

## 七种鲟鱼基因组草图的测序和分析\*

黄沁怡<sup>1,2</sup>, 刘洁<sup>1,2</sup>, 孔玲玉<sup>1,2</sup>, 周艳<sup>1,2</sup>, 齐茜<sup>3</sup>, 陆颖<sup>1,2</sup>

(1. 上海海洋大学水产种质资源发掘与利用教育部重点实验室, 上海 201306; 2. 上海海洋大学农业农村部淡水水产种质资源重点实验室, 上海 201306; 3. 河南科技大学动物科技学院, 河南 洛阳 471023)

**摘要:** 鲟鱼是全球现存鱼类中最为古老的种类之一, 全基因组复制(Whole genome duplication, WGD)事件在鲟形目祖先中多次发生, 导致鲟科物种普遍具有多倍性的复杂基因组, 这使得鲟鱼成为研究基因组多倍化的理想模型。本研究使用流式细胞仪测定闪光鲟(*Acipenser stellatus*)、小体鲟(*A. ruthenus*)和欧鳊(*Huso huso*)3个四倍体鲟鱼的基因组大小分别为2.30、2.16和2.15 Gb, 达氏鳊(*H. dauricus*)、施氏鲟(*A. schrenckii*)、俄罗斯鲟(*A. gueldenstaedtii*)和中华鲟(*A. sinensis*)4个八倍体鲟鱼的基因组大小分别为4.36、4.56、4.51和4.33 Gb。使用Illumina高通量测序对7种鲟鱼进行了全基因组survey测序和组装。基于这些鲟鱼survey基因组进行的系统发育分析表明, 鲟科起源于3亿年前, 在约1.55亿年前分化为大西洋和太平洋两个分支, 不同地理分布的鲟鱼基因组中, 免疫相关的基因都发生了快速扩增。基于不同物种多个同源基因和携带杂合单核苷酸多态性(Single nucleotide polymorphism, SNP)的同一基因编码序列的同义替换率计算结果证明, 鲟鱼特有的第三轮WGD发生于1.18亿~1.96亿年前。本研究为鲟科基因组研究提供了重要的基础数据, 为鲟科鱼类的起源与进化提供了新的研究视角。

**关键词:** 鲟鱼; 基因组测序; 比较基因组学; 全基因组复制; 多倍体

**中图分类号:** Q951 **文献标志码:** A

鲟鱼属于辐鳍鱼纲(Actinopterygii)、软骨硬鳞亚纲(Chondrostei)、鲟形目(Acipenseriformes), 是世界上最古老的鱼类之一, 被誉为“水中的活化石”, 有着长达2亿年的进化历史<sup>[1]</sup>。所有现存的鲟鱼物种均被列入世界自然保护联盟(International union for conservation of nature, IUCN)的“濒危物种红色名录”<sup>[2]</sup>。鲟鱼主要分布在北半球欧亚大陆和北美的流域中, 是洄游性冷水鱼, 体型大、寿命长、性成熟较晚<sup>[3-4]</sup>。鲟鱼因其极高的经济价值得到了广泛的人工养殖, 除了鱼卵可以被加工为名贵的鱼子酱外, 还被用于医疗保健品、化妆品和皮革生产等增值领域<sup>[5]</sup>。

近年来, 有关鲟鱼的分子细胞学研究证明了其具有多种不同倍性的染色体核型<sup>[6]</sup>。这些研究结合核型分析<sup>[7-8]</sup>、流式细胞术<sup>[9-10]</sup>、微卫星分析<sup>[10-11]</sup>和保守基因家族<sup>[12-13]</sup>等证据, 确定了鲟鱼的染色体基数约为120、250和360条<sup>[10, 14-15]</sup>, 并分别被定义为四倍体(功能性二倍体)、八倍体(功能性四倍体)和十二倍体(功能性六倍体)。其中十二倍体核型的鲟鱼最少, 仅见于短吻鲟(*Acipenser brevirostrum*)和达氏鲟(*A. dabryanus*)<sup>[16]</sup>两种, 四倍体和八倍体是鲟鱼中常见的染色体核型, 其中四倍体鲟鱼主要包括小体鲟(*A. ruthenus*,  $2n=4x=118\pm 4$ )<sup>[17]</sup>、闪光鲟(*A. stellatus*)、欧鳊(*Huso huso*,  $2n=4x=$

\* 收稿日期: 2025-04-24; 修回日期: 2025-05-07

基金项目: 河南省自然科学基金项目(232300420201)

