中国水产科学 2025 年 6 月, 32(6): 753–765
Journal of Fishery Sciences of China

研究论文

DOI: 10.12264/JFSC2025-0004

淡水定居型刀鲚无间隙基因组组装

马凤娇[1], 王慧[1], 任泷[2], 刘凯[1, 2*]

1. 南京农业大学无锡渔业学院, 江苏 无锡 214081;
2. 中国水产科学研究院淡水渔业研究中心, 农业农村部淡水渔业和种质资源利用重点实验室, 江苏 无锡 214081

摘要: 淡水定居型刀鲚(*Coilia nasus taihuensis*)又称为湖鲚, 终生生活在淡水湖泊中, 不需要进行江海洄游也能完成整个生活史过程。栖息生境的差异及长期的地理分割, 使得湖鲚与洄游型刀鲚(*Coilia nasus*)选择了不同的环境适应性进化机制。由于目前缺乏湖鲚的基因组信息, 与其环境适应相关的遗传机制缺乏系统研究。本研究以太湖的定居型个体为实验对象, 通过 Pacific Biosciences (PacBio) high-fidelity (HiFi)测序数据组装得到高质量基因组骨架, 利用 Hi-C 测序数据实现基因组染色体水平的组装, 结合 Nanopore 测序数据进行基因组补洞, 最终获得完整的、无间隙的湖鲚参考基因组。湖鲚基因组大小约为 834.09 Mb, contig N50 高达 35.45 Mb, 挂载到 24 条染色体上, 挂载率为 99.83%。基因组组装质量评估显示, BUSCO (Benchmarking Universal Single-Copy Orthologs)评估值为 91.90%, 表明基因组组装完整性较高。基因组注释特征显示, 基因组重复序列总长度为 382.39 Mb, 占基因组的 45.85%, 基因结构注释共鉴定 21730 个蛋白编码基因, 其中 21666 个基因(99.71%)被功能注释。基因组共线性结果显示, 刀鲚和湖鲚之间具有极高的基因组共线性比率(96.95%), 共鉴定到 48852 个共线性区块, 表明两者之间遗传关系密切。本研究为后续刀鲚适应性机制研究提供素材, 为深入开展湖鲚群体遗传学研究提供重要的基因组资源。

关键词: 淡水定居型; 刀鲚; 无间隙基因组; 基因注释; 共线性分析
中图分类号: S917 **文献标志码:** A **文章编号:** 1005–8737–(2025)06–0753–13

刀鲚(*Coilia nasus*)是一种重要的溯河洄游性经济鱼类, 主要分布在我国黄海、渤海、东海沿岸及长江、黄河、钱塘江等通海江河[1]。每年春季, 溯河洄游型刀鲚进入河口, 沿长江上溯至长江干流或鄱阳湖、洞庭湖等通江湖泊中进行产卵繁殖[2]。为了适应不断变化的环境, 刀鲚已形成两种不同的生活史形式, 一种是江海洄游型, 另一种就是淡水定居型, 如终生生活于太湖中的湖鲚(*Coilia nasus taihuensis*)。湖鲚不进行生殖洄游, 能够在太湖、洪泽湖、巢湖等陆封型淡水湖泊中完成整个生活史过程。由于长期的地理隔离及栖息生境差异, 与洄游型刀鲚相比, 湖鲚在形态特征、渗透压调节、摄食行为等方面发生了显著的变化[3], 袁传宓等将淡水定居型群体定为刀鲚的新亚种, 即湖鲚[4]。

目前, 关于湖鲚和刀鲚种群形态特征的研究, 主要是通过形态度量学、聚类分析等方法探讨刀鲚和湖鲚之间的系统发育关系, 结果显示两者之间的亲缘关系较近, 形态差异未达到亚种水平, 湖鲚是刀鲚为适应环境而产生形态差异的一种生态型[5-6]。还有研究基于形态学方法分析不同湖泊湖鲚种群之间形态变异程度, 发现太湖种群较其他湖泊种群出现了一定程度的形态分化[7]。对于湖鲚和刀鲚分子遗传学的研究, 主要是基于线粒体基因、核基因及少量微卫星位点等分子标记, 分析刀鲚和湖鲚种群遗传变异, 结果表明湖鲚和

收稿日期: 2025-01-18; **修订日期:** 2025-04-17.
基金项目: 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资金项目(2023TD11, 2023TD65); 江苏省农业农村厅渔业生态与资源监测专项; 江苏省研究生科研与实践创新计划项目(KYCX23_0757).
作者简介: 马凤娇, 女, 博士研究生, 从事鱼类基因组学研究. E-mail: mafengjiao2019@163.com
通信作者: 刘凯, 研究员, 从事鱼类生态学与物种保护研究. E-mail: liuk@ffrc.cn

刀鲚种群之间遗传分化不显著，尚未达到种或亚种的分化，湖鲚是刀鲚的淡水生态型，并非有效物种[8-13]。最近的研究利用简化基因组测序(RAD-seq)技术分析刀鲚和湖鲚之间的遗传关系，显示两者之间的遗传分化较低，证明湖鲚是刀鲚的一种生态型[14]。另外，还有一些关于湖鲚群体遗传多样性的研究，通过线粒体控制区(D-loop)分析不同湖泊湖鲚群体的遗传多样性水平和遗传结构特征，发现不同湖泊湖鲚种群间没有明显的遗传分化[15-16]。综上，现有研究仅局限于形态特征、线粒体基因以及核基因等分子标记，解析力度较低，对湖鲚相关遗传信息挖掘还不够深入。目前，湖鲚的参考基因组尚未有报道，这制约了对其适应性进化遗传机制的解析。

随着技术不断发展，以 Pacific Biosciences (简称 PacBio)公司开发的高精度长读长(HiFi)测序和 Oxford Nanopore Technologies 公司(简称 ONT)开发的 Nanopore 纳米孔测序为代表的第三代测序技术，呈现出长读长、准确度高、高通量、实时读取等多种优势，能够解锁基因组的复杂区域，使得无间隙基因的构建成为可能。ONT ultra-long 测序因其超长读长，轻松跨越基因组复杂区域，有效减少基因组的 gap，提高了基因组的完整性和连续性。PacBio HiFi 测序兼具超长读长和超高精度，成为构建无间隙基因组的理想工具。本研究首先利用 PacBio HiFi、Nanopore 测序以及 Hi-C 辅助组装技术构建湖鲚无间隙基因组，以期获得高质量湖鲚 gap-free 基因组序列。先前研究综合利用 PacBio HiFi、ONT ultra-long 和 Hi-C 等多种基因组测序技术及多种组装策略，成功获得了完全连续、无间隙(gap-free)刀鲚高质量基因组[17]。在此基础上，对湖鲚和刀鲚基因组进行共线性比对分析，鉴定基因组间的共线性区块。本研究结果将为丰富湖鲚基因组信息及鲚属鱼类遗传学研究积累基础资料，并为后续开展湖鲚适应性进化遗传机制提供理论依据。

