
至少在一个数据库中注释到读段数量（注释到的读段数量占比） Number of reads annotated by at least one database (percentage of annotated reads)	1 214 (66.59%)	862 (45.78%)

2.2　*P. carbunculus* 和 *P. micromegethes* 全长转录组基因结构分析

基于 hmmsearch 从头预测，在 *P. carbunculus* 和 *P. micromegethes* 中分别鉴定到 1 040 个和 385 个转录因子。这些转录因子被归类到不同的家族中，其中最丰富的类型包括 zf-C2H2（分别为 314 和 112 个转录因子）、ZBTB（分别为 115 和 45 个转录因子）、THR-like（分别为 65 和 15 个转录因子）、Homeobox（分别为 58 和 20 个转录因子）和 TF_bZIP（分别为 49 和 27 个转录因子）（图 1）。在 *P. carbunculus* 和 *P. micromegethes* 中分别鉴定出 9 404 和 6 068 个可变剪切事件。其中，5′端可变剪接比例最高，在 *P. carbunculus* 和 *P. micromegethes* 中分别包含 2 375 个（25.26%）和 1 267 个（20.88%），终止外显子可变剪接的比例最低，分别只有 126 个（1.34%）和 61 个（1.01%）（表 6）。此外，在 7 581 个 *P. carbunculus* 与 4 185 个 *P. micromegethes* 基因上分别鉴定出了 10 278 和 5 043 个多聚腺苷酸化位点。

在 *P. carbunculus* 和 *P. micromegethes* 中分别鉴定到 7 345 和 5 995 个 LncRNAs。根据这些 LncRNAs 的基因组位置，可分为四个类别：基因间区长非编码 RNA（LincRNA）、反义链（Antisense）LncRNA、基因内含子区（Sense intronic）LncRNA 和基因内与外显子有重叠（Sense overlapping）的 LncRNA（图 2）。其中 LincRNA 是两个物种中最丰富的 LncRNA 类型，在 *P. carbunculus* 和 *P. micromegethes* 中分别占比 63.31%和 58.1%。其次是 Antisense LncRNA 和 Sense overlapping LncRNA 在 *P. carbunculus* 中分别占比 13.68%和 16.71%，在 *P. micromegethes* 中分别占比 15.4%和 14.68%。相比之下，Sense intronic LncRNA 是最稀缺的类型，在 *P. carbunculus* 和 *P. micromegethes* 中分别占比 6.3%和 11.83%。这些研究结果展示了在婴鲤属基因组中不同类型 LncRNA 的分布和丰度，为深入探讨 LncRNA 在婴鲤属中的功能角色及其进化学意义提供了数据基础。