作者简介: 黄沁怡(2000—), 女, 上海人, 2022级硕士研究生。E-mail: 787335365@qq.com

通信作者: 齐茜(1983—), 女, 河南洛阳人, 副教授, 博士。主要从事鱼类育种研究。E-mail: qiqian@haust.edu.cn

陆颖(1973—), 男, 上海人, 研究员, 博士。主要从事动植物基因组学及分子遗传育种研究。E-mail: yinglu@shou.edu.cn

118±2)<sup>[18]</sup>等;八倍体鲟鱼主要包括施氏鲟(*A. schrenckii*)、中华鲟(*A. sinensis*,  $2n=8x=264$ )<sup>[19]</sup>、俄罗斯鲟(*A. gueldenstaedtii*,  $2n=8x=250±8$ )<sup>[18]</sup>和达氏鳇(*H. dauricus*,  $2n=8x=268±4$ )<sup>[7]</sup>等。现存的26种多倍体鲟鱼起源于已经灭绝的鲟鱼二倍体祖先( $2n=60$ )<sup>[20]</sup>,鲟鱼丰富的染色体倍性使其成为研究鱼类全基因组复制(Whole genome duplication, WGD)和远缘杂交的理想材料。除此之外,鲟鱼的细胞核内有大量的微小染色体,这些微小染色体的数目几乎占总染色体数目的一半<sup>[8]</sup>。

全基因组复制事件是生物适应性进化的重要生物学过程。脊椎动物历史上共发生了两轮WGD,称为1R和2R。在WGD之后,会产生大量冗余的基因,这些多余的基因拷贝不受选择压力的显著影响,可以产生基因功能的分化,进而导致物种的不断进化,因此,WGD是造成物种多样性增加和趋异性进化的重要因素<sup>[21]</sup>。在辐鳍鱼纲中,自真骨鱼类与以鳄雀鳝(*Atractosteus spatula*)为代表的全骨鱼类分离后,它们在大约3.2亿至4亿年前发生了硬骨鱼特有的第三轮WGD(Ts3R)<sup>[22]</sup>。在Ts3R之前,鲟形目的祖先就已经从硬骨鱼中分化出来。与鲤科鱼类特有的第四轮WGD(Cr4R)<sup>[23-24]</sup>和单独发生在鲑科鱼类中的第四轮WGD(Ss4R)<sup>[22]</sup>相似,鲟科鱼类中也发生了自己特异的第三轮(或第四轮)WGD。

鲟鱼多倍化且复杂的基因组结构给基因组组装造成很大困难。目前已发表的染色体水平的鲟鱼高质量基因组较少,这限制了大规模的鲟形目基因组研究进程,导致对鲟鱼特有的第三轮WGD(As3R)发生时间的估算结果存在不少分歧,尚未形成统一结论<sup>[25-26]</sup>。除了一些线粒体基因组已经测序之外,仅有小体鲟<sup>[25]</sup>和中华鲟<sup>[26]</sup>等鲟鱼基因组发表。考虑到鲟鱼基因组结构的特点,本研究采用二代高通量测序,对7种典型的鲟鱼物种(3个四倍体和4个八倍体)进行了全基因组低深度测序(Survey sequencing)和初步组装,利用这些组装的基因组对鲟鱼物种的进化进行了初步分析。同时,通过基因家族的同义替换率计算和杂合SNP计算这两种不同的方法,基本确定了鲟鱼特有的第三轮WGD发生的时间。此外,通过对物种间发生快速扩增的基因进行功能注释,揭示了免疫相关基因的复制在鲟鱼环境适应性过程中的重要作用。

## 1 研究方法

### 1.1 样本采集与流式细胞测定

样本采集于我国河南省安阳市的某鲟鱼养殖场中,分别从鲟形目鲟科鲟属的施氏鲟(NCBI:txid111304)、俄罗斯鲟(NCBI:txid7902)、中华鲟(NCBI:txid61970)、闪光鲟(NCBI:txid7903)和小体鲟(NCBI:txid7906),以及鲟形目鲟科鳇属的达氏鳇(NCBI:txid55293)和欧鳇(NCBI:txid61971)共7个物种中选取了7尾3~5龄雌性鲟鱼,每尾鲟鱼抽取0.5 mL尾鳍静脉血,保存至EDTA-K2抗凝采血管中。将采集到的部分鲟鱼尾鳍血液作为血液样本进行流式细胞术检测。以ECD(PE-TR)作为荧光染料,取60  $\mu$ L悬浮血液样品加入2 mL 0.01 mol/L的磷酸盐缓冲液PBS(成分为NaCl 9.0 g,  $\text{Na}_2\text{HPO}_4 \cdot 7\text{H}_2\text{O}$  0.795 g,  $\text{KH}_2\text{PO}_4$  0.144 g, ddH<sub>2</sub>O 1 000 mL, pH值为7.4), 300×g离心5 min(Mini-6KS,奥盛)分离白细胞和红细胞。弃上清,PBS重悬红细胞,每个样品稀释至 $10^6$ 个细胞/mL。向细胞悬液中加入0.2 mL ECD荧光染料,混匀后,4℃黑暗孵育20~30 min。用同样的方法来处理草鱼(*Ctenopharyngodon idella*)的红细胞。染色后的鲟鱼细胞和草鱼细胞样本按体积比1:1混合,使用流式细胞仪Cell Lab Quanta™ SC(Beckman Coulter)进行测定。每个样本不少于20 000个细胞进入FL3通道(橙色荧光范围:610~620 nm)。最后,以草鱼基因组(大小约为0.9 Gb<sup>[27]</sup>)作为参照,估计鲟鱼的基因组大小。