1 材料与方法

1.1 样本采集及生活史验证

2021 年 9 月 17 日在中国江苏省苏州太湖水域(120°17′4.44298″E，31°19′27.28645″N)采集了 1 尾雌性湖鲚成鱼样本用于全基因组测序。首先将新鲜的湖鲚样本置于冰盘上，进行上颌骨长度、头长、全长、体重等表观生物学测量，保证所采集的样本为典型的长颌型(上颌骨长度与头长的比值大于 1) (图 1a)。随后用无菌剪刀和镊子取 2 g 肌肉组织装入 2 mL 冻存管中，立即置于液氮中迅速冷冻，并保存于−80 ℃超低温冰箱中，直至进行 DNA 提取。耳石微化学技术通过对湖鲚左矢耳石中锶 Sr 和钙 Ca 进行定量线分析，并将 Sr: Ca 值标准化为 Sr/Ca×1000，来反演其生活史特征及生境履历[18]，结果显示所采集的湖鲚样本耳石 Sr/Ca×1000 比值均处于小于 3 的低水平，表现为淡水履历的生境特征，保证本研究用于全基因组测序的样本为淡水定居型个体(图 1b)。

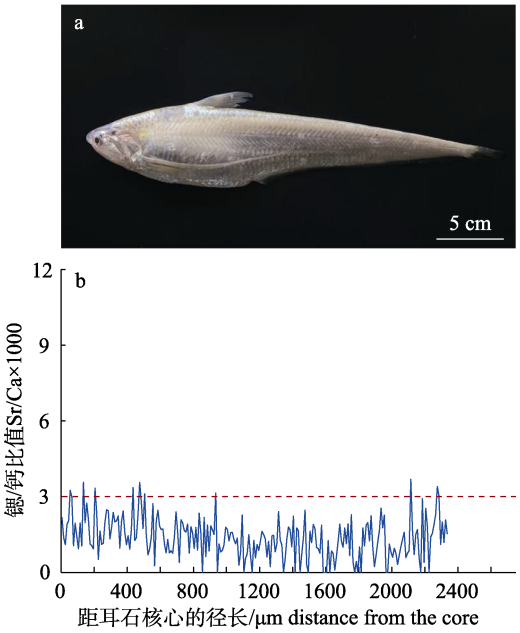

图 1 湖鲚样本形态特征(a)及其耳石从核心到边缘 Sr/Ca 比值(b)

Fig. 1 Morphological characteristics of sampled *Coilia nasus taihuensis* (a), and Sr/Ca ratio of its otolith from core to the edge (b)

1.2 文库构建和基因组测序

首先通过 DNeasy Blood & tissue Kit (Qiagen)试剂盒进行基因组 DNA 的提取，并利用 Qubit 荧光定量仪和 1%琼脂糖凝胶电泳检测 DNA 样品的浓度和完整性。对质检合格的 DNA 样本使用

Covaris E220 超声发生器(Covaris, Brighton, UK)进行片段化处理。选择 300~400 bp 左右的 DNA 片段经末端修复后在 3′端加上“A”碱基, 接入测序接头, 随后对片段进行单链分离、环化处理和滚环复制(rolling circle amplification, RCA), 生成 DNA 纳米球(DNA nano ball, DNB)。对质控合格的短读长(short-read)文库在 DNBSEQ 平台上进行高通量测序, 并保证每个样品的数据量达到要求。

HiFi 文库构建流程与二代测序大致相似, 利用 Megaruptor 随机打断全基因组 DNA, 筛选出长度为 13~16 kb 的 DNA 片段, 并在片段两端连接环状单链 PacBi 接头(adaptor), 形成“哑铃形”SMRT bell 文库, 在三代测序平台 Pacbio Sequel II (Pacific Biosciences, USA)上应用 SMRT 测序技术以环状共有序列(CCS)模式进行单分子荧光测序。Hi-C 文库构建不需要提取 DNA, 直接对细胞进行甲醛交联固定, 然后完成限制性内切酶酶切、末端修复、DNA 连接酶环化、DNA 纯化及目的 DNA 片段捕获等操作过程, 经 PCR 扩增后在 DNBSEQ 平台上测序。对于 ONT ultra-long 文库的构建, 使用 BluePippin/Pippin HT/Sage HLS 全自动核酸回收系统筛选大片段 DNA, 对 DNA 进行损伤修复和末端修复; 在纯化后的 DNA 末端加“A”碱基, 并对 DNA 进行接头连接, 使 DNA 片段带有可被识别的特异接头序列; 最后使用 Qubit 荧光仪(Invitrogen, 美国)对基因组文库进行精确定量, 测定浓度并计算回收率。建库完成后, 在华大基因研究中心 Nanopore PromethION 平台(Oxford Nanopore Technologies)上测序。

1.3　基因组大小和杂合度评估

基于二代短读长数据, 采用 K-mer 分析方法评估湖鲚基因组大小和杂合度。首先使用 Jellyfish (v2.2.6)软件快速统计 K-mer 频数(K=21), 再利用 GenomeScope 软件评估基因组大小、杂合度、重复序列、测序深度等基因组特征[19-21]。

1.4　基因组组装

基因组组装策略为: 首先利用 Pacbio HiFi 测序数据完成基因组的初步组装, 结合 Hi-C 测序技术获得基因在染色体上的相对位置, 完成基因组染色体水平的组装, 再结合 Nanopore 的超长读长数据进行基因组补洞, 最终获得 Gap-free 的基因组(图 2)。使用 CCS v4.0.0 和 SMRTLink v8.0.0 算法对 PacBio Sequel II CCS 测序下机的原始数据进行过滤处理, 参数为“--minPasses 3--min Predicted Accuracy 0.99--minLength 500”。利用 Hifiasm (v0.15.1)的默认参数进行初步组装, 随后使用 Purge Haplotigs 程序去除冗余序列, 获得 contigs 水平基因组草图。对 Hi-C 测序下机得到的原始数据(raw data)进行质控, 得到高质量的 clean data[22-23]。利用 Juicer v1.5 软件通过 BWA 将 clean data 比对到组装的基因组上, 进行比对分析[24]。再根据比对结果, 使用 3D DNA v180922 软件对组装好的 contigs 进行聚类、排序和定向, 辅助基因组锚定在染色体上, 完成 Hi-C 技术辅助基因组染色

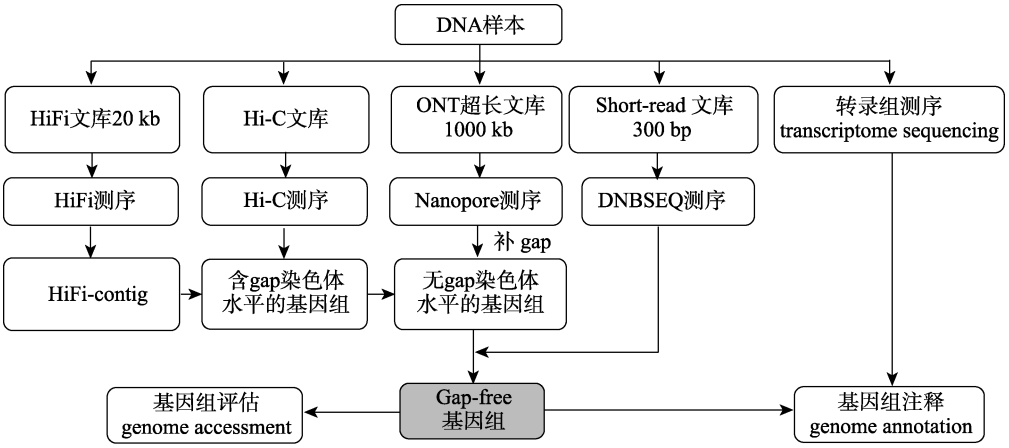

图 2　湖鲚基因组组装流程图

Fig. 2　Genome assembly flow chart of *Coilia nasus taihuensis*

体水平组装, 获得染色体长度的 Scaffolds[25-26]。对于基因组中存在的少量 gap 区域, 使用 LR_gapcloser 和 TGS_gapcloser 补洞软件通过 Nanopore 测序获得的超长序列及 Necat 组装结果来填充, 得到湖鲚 gap-free 基因组[27-28]。最后, 基于有胚植物数据库 actinopterygii_odb10, 利用 BUSCO (Benchmarking Universal Single-Copy Orthologs) v5.2.2 对基因组完整性进行评估[29]。