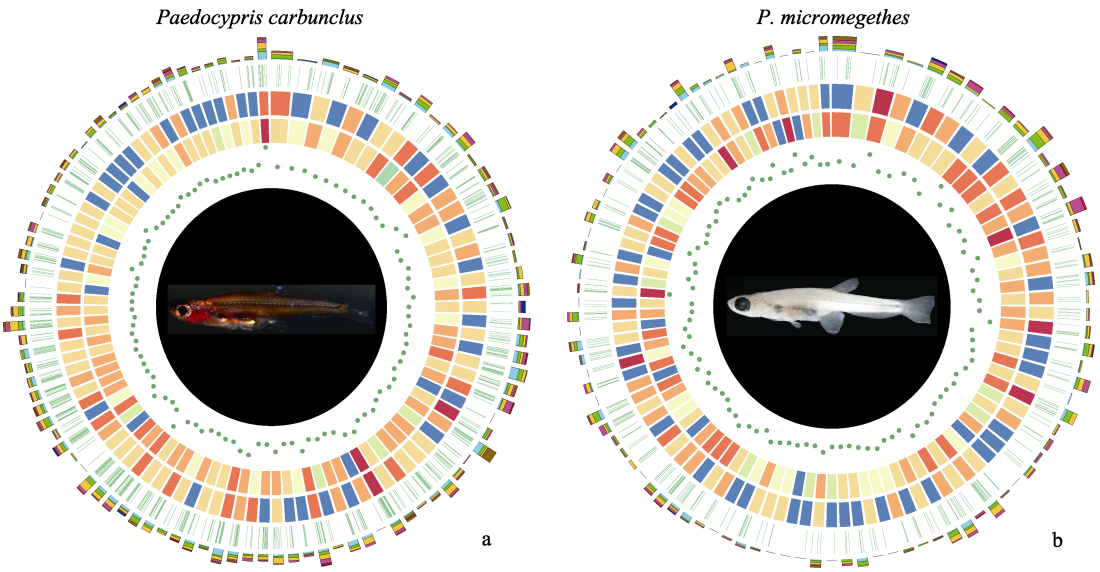

图 1　婴鲤全长转录组基因结构分析

Fig. 1　Gene structure analysis of full-length transcriptome of *Paedocypris*

a 和 b 分别是 *P. carbunculus* 和 *P. micromegethes* 的基因结构分析结果，由外及内依次为：1. 可变剪接位点（堆积柱状图，不同可变剪接类型用不同颜色表示，浅蓝色为内含子滞留，绿色为 3′端可变剪接，黄色为 5′端可变剪接，紫色为外显子跳跃，红色为外显子互斥，棕色为起始外显子，深蓝色为终止外显子）；2. 多聚腺苷酸化位点；3. 新转录本分布图，颜色越接近于红色表示密度越大；4. 新基因分布图，颜色越接近于红色表示密度越大；5. 长非编码 RNA 密度分布；6. 两种婴鲤图示。

a and b show the gene structure analysis results for *P. carbunculus* and *P. micromegethes*, respectively. The layers of the figure from outside inward are as follows: 1. Alternative splicing sites (stacked bar chart, different types of alternative splicing are indicated by different colors; light blue for intron retention, green for 3′ alternative splicing, yellow for 5′ alternative splicing, purple for exon skipping, red for mutually exclusive exons, brown for alternative first exons, dark blue for alternative last exons); 2. Alternative polyadenylation sites; 3. Distribution map of novel transcripts, with colors closer to red indicating higher density; 4. Distribution map of novel genes, with colors closer to red indicating higher density; 5. Density distribution of long non-coding RNAs; 6. Photos of two *Paedocypris* species.

表 6　可变剪切位点类型

Table 6　Classification of alternative splicing

指标 Index	*Paedocypris carbunculus*	*P. micromegethes*
可变剪切基因总数 Total number of genes with alternative splicing	9 404	6 068
外显子跳跃可变剪切基因数量（占所有可变剪切基因百分比） Number of genes with skipped exon alternative splicing (percentage of all alternative splicing genes)	1 428 (15.19%)	719 (11.85%)
外显子互斥可变剪切基因数量（占所有可变剪切基因百分比） Number of genes with mutually exclusive exon alternative splicing (percentage of all alternative splicing genes)	133 (1.41%)	70 (1.15%)
内含子滞留可变剪切基因数量（占所有可变剪切基因百分比） Number of genes with retained intron alternative splicing (percentage of all alternative splicing genes)	1 265 (13.45%)	752 (12.39%)
5'端可变剪接基因数量（占所有可变剪切基因百分比） Number of genes with alternative 5' splice site alternative splicing (percentage of all alternative splicing genes)	2 375 (25.26%)	1 267 (20.88%)
3'端可变剪接基因数量（占所有可变剪切基因百分比） Number of genes with alternative 3' splice site alternative splicing (percentage of all alternative splicing genes)	1 990 (21.16%)	1 079 (17.78%)
起始外显子可变剪接基因数量（占所有可变剪切基因百分比） Number of genes with alternative first exon alternative splicing (percentage of all alternative splicing genes)	495 (5.26%)	205 (3.38%)
终止外显子可变剪接基因数量（占所有可变剪切基因百分比） Number of genes with alternative last exon alternative splicing (percentage of all alternative splicing genes)	126 (1.34%)	61 (1.01%)