### 1.2 DNA提取和基因组测序

DNA提取根据血液/细胞/组织基因DNA提取试剂盒(TIANGEN, DP304)的操作手册进行,提取后的DNA按照标准Illumina测序文库构建方法,经片段化、纯化、末端修复、3'端加A、连接测序接头和片段大小筛选后,构建成平均长度为350 bp的测序文库,在Illumina HiSeq X Ten平台上进行双

端测序,生成长度为150 bp、平均深度约为100×的paired-end测序数据。使用fastp v0.23.2和Trimomatic v0.39软件对原始数据(Raw data)进行质控和过滤,去除读段中残留的5 bp以上的接头序列,过滤掉碱基N数量大于总长度10%、质量分数较低( $\leq 5$ )且碱基数量大于总长度50%的读段,最终获得高质量的读段(Clean reads)。

### 1.3 基因组的组装和注释

使用Jellyfish v2.3.0进行k-mer分析(k-mer设定值为51-mer),估计基因组大小和杂合度,结果在GenomeScope v2.0中可视化。用SOAP denovo v2.40软件分别在51-mer、61-mer、71-mer和81-mer长度下组装,生成contig和scaffold序列。通过比较组装序列N50长度和基因组总长度等参数,确定最佳组装结果用于后续分析。

组装后的survey基因组序列,挑选长度在500 bp以上的scaffold序列进行基因注释和分析,以减少因小片段中的组装错误而引起的分析误差。之后使用RepeatModeler v2.0.5构建了各物种自身的重复序列数据库。通过Rebase数据库(<https://www.girinst.org/rebase/>)进行同源重复序列的注释。每个基因组中重复序列的预测和屏蔽由RepeatMasker v4.1.5完成。Infernal v1.1.2软件预测非编码RNA,tRNAScan-SE v2.0软件预测tRNA基因,LTR harvest v1.6.5、LTR\_Finder v1.0.7和LTR\_retriver v2.9.4等软件预测和注释长末端逆转座子(Long terminal repeat retrotransposon, LTR-retrotransposon),TRF v4.10.0软件预测简单重复序列(Simple sequence repeat, SSR)。在随后的编码基因注释中,首先使用AUGUSTUS v3.3.1和FGENESH v7.2.2进行de novo基因结构预测。然后选取大西洋鲟(*A. oxyrinchus*, GCA\_030684275.1)、小体鲟(GCF\_010645085.1)、斑马鱼(*Danio rerio*, GCF\_000002035.6)、斑点雀鳢(*Lepisosteus oculatus*, GCF\_000242695.1)和美国匙吻鲟(*Polyodon spathula*, GCF\_017654505.1)5个近缘物种的氨基酸序列,运用Genewise v2.4.1软件确定最佳匹配的基因结构。接着使用EVIDENCEModeler v2.1.0软件整合上述两种方法的注释结果,获得各鲟鱼物种非冗余的最终基因集。最后采用InterProScan v5.52软件,参考InterPro(<http://www.ebi.ac.uk/interpro/>)、KOG(<https://ftp.ncbi.nih.gov/pub/COG/KOG/kyva>)和Uniprot(<https://www.uniprot.org>)等数据库对基因进行功能注释。