1.5 基因组注释

首先, 利用同源序列比对(homolog)和从头预测(*de novo*)两种方法对湖鲚基因组进行重复序列识别。使用 Tandem Repeat Finder v 4.10.0 软件和 LTR-FINDER v1.0.79 软件分别鉴定出基因组序列中的串联重复序列和长末端重复序列[30-31]。利用基于 RepBase 数据库的 RepeatMasker v4.0.7 软件和 RepeatProteinMasker v4.0.7 软件分别对基因组中的 DNA 和蛋白质转座元件进行筛选[32-33]。为了获得蛋白质编码基因, 结合同源预测、从头预测和 RNA-Seq 辅助预测三种预测方法进行基因结构预测。对于同源注释, 利用从 NCBI 数据库下载的 6 个近缘物种[大西洋鲱(*Clupea harengus*)、斑马鱼(*Danio rerio*)、齿鲱(*Denticeps clupeoides*)、电鳗(*Electrophorus electricus*)、虹鳟(*Oncorhynchus mykiss*)、沙丁鱼(*Sardina pilchardus*)]基因组组装和基因注释文件, 使用 GeMoMa 软件进行基因结构预测[34]。对于肝、脑、胃的转录组数据, 使用 HISAT2 将 RNA-seq 数据比对到基因组, 然后使用 StringTie 进行转录本的拼接。使用 GeMoMa 软件整合以上三种策略获取的所有结果, 生成非冗余基因集[35-36]。另外, 利用本课题组组装的刀鲚、短颌鲚、凤鲚和七丝鲚基因组(数据未发表), 评估鲚属物种在基因、CDS、外显子和内含子水平上的分布特征。最后, 对获得的基因进行基因功能注释, 通过将注释得到的基因集与功能数据库 KEGG (Kyoto Encyclopedia of Genes and Genome, http://www.genome.jp/kegg/)、Swissport (http://www.gpmaw.com/html/swiss-prot.html)、TrEMBL (http://www.gpmaw.com/html/swiss-prot.html)进行比对, 注释蛋白质的生物学通路和功能, 使用 InterProScan 注释蛋白质结构域和基序[37]。

1.6 共线性分析

为了评估两个基因组的染色体水平共线性, 根据先前研究得到的刀鲚全基因组序列(https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_027475355.1/)[17], 利用 MUMmer 软件(http://mummer.sourceforge.net/)中的 nucmer 程序对湖鲚和刀鲚基因组进行比对, 寻找基因组间的共线性区块[38]。通过共线性分析获得的 collinearity 文件, 利用 MUMmer 软件工具中的 mummerplot 对基因组间的共线性区块进行可视化。

2 结果与分析

2.1 基因组测序数据统计

为了构建高质量的湖鲚基因组, 本研究共获取了 599.32Gb 原始测序数据量, 总测序深度为 230×。构建 1 个 short-reads 文库, 过滤得到 57.22Gb clean reads (25×)。构建 1 个 20 kb HiFi 文库, 通过 CCS 过滤得到 1.48M 条 HiFi reads, 获得 21.16Gb 的 PacBio HiFi reads (67×)。构建 1 个 ONT ultra-long 文库, 过滤得到 0.35M 条 clean reads, 获得 15.33Gb 的 ONT ultra-long reads (18×)。构建 1 个 Hi-C 文库, 过滤得到 101.75Gb Hi-C reads (120×) (表 1)。

2.2 Survey 分析及基因组大小

对于 DNBseq 测序平台获取的短读长数据进

表 1 湖鲚基因组测序数据统计

Tab. 1 Sequencing data for genome assembly of *Coilia nasus taihuensis*

测序文库 sequencing library	插入片段大小/bp insert size	过滤后数据量/Gb clean data	平均长度/bp average length	测序深度/× sequence coverage
WGS	300–400	57.22	150	25
PacBio HiFi	20000	21.16	14297	67
Hi-C	300–400	101.75	150	120
ONT		15.33	61129	18

行基因组 survey, 评估基因组大小。经 21-mer 分析估计基因组大小为 612.76 Mb, 杂合度为 1.31%, 属于高杂合基因组(图 3)。

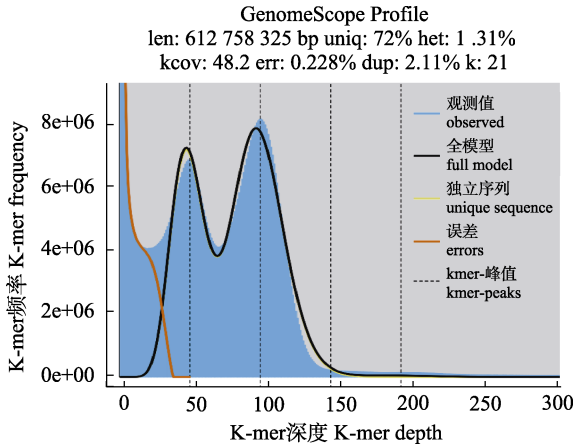

图 3　基于 K-mer 频率分布估计湖鲚基因组大小

最高峰为纯合峰, 所对应的为 K-mer 分布最多的深度. 蓝色区域代表实际观测值; 黑色拟合线代表除去错误部分之后剩下的所有 K-mer; 黄色拟合线代表基因组非重复区域的 K-mer 分布; 橙红色拟合线代表深度过低的 K-mer; 垂直的黑色虚线代表预测最低深度峰的整倍数覆盖度.

Fig. 3　K-mer distribution analysis to estimate the genome size of *Coilia nasus taihuensis*

The tallest peak represents the homozygous peak, corresponding to the depth where the K-mer distribution is the most abundant. Blue area: observed K-mer frequencies (raw data); black fitted curve: filtered k-mer distribution after error correction; yellow fitted curve: K-mer profile of non-repetitive genomic regions (unique sequence); orange-red fitted curve: low-coverage k-mers (errors); vertical black dashed line: predicted whole-genome coverage multiples (x-fold).

2.3　基因组组装及组装质量评估

结合 PacBio HiFi、ONT ultra-long 和 Hi-C 测序数据进行基因组组装, 获得了一个 24 条染色体的高质量基因组(图 4)。基因组 contig 水平基因组大小为 829.28 Mb, contig N50 为 0.68 Mb, 经过 Hi-C 挂载, 其中挂载到染色体的基因组 827.94 Mb, 染色体挂载率为 99.83%。获得染色体级别的基因组后, 利用 Nanopore 测序得到的 ONT ultra-long reads 填补 HiFi 组装的 gap 区域, 得到湖鲚 gap-free 基因组大小为 834.09 Mb, contig N50 为 35.46 Mb, 基因组的 GC 含量为 44%, 与硬骨鱼类的基因组含量基本一致(表 2)。使用 BUSCO 对基因组组装质量进行评估, 共组装出 3344 个(91.90%)单拷贝直系同源基因, 只有 208 个(5.7%)基因未检测到, 表明基因组组装完整性较好(表 3)。基因组测序原始数据上传至美国国家生物技术信息中心(National Center for Biotechnology Information)Sequence Read Archive (SRA) 数据库中, 序列号为 PRJNA1054695。