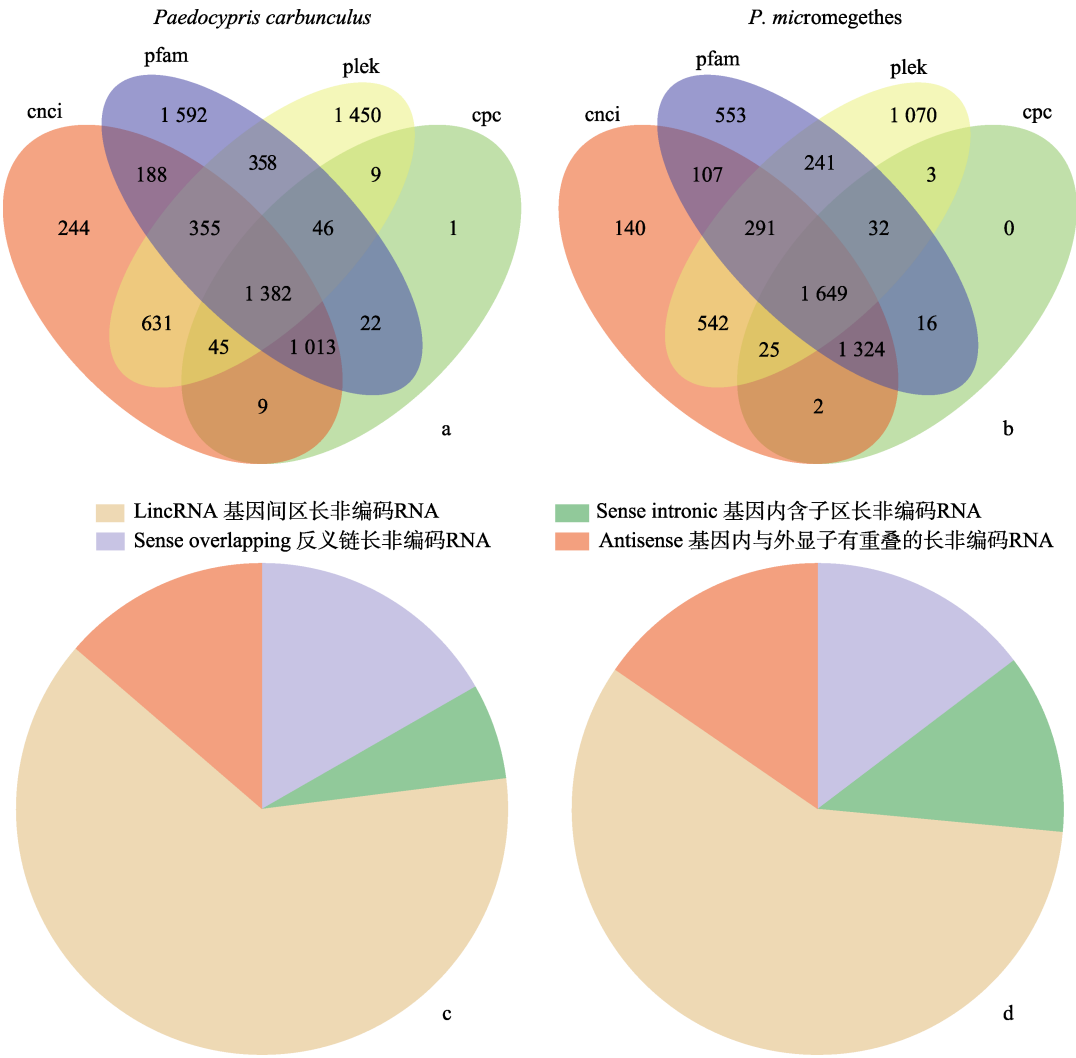

图 2　长非编码 RNA 预测结果

Fig. 2　LncRNA prediction results

a 和 b 分别是 *Paedocypris carbunculus* 和 *P. micromegethes* 的全长转录组由 4 种不同软件预测的长非编码 RNA 结果绘制的韦恩图；c 和 d 分别是 *P. carbunculus* 和 *P. micromegethes* 的长非编码 RNA 分类。Antisense. 反义链长非编码 RNA；LincRNA. 基因间区长非编码 RNA；Sense intronic. 基因内含子区长非编码 RNA；Sense overlapping. 基因内与外显子有重叠的长非编码 RNA

a and b are the Venn diagrams of the full-length transcriptomes of *P. carbunculus* and *P. micromegethes*, respectively, drawn from the LncRNA results predicted by four different software; c and d are the classification of LncRNA of *P. carbunculus* and *P. micromegethes* respectively. Antisense. Antisense strand; LincRNA. Intergenic region; Sense intronic. Gene intron region; Sense overlapping. Gene overlaps with exons

2.3　*Hox* 家族基因和发育相关基因的鉴定

已知在婴鲤属的基因组中存在 *Hox* 家族基因的丢失，但对婴鲤基因组中 *Hox* 家族基因的表达情况却缺乏了解。为鉴定婴鲤属中 *Hox* 家族基因的转录表达情况，下载了两个与婴鲤同为鲤科的物种（鲤和鲢）和两个近亲鲑科的物种（银大麻哈鱼和大西洋鲑）的全长转录组数据，将这 4 个物种和婴鲤属的全长转录组数据通过 Blastn 与斑马鱼的 48 个 *Hox* 家族基因 cds 序列进行比对（图 3）。通过阈值（e-value）小

于 10^{-20} 进行筛选，分别在婴鲤、鲤、鲢、大西洋鲑和银大麻哈鱼中检测到 8、25、20、17 和 21 个 *Hox* 家族基因（图 3a）。假设斑马鱼的 48 个 *Hox* 家族基因全部存在于其他四种鱼类的基因组中，由此可推断婴鲤、鲤、鲢、大西洋鲑和银大麻哈鱼的 *Hox* 家族基因发生转录表达的比例分别为 21%、52%、42%、35%和 44%。这表明 *Hox* 家族基因不仅在婴鲤属基因组水平发生丢失，而且在转录表达水平也显著低于其他鱼类。例如，发现有 7 个 *Hox* 家族基因（*hoxb2a*、*hoxb4a*、*hoxb5a*、*hoxc5a*、*hoxc10a*、*hoxc13a*、*hoxd10a*）在婴鲤属全长转录组数据中未检测到表达（图 3c），但在上述 4 种鱼中最少有 3 种鱼表达这些基因，

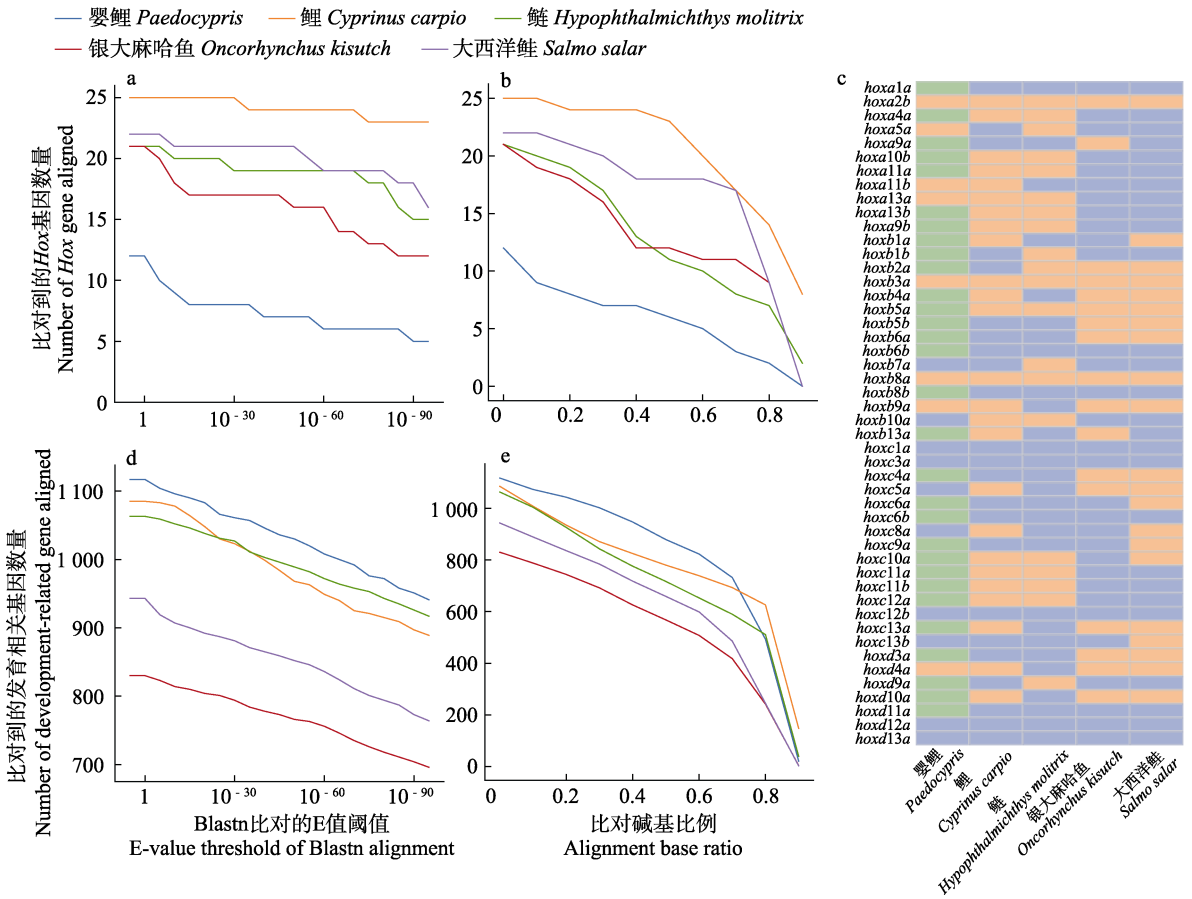

图 3　*Hox* 家族基因和发育相关基因比对结果

Fig. 3　Comparative results for *Hox* family genes and developmental related genes

a. 不同 e-value 阈值下比对到的 *Hox* 家族基因数量；b. *Hox* 家族基因比对覆盖度与比对数量的关系；c. 在 e-value 设置为 10^{-20} 时，婴鲤和四个物种全长转录组检索到的 *Hox* 家族基因（橙色表示未比对到相应基因；蓝色表示比对到相应基因；绿色表示在婴鲤的参考基因组中存在，但在婴鲤的全长转录组中没有比对到的 *Hox* 基因）；d. 不同 e-value 阈值下比对到的发育相关基因数量；e. 发育相关基因比对覆盖度与比对数量的关系。