2.4　基因组注释

本研究对湖鲚基因组进行了重复序列注释、基因结构预测和基因功能注释。基因组重复序列长度为 382.39 Mb, 占基因组的 45.85%, 其中长末端重复序列(LTR, 11.44%)、长散在重复序列(LINE, 11.03%)和短散在重复序列(SINE, 0.4%)共同构成了较大的比例, 占比为 22.87%。此外, 对注释出的重复序列进行分类统计发现 DNA 转座子占基因组的 15.76%, 还包含 1.73%未被分类的重复序列(表 4)。结合从头预测、同源比对和转录组数据辅助注释对湖鲚基因组进行注释, 共注释出 21730 个蛋白质编码基因, 基因平均长度、CDS 平均长度、外显子平均长度和内含子平均长度分别为 22130 bp、1560 bp、172 bp 和 2542 bp (表 5)。与近缘物种刀鲚、短颌鲚、凤鲚和七丝鲚相比, 基因模型的长度在基因、CDS、外显子和内含子水平上具有相似的分布趋势(图 5)。

将基因结构预测得到的基因集与 Nr、Interpro、KEGG、Swissport、TrEMBL 和 KOG 等数据库进行比对, 对基因进行功能注释。其中 Nr 数据库基因组注释基因最多, 占总基因数的 99.66%。Swissport 数据库注释 20501 个基因, 占比 94.34%。KEGG 数据库注释 17625 个基因, 占比 90.92%。GO 数据库共比对出 15588 个基因, 占比 71.73%。综合所有数据库的注释结果, 21666 个基因(99.71%)被至少一个数据库注释(表 6)。其中 16976 个基因在五个数据库中均被注释出, 占总注释基因数的 78.35% (图 6)。

2.5　基因组共线性分析

利用 MUMmer 软件将刀鲚和湖鲚基因组相互比对来进行全基因组共线性分析, 结果显示, 刀鲚和湖鲚之间具有极高的基因组共线性比率(96.95%), 共鉴定到 48852 个共线性区块。根据刀鲚的基因组序列绘制共线性比对图, 每个共线性区域为 1∶1 的比对, *x* 轴为刀鲚基因组的染色体

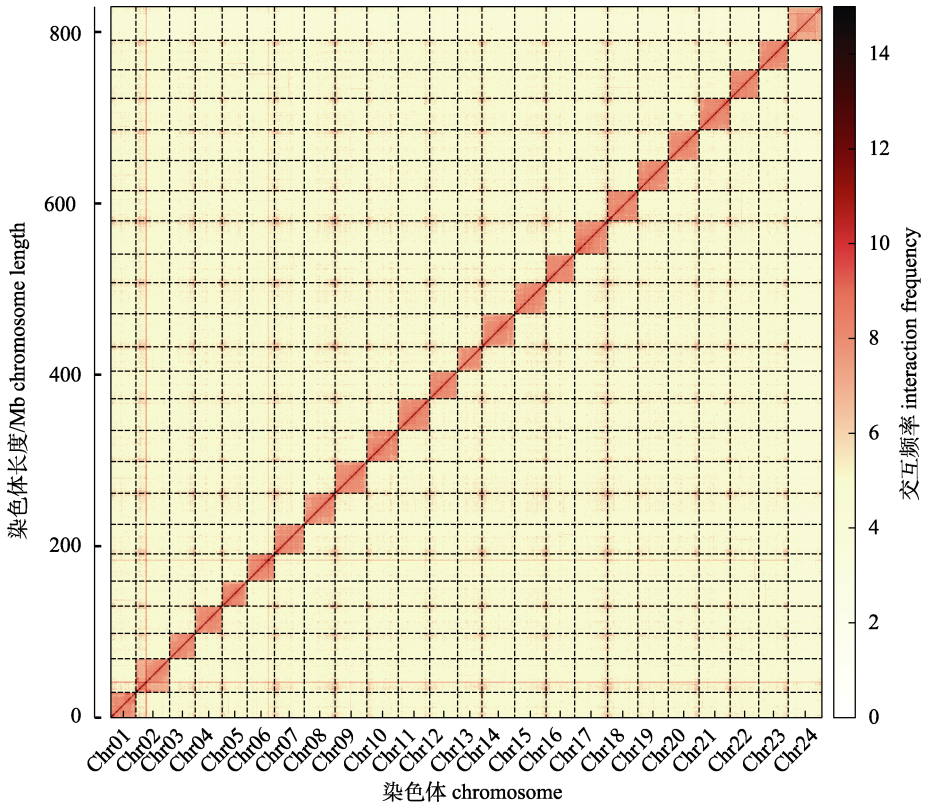

图 4 湖鲚染色体 Hi-C 互作图(带宽为 500 kb)

颜色由浅到深表示互作强度增加, 对角线表示染色体内的相互作用最强.

Fig. 4 Genomic-wide Hi-C heatmap of *Coilia nasus taihuensis* (bin=500 kb)

The color from light to dark represents the contact density from low to high, with the diagonal line reflecting the highest-intensity intra-chromosomal interactions.

表 2 湖鲚基因组组装信息统计

Tab. 2 Statistics of the assembled contigs and scaffolds for *Coilia nasus taihuensis* genome

基因组组装指标 genome assembly	无间隙组装 gap-free assembly		初步组装 primal contig	
	长度/bp length	数量 number	长度/bp length	数量 number
N50	35455198	12	683430	279
N90	29405414	22	102740	1498
最大长度 maximum length	43425900		8165644	
gap 数量 gap number	0	0	0	0
总长 total length	834090801		829283070	
总数量 total number		72		3193
GC 含量/% GC content		44.0		44.0

表 3 湖鲚基因组 BUSCO 评估

Tab. 3 BUSCO scores of the assembled *Coilia nasus taihuensis* genome

类型 type	数量 count	比例/% percentage
完整的 BUSCOs (C) complete BUSCOs (C)	3344	91.9
完整的单拷贝 BUSCOs (S) complete and single-copy BUSCOs (S)	3276	90
完整的重复序列 BUSCOs (D) complete and duplicated BUSCOs (D)	68	1.9
碎片 BUSCOs (F) fragmented BUSCOs (F)	88	2.4
未比对上的 BUSCOs (M) missing BUSCOs (M)	208	5.7
总的 BUSCO total BUSCO	3640	100

表 4　湖鲚基因组重复序列信息

Tab. 4　Information on repetitive sequence of *Coilia nasus taihuensis* genome

类型 type		序列长度/bp sequence length	百分比/% percentage
反转录转座子 retrotransposon	LTR/Copia	1836103	0.22
	LTR/Gypsy	31820184	3.82
	LTR/Other	61757601	7.40
	SINE	3306257	0.40
	LINE	91992097	11.03
DNA 转座子 DNA transposon	EnSpm	131421912	15.76
	Harbinger	7571243	0.90
	hAT	47469205	5.69
	Helitron	25540061	3.06
	Mariner	2150949	0.26
	MuDR	1488226	0.18
	P	1934830	0.23
	other	157778939	18.92
其他 other		31822582	3.82
未知 unknown		14464702	1.73

表 5　湖鲚基因结构注释信息

Tab. 5　General statistics of predicted protein-coding genes in *Coilia nasus taihuensis* genome