a. Number of *Hox* family genes aligned at various e-value thresholds; b. The relationship between alignment coverage and the number of aligned *Hox* family genes; c. When the e-value is set to 10^{-20}, five full-length transcriptomes searched for *Hox* family genes (orange means no alignment, blue means aligned, green means the gene exists in the reference genome, but no aligned in the full-length transcriptome; d. Number of developmental related genes aligned at various e-value thresholds; e. The relationship between alignment coverage and the number of aligned developmental related genes.

其中 *hoxb5a* 在上述 4 种鱼的全长转录组数据中均检测到表达，仅在婴鲤中未检测到表达。*hoxc5a* 被发现在婴鲤属的参考基因组发生了丢失，在婴鲤属两个物种的全长转录组数据中同样未检测到该基因的表达。

发育相关基因在动物生长和体型塑造中起着非常重要的作用。为验证婴鲤属中发育相关基因的转录表达情况，将婴鲤和上述 4 个物种的全长转录组数据与斑马鱼的 1 581 个发育相关基因 cds 序列进行 Blastn 比对（图 3c，d），相较于其他 4 个物种，婴鲤属两个物种在转录层面上表达更多的发育相关基因。例如，在阈值设为 10^{-20} 时，通过比对分别在银大麻哈鱼、大西洋鲑、鲤和鲢中鉴别到约 800、900、1 050 和 1 060 个发育相关基因存在表达，但在婴鲤属中鉴别到约 1 100 个发育相关基因存在表达（图 3c）。

2.4　鉴定婴鲤的体型小型化与基因结构的关系

为探究婴鲤属中位于关键发育 GO 功能通路（如前后轴形成、背腹轴形成、骨骼发育等）中的发育相关基因是否在基因结构上与其他基因具有显著差异，下载斑马鱼中 1 581 个上述 GO 功能通路的基因 cds 序列与婴鲤两个物种的全长转录组数据进行比对，依据婴鲤属参考基因组的注释信息，经 *t* 检验，发现婴鲤属中这些与发育相关基因在可变剪切和多聚腺苷酸化上与基因组其他基因无显著差异，在 *P. carbunculus* 中，与发育相关的每个基因平均有 2.0 个剪切位点，其他基因平均有 2.4 个剪切位点（$t = -1.98$，$df = 37$，$P = 0.056$，图 4a），*P. micromegethes* 中可变剪切位点在发育相关基因和其他基因中分别为 2.5 与 2.6 个剪切位点（$t = -0.11$，$df = 30$，$P = 0.914$，图 4a）。在 *P. carbunculus* 中发育相关基因与其他基因均包含 1.4 个多聚腺苷酸化位点（$t = 0.05$，$df = 76$，$P = 0.958$，图 4b），而在 *P. micromegethes* 中分别为 1.1 和 1.2（$t = -1.19$，$df = 53$，$P = 0.239$，图 4b）。在 *P. carbunculus* 中观察到，其他基因的平均外显子数量（18.6）显著高于发育相关基因（13.2，$t = -2.83$，$df = 85$，$P = 0.006$，图 4c），然而在 *P. micromegethes* 中发育相关基因的平均外显子数量（14.2）低于其他基因（17.2，$t = 1.06$，$df = 61$，$P = 0.292$，图 4c）。针对基因融合事件，在 *P. carbunculus* 中，发育相关基因和其他基因分别检测到 1.5 次和 2.4 次融合事件（$t = -2.83$，$df = 11$，$P = 0.017$，图 4d），在 *P. micromegethes* 中，分别检测到 1.4 次和 2.2 次（$t = -2.78$，$df = 6$，$P = 0.032$，图 4d），这表明在婴鲤属物种中，发育相关基因的融合事件均显著低于基因组上的其他基因。

3　讨论

本研究通过对 *P. carbunculus* 和 *P. micromegethes* 开展全长转录组测序，构建了婴鲤属这两个物种的全长转录组文库，通过评估全长转录组测序质量、比对参考基因组信息、对未比对上的全长转录组进行注释，分别在 *P. carbunculus* 和 *P. micromegethes* 中获得 19 352 和 11 139 个有效全长转录本，随后对这些转录本开展了特征分类和新转录本功能注释。基于参考基因组及注释信息对获得的全长转录组数据进一步开展基因结构分析、数据库注释、转录因子分析、可变剪接分析、长非编码 RNA 分析和融合基因分析，发现婴鲤属这两个物种在转录表达层面上存在大量的新基因、新转录本和新剪接位点，为探索婴鲤属的体型小型化的分子机制提供了新的数据支撑。

Malmstrøm 等（2018）研究发现，婴鲤属在基因组水平丢失多个驱动肢体发育和其他关键发育通路的 *Hox* 家族基因。*Hox* 家族基因主要调控细胞分裂、纺锤体方向以及硬毛、附肢等部位的发育，是生物体中一类专门调控生物形体的基因，一旦这些基因发生突变，会导致生物体发育畸形或者部分组织或器官缺失（Krumlauf 1994，Sordino et al. 1995，Santini et al. 2005，Hejnol et al. 2017）。本研究发现，婴鲤属在基因组上保留的 *Hox* 家族基因存在大量的未表达现象，结合此前研究发现婴鲤属在基

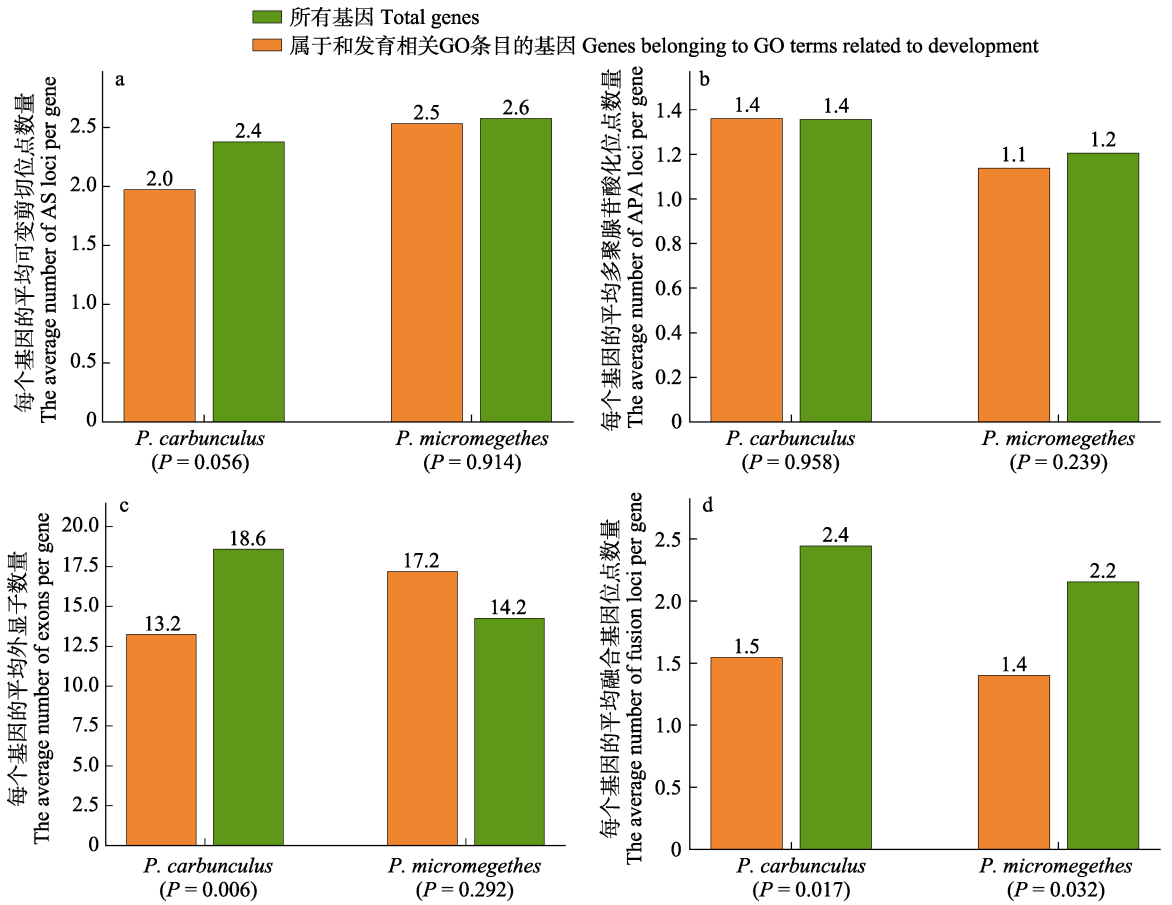

图 4　婴鲤转录组基因结构差异结果

Fig. 4　Transcriptome gene structure differences of *Paedocypris*

a ~ d. 被注释到与发育相关点 7 个 GO 的基因和所有基因拥有的每个基因的平均可变剪切位点数量（a）；每个基因的平均多聚腺苷酸化位点数量（b）；每个基因的平均外显子数量（c）；每个基因的平均融合基因位点数量（d）。

a - d. The average number of alternative splicing sites (AS) per gene (a); the average number of the alternative polyadenylation (APA) site per gene (b); the average number of exons per gene (c); the average number of fusion gene sites per gene (d) possessed by the genes annotated to the 7 GOs related to development and all genes.