注释方法 annotation method	物种 species	基因数量 number of gene	基因平均长度/bp average gene length	CDS 平均长度/bp average CDS length	基因平均外显子数量 average exon per gene	外显子平均长度/bp average exon length	内含子平均长度/bp average intron length
从头预测 *de novo*		51617	8555	1150	6	204	1596
同源预测 homolog	大西洋鲱 *Clupea harengus*	24957	22870	1716	10	181	2488
	齿鲱 *Denticeps clupeoides*	20461	25268	1780	10	173	2533
	斑马鱼 *Danio rerio*	21311	23686	1745	10	183	2571
	电鳗 *Electrophorus electricus*	19583	25302	1762	10	172	2542
	虹鳟 *Oncorhynchus mykiss*	22229	25123	1755	10	174	2576
	沙丁鱼 *Sardina pilchardus*	27664	14707	1185	5	217	3033
转录预测 transcript		29918	22030	3224	10	320	2074
总计 total		21730	22130	1560	9	172	2542

尺度, *y* 轴为湖鲚基因组的染色体尺度, 染色体的链接依次为 Chr1 到 Chr24 (图 7), 接近平滑的直线显示出刀鲚与湖鲚基因组间的高度相似的共线性, 证明两者基因组的相似性高度一致。

3　讨论

3.1　湖鲚基因组组装质量

长读长测序技术(long-read sequencing, LRS)

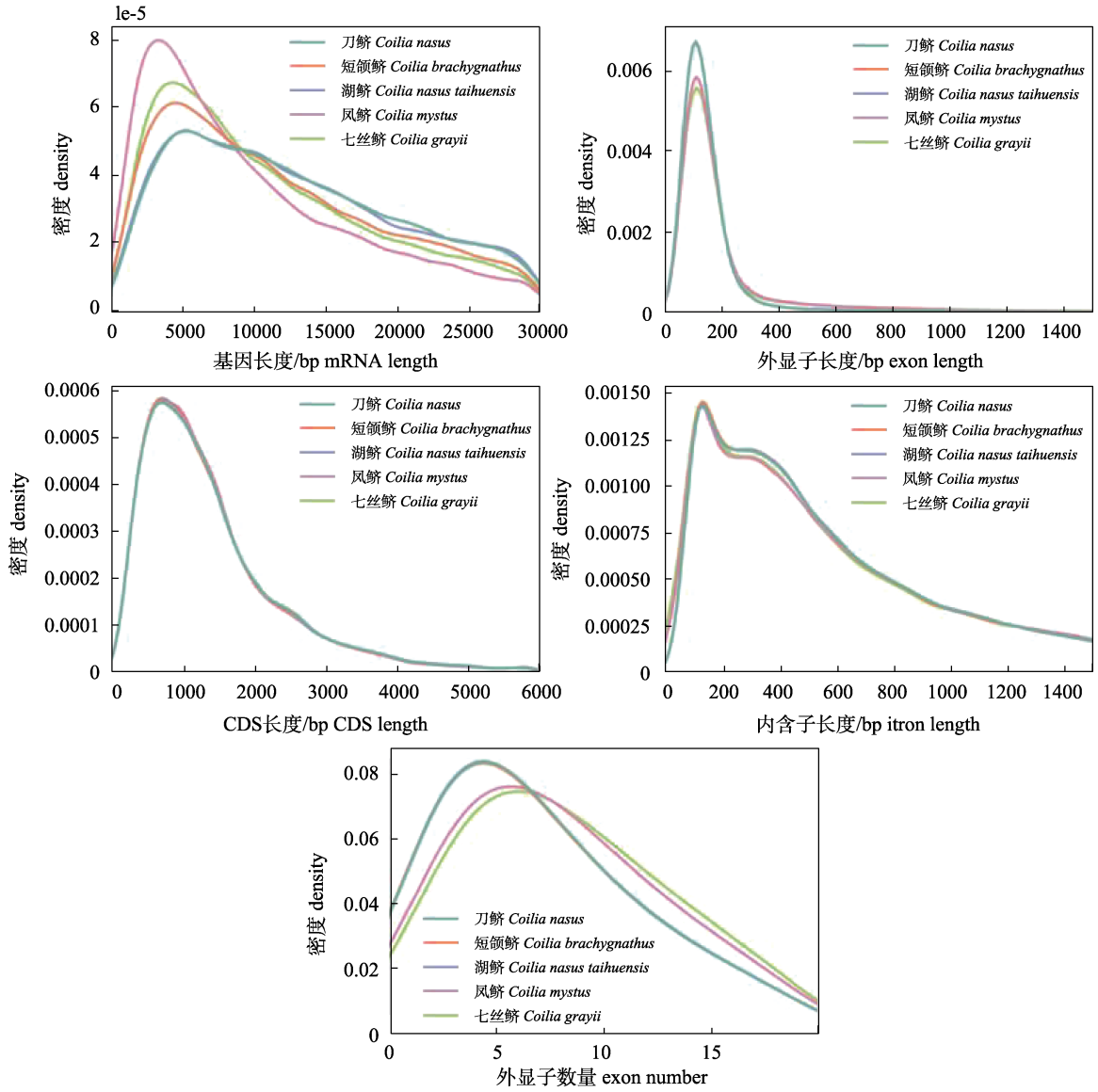

图 5 湖鲚与其他近缘物种基因组的基因结构注释信息比较

Fig. 5 Comparison of gene structure features of *Coilia nasus taihuensis* with other fish species

表 6 湖鲚基因功能注释表

Tab. 6 Functional annotations of *Coilia nasus taihuensis*

数据库 database	数量 number	百分比/% percentage
Nr	21657	99.66
Swissport	20501	94.34
KEGG	19758	90.92
KOG	17652	81.23
TrEMBL	21656	99.66
Interpro	21233	97.71
GO	15588	71.73
总计 total	21666	99.71

可以产生长度≥10 kb 的连续序列，为基因组探索提供了强有力的工具，使得无间隙基因组组装成为可能。无间隙基因组能够全面地识别基因组信息，被广泛视为基因组组装的最终目标。先前的研究已经在多个物种中发布了无间隙基因组，包括刀鲚(*Coilia nasus*)、东亚江豚(*Neophocaena asiaeorientalis sunameri*)、水稻(*Oryza sativa*)、西瓜(*Citrullus lanatus*)等[17,39-41]。与二代短读长测序技术相比，三代测序能够有效跨越高度重复的基因组区域，并且能够填补二代测序技术在组装中难以解决的空缺，从而显著提高基因组的整体连

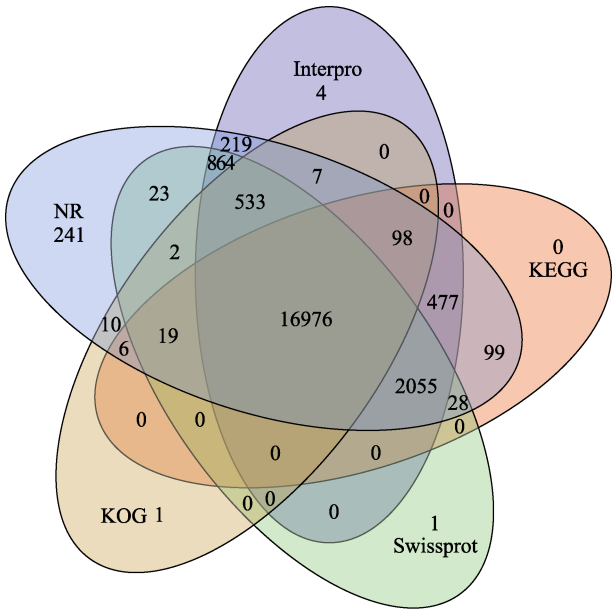

图 6　基因功能注释结果 Venn 图

Fig. 6　Venn diagram of gene function annotation results in the five databases of NR, Interpro, KEGG, Swissprot and KOG

续性。本研究以太湖湖鲚为材料，整合 PacBio HiFi、ONT ultra-long 和 Hi-C 等多种测序技术以及多种组装策略完成了湖鲚的 gap-free 基因组组装。湖鲚基因组大小为 834.09 Mb, contig N50 达到了 35.46 Mb, BUSCO 评估完整性为 91.90%。PacBio 和 Nanopore 平台提供的长读长测序技术促进了基因组的组装，显著提高了基因组连续性或完整性[42]。本研究获得的湖鲚基因组大小为 834.09 Mb，与先前发表的洄游型刀鲚(851.67 Mb)相比，基因组长度存在差异[17]。随着 PacBio 和 Nanopore 三代测序技术的不断升级，基因组组装指标 contig N50 有明显提升, contig N50 由数 Mb 提升至上百 Mb 的水平。本研究获得的湖鲚 gap-free 参考基因组 contig N50 达到了 35.46 Mb，与刀鲚基因组(N50 35.42 Mb)质量处于同一水平，表明湖鲚基因组组装结果的可靠性[17]。从基因组组装完整性来看, BUSCO 评价湖鲚的基因组完整度为 91.9%，表明湖鲚基因组装的高度完整性。另外，与 2020 年公布的养殖刀鲚基因组相比[43]，本研究组装的基因组在连续性(contig N50 值)和完整度上有明显提升，其中 contig N50 从 1.6 Mb 提

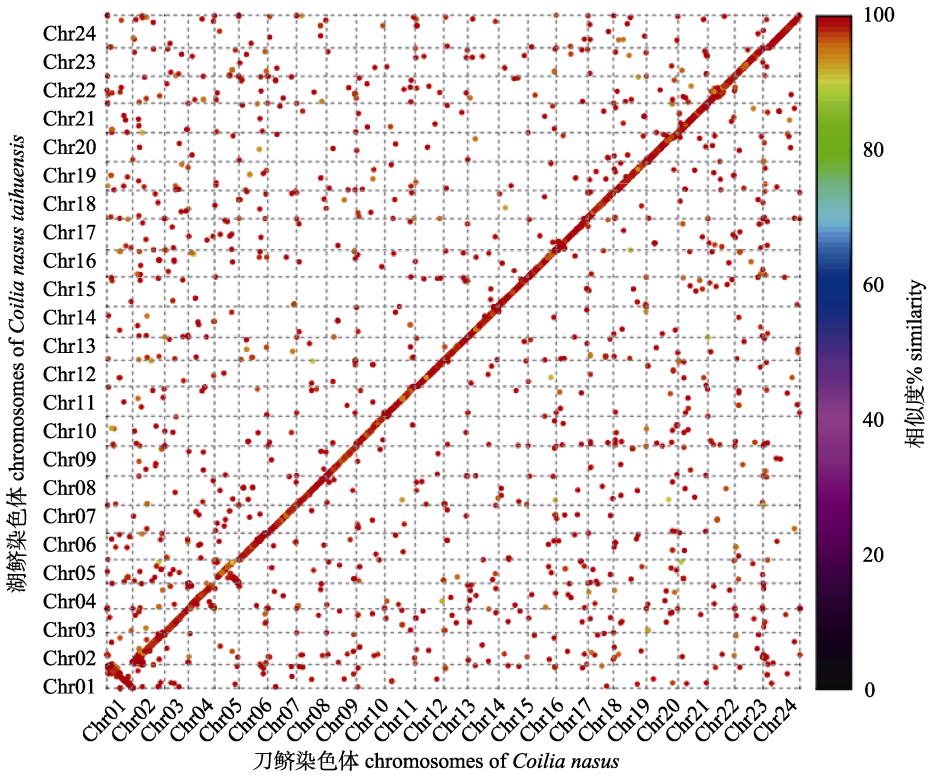

图 7　湖鲚与刀鲚基因组序列共线性比较

Fig. 7　Collinearity of the genomes of *Coilia nasus* and *Coilia nasus taihuensis*

升至 35.46 Mb, 完整度也由 87.1%提高到 91.9%, 表明湖鲚基因组实现了高质量组装。Hi-C 挂载到了 24 条染色体上, 与染色体核型分析的染色体数目(2n=48)一致, 挂载率达 99.83%及 Hi-C 互作图均反映出染色体水平的基因组组装效果良好。

3.2 湖鲚基因组注释特征

与大多数鱼类基因组一样, 湖鲚基因组的重复元件占比较高。基因组重复序列注释显示, 湖鲚基因组重复序列主要由 SINE、LINE、LTR 和 DNA 转座子四种类型组成, 序列总长度为 382.39 Mb, 占基因组的 45.85%, 揭示了湖鲚基因组内重复元件的复杂组成。湖鲚基因组重复序列的总长度和比例显著高于养殖刀鲚注释到的重复序列(长度为 245.66Mb, 占基因组的 31.24%), 表明本研究获得的湖鲚基因组重复序列的注释更加完善[43]。结合同源预测、从头预测和 RNA-Seq 辅助预测多种策略注释得到 21730 个蛋白质编码基因, 与其他硬骨鱼基因数目相比, 湖鲚编码基因的数目接近于鲤科鱼类黄鲫(*Setipinna tenuifilis*), 低于大口黑鲈(*Micropterus salmoides*)[44-45]。从编码基因的结构来说, 湖鲚基因平均长度和 CDS 平均长度分别为 22130 bp 和 1560 bp, 与刀鲚、短颌鲚、凤鲚和七丝鲚基因组注释的基因平均长度、CDS 平均长度、外显子平均长度和内含子平均长度分布具有相似性, 表明湖鲚与已发表的刀鲚基因组具有相似的基因结构分布模式[17]。基因功能注释得到 21666 个功能基因, 占比 99.71%, 为进一步的基因功能研究和通路分析提供了丰富的信息资源。以上结果均表明, 本研究获得的湖鲚基因组注释结果质量较高, 将为进一步阐明湖鲚基因组层面的分子机制提供重要的基因组资源。

3.3 基因组共线性分析

共线性分析是比较基因组学中的重要内容之一, 在共线性区域内, 基因的序列及其相对位置在一定程度上是保守的[46]。通过对湖鲚和刀鲚全基因组序列进行比对, 结果显示, 两基因组之间表现出极高的共线性(相似度 96.95%), 表明两者有着极近的亲缘关系, 进一步支持了湖鲚为刀鲚淡水生态型这一结论[12]。共线性片段的大小与分化时间密切相关, 分化时间较短的群体间由于积累的变异较少能够保留更多祖先遗传的特征; 而分化时间较长的群体间则因变异累积较多, 导致共有的特征减少, 共线性片段较短。湖鲚和刀鲚共线性比对图显示, 染色体之间除少数基因存在倒置或插入外, 各染色体整体呈一一对应关系, 表明两者在整个基因组水平上表现出高度的保守性。本研究通过刀鲚和湖鲚比较基因组共线性分析, 鉴定出 48852 个共线性区块。基于先前研究揭示的湖鲚和刀鲚较短的分化时间(约 3 千年前)[12], 两者之间较大的共线性区块表明湖鲚积累的变异较少, 在基因组序列和功能上具有高度保守性, 也从侧面印证了湖鲚良好的基因组组装效果。

参考文献:

[1] Jiang T, Liu H B, Xuan Z Y, et al. Classification of ecomorphotypes of *Coilia nasus* from the middle and lower reaches of the Yangtze River Basin[J]. Journal of Lake Science, 2020, 32(2): 518-527. [姜涛, 刘洪波, 轩中亚, 等. 长江中下游流域刀鲚(*Coilia nasus*)生态表型的划分[J]. 湖泊科学, 2020, 32(2): 518-527.]