因组层面上丢失多个 *Hox* 家族基因，表明该属在基因组水平和转录表达水平上均存在 *Hox* 家族基因功能丢失，这进一步提示 *Hox* 家族基因在转录表达层面上的功能丢失可能是导致婴鲤体型小型化的一个重要原因。

此前研究表明，*P. carbunculus* 和 *P. micromegethes* 组装的基因组大小分别为 430.79 Mb 和 414.70 Mb，注释到 25 567 和 25 453 个结构基因（Malmstrøm et al. 2018）。暹罗斗鱼（*Betta splendens*）基因组大小与婴鲤属相当，约为 452 Mb，包含 25 104 个结构基因（Zhang et al. 2022），表明婴鲤属中结构基因的总数与其他物种相比并没有发生大量丢失，这提示婴鲤属的体型小型化可能还与基因结构有关。本研究采用基于 PacBio 的全长转录组测序技术揭示了婴鲤属两个物种全组织的转录本结构特征，包括基因的外显子数量、可变剪切位点、可变多聚腺苷酸化以及基因融合。对注释到 7 个重要发育相关 GO 功能通路的转录本与所有转录本进行了比较，以评估发育相

关基因与其他所有基因在结构方面的差异。结果显示，婴鲤属中与发育相关的基因发生融合次数显著低于其他所有基因，这表明婴鲤属中发育相关基因更少发生基因融合事件。基因融合在基因结构进化中发挥着关键作用，它既可以促成新基因的形成，从而产生新功能，也可能导致原有基因功能丧失。推测这可能是由于婴鲤属的发育相关基因需要快速发育以实现生殖成熟并维持种群连续性，维持更为保守的功能，从而保证个体的正常发育。

本研究通过对婴鲤全长转录组开展深入分析，为进一步理解婴鲤属的体型发育机制提供了重要线索，并突显了 *Hox* 家族基因以及基因融合事件在其适应性演化中的重要性。

参 考 文 献

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

Agerholm J S, Menzi F, McEvoy F J, et al. 2016. Lethal chondrodysplasia in a family of Holstein cattle is associated with a *de novo* splice site variant of *COL2A1*. BMC Veterinary Research, 12: 100.

Alamancos G P, Pagès A, Trincado J L, et al. 2015. Leveraging transcript quantification for fast computation of alternative splicing profiles. RNA, 21(9): 1521–1531.

Ashburner M, Ball C A, Blake J A, et al. 2000. Gene Ontology: tool for the unification of biology. Nature Genetics, 25(1): 25–29.

Bairoch A, Boeckmann B. 1991. The SWISS-PROT protein sequence data bank. Nucleic Acids Research, 19(Suppl): 2247–2249.

Bateman A, Coin L, Durbin R, et al. 2004. The Pfam protein families database. Nucleic Acids Research, 32(Database issue): D138–D141.

Boegheim I J M, Leegwater P A J, van Lith H A, et al. 2017. Current insights into the molecular genetic basis of dwarfism in livestock. Veterinary Journal, 224: 64–75.

Bouwman A C, Daetwyler H D, Chamberlain A J, et al. 2018. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nature Genetics, 50(3): 362–367.

Britz R. 2008. *Paedocypris carbunculus*, a new species of miniature fish from Borneo (Teleostei: Cypriniformes: Cyprinidae). Raffles Bulletin of Zoology, 56(2): 415–422.

Britz R, Conway K W. 2009. Osteology of *Paedocypris*, a miniature and highly developmentally truncated fish (Teleostei: Ostariophysi: Cyprinidae). Journal of Morphology, 270(4): 389–412.

Britz R, Conway K W, Rüber L. 2014. Miniatures, morphology and molecules: *Paedocypris* and its phylogenetic position (Teleostei, Cypriniformes). Zoological Journal of the Linnean Society, 172(3): 556–615.

Cooper N, Purvis A. 2010. Body size evolution in mammals: complexity in tempo and mode. The American Naturalist, 175(6): 727–738.

Galdzicka M, Patnala S, Hirshman M G, et al. 2002. A new gene, *EVC2*, is mutated in Ellis–van Creveld syndrome. Molecular Genetics and Metabolism, 77(4): 291–295.