[2] Yuan C M, Qin A L, Liu R H, et al. On the classification of the anchovies, *Coilia*, from the lower Yangtze River and the southeast coast of China[J]. Journal of Nanjing University (Natural Science), 1980, 3: 67-77 [袁传宓, 秦安舲, 刘仁华, 等. 关于长江中下游及东南沿海各省的鲚属鱼类种下分类的探讨[J]. 南京大学学报(自然科学版), 1980, 3: 67-77.]

[3] Cheng Q Q, Li S F. Morphological discrimination between two populations of *Coilia ectenes*[J]. Marine Sciences, 2004, 28(11): 39-43. [程起群, 李思发. 刀鲚和湖鲚种群的形态判别[J]. 海洋科学, 2004, 28(11): 39-43.]

[4] Yuan C B, Lin J B, Qin A L, et al. The history and current situation of the classification of Chinese genus *Coilia*-some experience on the transformation of the old taxonomy of fishes[J]. Journal of Nanjing University (Natural Sciences), 1976, 2: 1-12. [袁传宓, 林金榜, 秦安舲, 等.关于我国鲚属鱼类分类的历史和现状-兼谈改造旧鱼类分类学的几点体会[J]. 南京大学学报(自然科学版), 1976, 2: 1-12.]

[5] Yang Q L. Phylogenetic analysis of genus *Coilia* in China and molecular phylogeography of *C. nasus* and *C. mystus*[D]. Qingdao: Ocean University of China, 2012. [杨巧莉. 中国鲚属鱼类进化关系及刀鲚、凤鲚的分子系统地理学研究[D]. 青岛: 中国海洋大学, 2012.]

[6] Liu W B. Biochemical and morphological comparison and interspecific relationships of four species of the genus *Coilia*

in China[J]. Oceanologia et Limnologia Sinica, 1995, 26(5): 558-565. [刘文斌. 中国鲚属 4 种鱼的生化和形态比较及其系统发育的研究[J]. 海洋与湖沼, 1995, 26(5): 558-565.]

[7] Xiang W D, Xie J Y, Lin J. Morphological variations of *Coilia ectenes* in different lakes[J]. Hubei Agricultural Sciences, 2011, 50(21): 4445-4447. [向文殿, 谢佳燕, 林佳. 不同湖泊湖鲚种群形态差异的研究[J]. 湖北农业科学, 2011, 50(21): 4445-4447.]

[8] Cheng Q Q, Zhang Q Y, Ma C Y, et al. Genetic structure and differentiation of four lake populations of *Coilia ectenes* (Clupeiformes: Engraulidae) based on mtDNA control region sequences[J]. Biochemical Systematics and Ecology, 2011, 39(4): 544-552.

[9] Zhou X D, Yang J Q, Tang W Q, et al. Species validities analyses of Chinese *Coilia* fishes based on mtDNA COI barcoding[J]. Acta Zootaxonomica Sinica, 2010, 35(4): 819-826. [周晓犊, 杨金权, 唐文乔, 等. 基于线粒体 COI 基因 DNA 条形码的中国鲚属物种有效性分析[J]. 动物分类学报, 2010, 35(4): 819-826.]

[10] Yang J Q, Hu X L, Tang W Q, et al. mtDNA control region sequence variation and genetic diversity of *Coilia nasus* in Yangtze River estuary and its adjacent waters[J]. Chinese Journal of Zoology, 2008, 43(1): 8-15.[杨金权, 胡雪莲, 唐文乔, 等. 长江口邻近水域刀鲚的线粒体控制区序列变异与遗传多样性[J]. 动物学杂志, 2008, 43(1): 8-15.]

[11] Cheng Q Q, Wen J E, Wang Y L, et al. Genetic diversity and genetic differentiation between *Coilia ectenes* and *Coilia ectenes taihuensis* inferred from cytochrome b gene segment sequence of mitochondrial DNA[J]. Journal of Lake Science, 2006, 18(4): 425-430. [程起群, 温俊娥, 王云龙, 等. 刀鲚与湖鲚线粒体细胞色素 b 基因片段多态性及遗传关系[J]. 湖泊科学, 2006, 18(4): 425-430.]

[12] Cheng F Y, Wang Q, Maisano Delser P, et al. Multiple freshwater invasions of the tapertail anchovy (Clupeiformes: Engraulidae) of the Yangtze River[J]. Ecology and Evolution, 2019, 9(21): 12202-12215.

[13] Xuan Z Y, Jiang T, Liu H B, et al. Mitochondrial DNA and microsatellite analyses reveal strong genetic differentiation between two types of estuarine tapertail anchovies (*Coilia*) in Yangtze River Basin, China[J]. Hydrobiologia, 2021, 848: 1409-1431.

[14] Zong S B, Li Y L, Liu J X. Genomic architecture of rapid parallel adaptation to fresh water in a wild fish[J]. Molecular Biology and Evolution, 2021, 38(4): 1317-1329.

[15] Li D M, Tang S K, Liu Y S, et al. Genetic diversity and genetic structure of *Coilia nasus taihuensis* populations in Jiangsu Province based on mtDNA control region sequences [J]. Marine Fisheries, 2021, 43(2): 149-159. [李大命, 唐晟凯, 刘燕山, 等. 基于线粒体控制区的江苏湖鲚群体遗传多样性和遗传结构分析[J]. 海洋渔业, 2021, 43(2): 149-159.]

[16] Li X Q, Liu F, Leng C M, et al. Genetic structure and diffusion of population of *Coilia ectenes taihuensis* in Lake Nansi inferred from the mitochondrial control region [J]. Journal of Lake Science, 2015, 27(4): 686-692. [李秀启, 刘峰, 冷春梅, 等. 基于线粒体 DNA 控制区的南四湖湖鲚(*Coilia ectenes taihuensis*)群体遗传结构和种群扩散[J]. 湖泊科学, 2015, 27(4): 686-692.]

[17] Ma F J, Wang Y P, Su B X, et al. Gap-free genome assembly of anadromous *Coilia nasus*[J]. Scientific Data, 2023, 10(1): 360.

[18] Jiang T, Liu H B, Hu Y H, et al. Revealing population connectivity of the estuarine tapertail anchovy *Coilia nasus* in the Changjiang River estuary and its adjacent waters using otolith microchemistry[J]. Fishes, 2022, 7(4): 147.

[19] Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers[J]. Bioinformatics, 2011, 27(6): 764-770.

[20] Liu B H, Shi Y J, Yuan J Y, et al. Estimation of genomic characteristics by analyzing k-mer frequency in *de novo* genome projects[J]. Quantitative Biology, 2013, 35(1): 62-67.

[21] Vurture G W, Sedlazeck F J, Nattestad M, et al. GenomeScope: fast reference-free genome profiling from short reads [J]. Bioinformatics, 2017, 33(14): 2202-2204.

[22] Cheng H Y, Concepcion G T, Feng X W, et al. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm[J]. Nature Methods, 2021, 18(2): 170-175.

[23] Roach M J, Schmidt S A, Borneman A R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies[J]. BMC Bioinformatics, 2018, 19: 1-10.

[24] Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform[J]. Bioinformatics, 2010, 26(5): 589-595.

[25] Dudchenko Olga, Batra Sanjit S, Omer Arina D, et al. *De novo* assembly of the genome using Hi-C yields chromo. some-length scaffolds[J]. Science, 2017, 356(6333): 92-95.

[26] Rao S P, Huntley M H, Durand N C, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping[J]. Cell, 2014, 159(7): 1665-1680.

[27] Xu G C, Xu T J, Zhu R, et al. LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly[J]. Gigascience, 2019, 8(1): 157.

[28] Xu M, Guo L, Gu S, et al. TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of

error-prone long reads[J]. GigaScience, 2020, 9(9): 94.

[29] Simão F A, Waterhouse R M, Ioannidis P, et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs[J]. Bioinformatics, 2015, 31(19): 3210-3212.

[30] Benson G. Tandem repeats finder: a program to analyze DNA sequences[J]. Nucleic Acids Research, 1999, 27(2): 573-580.

[31] Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons[J]. Nucleic Acids Research, 2007, 35: 265-268.

[32] Price A L, Jones N C, Pevzner P A. *De novo* identification of repeat families in large genomes[J]. Bioinformatics, 2005, 21: 351-358.

[33] Bao W D, Kojima K K, Kohany O, et al. Repbase Update, a database of repetitive elements in eukaryotic genomes[J]. Mobile DNA, 2015, 6: 1-6.

[34] Keilwagen J, Hartung F, Grau J. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data[J]. Gene prediction: Methods and Protocols, 2019, 1962: 161-177.

[35] Kim D, Paggi J M, Park C, et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype[J]. Nature Biotechnology, 2019, 37(8): 907-915.

[36] Kovaka S, Zimin A V, Pertea G M, et al. Transcriptome assembly from long-read RNA-seq alignments with StringTie2[J]. Genome Biology, 2019, 20: 1-13.

[37] Zdobnov E M, Apweiler R. InterProScan–an integration platform for the signature-recognition methods in InterPro[J]. Bioinformatics, 2001, 17(9): 847-848.

[38] Marçais G, Delcher A L, Phillippy A M, et al. MUMmer4: A fast and versatile genome alignment system[J]. PLoS Computational Biology, 2018, 14(1): 1005944.

[39] Yin D H, Chen C H, Lin D Q, et al. Gapless genome assembly of East Asian finless porpoise[J]. Scientific Data, 2022, 9(1): 765.

[40] Zhang Y L, Fu J, Wang K, et al. The telomere-to-telomere gap-free genome of four rice parents reveals SV and PAV patterns in hybrid rice breeding[J]. Plant Biotechnology Journal, 2022, 20(9): 1642-1644.

[41] Deng Y, Liu S, Zhang Y, et al. A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene discovery and breeding[J]. Molecular Plant, 2022, 15(8): 1268-1284.

[42] Pollard M O, Gurdasani D, Mentzer A J, et al. Long reads: their purpose and place[J]. Human Molecular Genetics, 2018, 27(2): 234-241.

[43] Xu G C, Bian C, Nie Z J, et al. Genome and population sequencing of a chromosome-level genome assembly of the Chinese tapertail anchovy (*Coilia nasus*) provides novel insights into migratory adaptation[J]. GigaScience, 2020, 9(1): 157.

[44] Liu B, Li J, Peng Y, et al. Chromosome-level genome assembly and population genomic analysis reveal evolution and local adaptation in common hairfin anchovy (*Setipinna tenuifilis*)[J]. Molecular Ecology, 2024, 33(10): 17067.

[45] Sun C, Li J, Dong J, et al. Chromosome-level genome assembly for the largemouth bass *Micropterus salmoides* provides insights into adaptation to fresh and brackish water [J]. Molecular Ecology Resources, 2021, 21(1): 301- 315.

[46] Tang H, Bowers J E, Wang X, et al. Synteny and collinearity in plant genomes[J]. Science, 2008, 320(5875): 486-488.

Gap-free genome assembly of freshwater resident *Coilia nasus*

MA Fengjiao[1], WANG Hui[1], REN Long[2], LIU Kai[1, 2*]

1. Wuxi Fisheries College, Nanjing Agricultural University, Wuxi 214081, China;
2. Key Laboratory of Freshwater Fisheries and Germplasm Resources Utilization, Ministry of Agriculture and Rural Affairs; Freshwater Fisheries Research Center, Chinese Academy of Fishery Sciences, Wuxi 214081, China

Abstract: As one of the ecotypes of anadromous *Coilia nasus*, the freshwater resident *Coilia nasus taihuensis* is unable to conduct anadromous behavior due to the existence of physical barriers. *C. nasus taihuensis* can grow and reproduce independently in freshwater habitats such as Lake Taihu. Genetic differences in the genome have been demonstrated between anadromous *C. nasus* and non-anadromous *C. nasus taihuensis* due to geographical isolation and the significant difference in living environment. The genetic mechanisms related to environmental adaptation of *C. nasus taihuensis* were hindered by the lack of reference genomes. To conduct comprehensive and systematic analysis of the genomic characteristics of *C. nasus taihuensis*, a freshwater resident individual from Lake Taihu was selected as the experimental object. The first complete, gap-free reference genome of *C. nasus taihuensis* was constructed using a combination of PacBio (Pacific Biosciences) high-fidelity reads and ONT (Oxford Nanopore Technologies) ultra-long reads and Hi-C datasets. Based on a K-mer value of 21, we estimated the genome size as 612.76 Mb with a heterozygosity rate of 1.31%. The results of genome assembly showed a gap-free *C. nasus taihuensis* genome with an assembly size of 834.09 Mb. A contig N50 of 35.45 Mb was obtained through library construction, sequencing, assembly, chromosome mounting and gaps filling. Compared with the genome of cultured *C. nasus*, our gap-free genome substantially improved contiguity and completeness, where contig N50 increased from 1.6 Mb to 35.46 Mb. The integrity also increased from 87.1% to 91.9%, representing the highest quality genome. By Hi-C data, the assembled sequences were anchored and oriented onto all the 24 chromosomes with a total length of 829.28 Mb, covering 99.83% of the scaffold-level genome. After gap filling using ONT ultra-long reads, all 24 chromosomes were assembled without gaps, representing the highest assembly quality. BUSCO analysis based on the actinopterygii_odb10 database showed that 91.9% of the expected actinopterygii_odb10 genes (single-copy genes: 90%; duplicated genes: 1.9%) were identified as complete, suggesting that the assembled *C. nasus taihuensis* genome is highly complete. Furthermore, a total of 382.393 Mb of repetitive sequence, accounted for 45.85% of the genome. Using a combination of *de novo* prediction, protein homology and RNA-seq annotation, a total of 21730 protein-coding genes were identified. Approximately 99.71% (21666 genes) of the total predicted genes were assigned with at least functional annotation, showing a more complete annotation. Furthermore, the conservation synteny between *C. nasus taihuensis* and *C. nasus* was compared to validate the chromosome assembly, and the results showed that the genomic sequences of *C. nasus* and *C. nasus taihuensis* were highly consistent (96.95%). Such highly conserved synteny and strict correspondence of chromosome assignment indicated a close genetic relationship between the two ecotypes. The high-quality gap-free assembled genome of *C. nasus taihuensis* can provide material for studying the freshwater adaptability of *C. nasus* and accumulate basic data for further studies on the population genetics of *C. nasus*.

Key words: freshwater resident; *Coilia nasus*; gap-free genome; gene annotation; collinear analysis

Corresponding author: LIU Kai. E-mail: liuk@ffrc.cn