Glaw F, Köhler J, Hawlitschek O, et al. 2021. Extreme miniaturization of a new amniote vertebrate and insights into the evolution of genital size in chameleons. Scientific Reports, 11(1): 2522.

Hanken J, Wake D B. 1993. Miniaturization of body size: organismal consequences and evolutionary significance. Annual Review of Ecology and Systematics, 24: 501–519.

Harada D, Yamanaka Y, Ueda K, et al. 2007. Sustained phosphorylation of mutated FGFR3 is a crucial feature of genetic dwarfism and induces apoptosis in the ATDC5 chondrogenic cell line via PLCγ-activated STAT1. Bone, 41(2): 273–281.

Hejnol A, Vellutini B C. 2017. Larval evolution: I'll tail you later…. Current Biology, 27(1): R21–R24.

Hunter J D. 2007. Matplotlib: a 2D graphics environment. Computing in Science & Engineering, 9(3): 90–95.

Kanehisa M. 2002. The KEGG Database. Novartis Foundation Symposium, 247: 91–103.

Kong L, Zhang Y, Ye Z Q, et al. 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Research, 35(Web Server issue): W345–W349.

Kottelat M, Britz R, Hui T H, et al. 2006. *Paedocypris*, a new genus of Southeast Asian cyprinid fish with a remarkable sexual

dimorphism, comprises the world's smallest vertebrate. Proceedings of the Royal Society B: Biological Sciences, 273(1589): 895–899.

Krumlauf R. 1994. *Hox* genes in vertebrate development. Cell, 78(2): 191–201.

Kyöstilä K, Niskanen J E, Arumilli M, et al. 2021. Intronic variant in *POU1F1* associated with canine pituitary dwarfism. Human Genetics, 140(11): 1553–1562.

Li A, Zhang J, Zhou Z. 2014. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. BMC Bioinformatics, 15(1): 311.

Liu S, Hui T H, Tan S L, et al. 2012. Chromosome evolution and genome miniaturization in minifish. PLoS One, 7(5): e37305.

Malmstrøm M, Britz R, Matschiner M, et al. 2018. The most developmentally truncated fishes show extensive *Hox* gene loss and miniaturized genomes. Genome Biology and Evolution, 10(4): 1088–1103.

Mayden R L, Chen W J. 2010. The world's smallest vertebrate species of the genus *Paedocypris*: *a* new family of freshwater fishes and the sister group to the world's most diverse clade of freshwater fishes (Teleostei: Cypriniformes). Molecular Phylogenetics and Evolution, 57(1): 152–175.

McGinnis S, Madden T L. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Research, 32(Web Server issue): W20–W25.

McKinney W. 2010. Data structures for statistical computing in Python // Austin, Texas: Proceedings of the 9th Python in Science Conference, 56–61.

Murgiano L, Jagannathan V, Benazzi C, et al. 2014. Deletion in the *EVC2* gene causes chondrodysplastic dwarfism in Tyrolean Grey cattle. PLoS One, 9(4): e94861.

Parker H G, VonHoldt B M, Quignon P, et al. 2009. An expressed *fgf4* retrogene is associated with breed-defining chondrodysplasia in domestic dogs. Science, 325(5943): 995–998.

Polilov A A. 2017. First record of *Megaphragma* (Hymenoptera, Trichogrammatidae) in Columbia, and third animal species known to have anucleate neurons. Journal of Hymenoptera Research, 60: 181–185.

Pruitt K D, Tatusova T, Maglott D R. 2005. NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research, 33(Database issue): D501–D504.

Rittmeyer E N, Allison A, Gründler M C, et al. 2012. Ecological guild evolution and the discovery of the world's smallest vertebrate. PLoS One, 7(1): e29797.

Santini S, Bernardi G. 2005. Organization and base composition of tilapia *Hox* genes: implications for the evolution of *Hox* clusters in fish. Gene, 346: 51–61.

Sordino P, van der Hoeven F, Duboule D. 1995. *Hox* gene expression in teleost fins and the origin of vertebrate digits. Nature, 375(6533): 678–681.

Sun L, Luo H, Bu D, et al. 2013. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Research, 41(17): e166.

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

Virtanen P, Gommers R, Oliphant T E, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods, 17(3): 261–272.

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

Zhang H M, Liu T, Liu C J, et al. 2015. AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. Nucleic Acids Research, 43(Database issue): D76–D81.

Zhang W, Wang H, Brandt D Y C, et al. 2022. The genetic architecture of phenotypic diversity in the Betta fish (*Betta splendens*). Science Advances, 8(38): eabm4955.