### 1.4 进化分析

将已经发表的四倍体小体鲟基因组作为参考基因组,使用OrthoFinder v2.2.7,以斑点雀鳢作为外类群建立了7种鲟鱼的同源基因簇。鉴于鲟鱼基因组的多倍体特征,选择7种鲟鱼间共有的低拷贝基因簇(不同物种中拷贝数小于5的基因簇),使用ClustalW v2.0.12软件进行多序列比较,再使用IQ-TREE v1.6.12构建了每个基因家族的系统发育树,均以斑点雀鳢同源基因作为外类群。最后在FigTree v1.4.4中进行可视化,以每个物种中最靠近进化树根部的拷贝作为该物种的直系同源基因。使用PAL2NAL v14将基因的氨基酸序列转化成相应的核苷酸编码序列后,将每个物种中所有比对整齐后的序列连接成一个超级DNA矩阵(DNA supermatrix)。再用IQ-TREE和RAxML v8.2.12软件基于最大似然法构建物种系统发育树。使用TimeTree(<http://www.timetree.org/>)数据库中的小体鲟~欧鲤(9.47~86.40 Ma)、中华鲟~施氏鲟(9.47~86.40 Ma)以及斑点雀鳢~鲟亚科(493.8~652.0 Ma)这3个化石时间点校正进化树上各节点的分化时间。之后使用CAFE v5.1.0估算了7种鲟鱼的基因家族扩张与收缩情况,并使用EggNOG对每种鲟鱼相关的扩张和收缩基因进行了功能注释。最后使用KOBAS v3.0对鲟鱼中快速膨胀的基因家族进行了KEGG(Kyoto encyclopedia of genes and genomes)代谢途径的富集分析。

目前还缺乏高质量的鲟鱼基因组,为了估计鲟亚科谱系中WGD发生的时间,采用WGDdetector pipeline<sup>[28]</sup>计算流程,基于OrthoFinder生成同源基因,计算同源基因间的同义替换率( $K_s$ ),根据已知的鲟形目进化速率 $r=1.9 \times 10^{-4}$  subs·site<sup>-1</sup>·Ma<sup>-1</sup><sup>[29]</sup>计算分化时间。此外,利用携带杂合单核苷酸多态性(SNPs)的编码序列(Coding DNA sequence, CDS)计算基因分化时间,具体流程如下:使用

BWA v0.7.16软件将7种鲟鱼测序reads比对到各自的参考基因组上,用Genome Analysis Toolkit v4.3.0鉴定SNP,再根据SNP的位置和碱基多态性,在参考基因的编码序列上进行碱基替换,替换后的CDS与基因的原CDS进行序列比较后,计算这两个基因间的同义替换率( $K_s$ ),最后根据 $K_s$ 值计算替换后的基因与原基因的分化时间,进而估计整个基因组的WGD时间。

## 2 结果

### 2.1 基因组大小测定和survey测序

流式细胞测定的结果显示(表1),4种八倍体鲟鱼如达氏鳇、施氏鲟、俄罗斯鲟和中华鲟的基因组大小范围约为4.3~4.6 Gb,其中施氏鲟的基因组最大;3种四倍体鲟鱼如闪光鲟、小体鲟和欧鳇的基因组大小范围约为2.1~2.3 Gb,其中欧鳇的基因组最小。根据流式细胞仪测定结果估计的基因组大小,对7种鲟鱼的基因组进行二代高通量测序,共生成数据2.27 Tb。其中,达氏鳇338.6 Gb(测序深度为77.7 $\times$ )、施氏鲟414.7 Gb(90.9 $\times$ )、俄罗斯鲟415.7 Gb(92.1 $\times$ )、中华鲟280.2 Gb(64.7 $\times$ )、闪光鲟273.4 Gb(118.9 $\times$ )、小体鲟257.6 Gb(119.3 $\times$ )和欧鳇290.2 Gb(135.0 $\times$ )。

表1 流式细胞仪测定的DNA含量和估计的基因组大小

Table 1 DNA contents and genome sizes determined by a flow cytometry analysis

物种	峰的位置(相对荧光强度)		鲟鱼峰与草鱼峰的位置比值	基因组大小/Gb
	鲟鱼	草鱼		
达氏鳇	4 495 194	927 810	4.84	4.36
施氏鲟	4 307 507	849 470	5.07	4.56
俄罗斯鲟	4 255 096	849 751	5.01	4.51
中华鲟	4 265 734	887 512	4.81	4.33
闪光鲟	1 988 451	777 642	2.56	2.30
小体鲟	2 146 105	894 208	2.40	2.16
欧鳇	2 150 404	898 019	2.39	2.15

从测序数据的 $k$ -mer( $k=51$ )频率分布曲线(图1)来看,八倍体鲟鱼的 $k$ -mer分布曲线除了主峰以外还会呈现一个次峰,而四倍体鲟鱼的曲线则更接近泊松分布。大多数鲟鱼的 $k$ -mer曲线下方都存在起伏拖尾的现象,表明基因组中重复序列的比例较高。 $k$ -mer分布曲线预测的基因组大小大多与测量值相差很大,说明这一方法不适合多倍体复杂基因组大小的估计。从预测的基因组杂合度来看,八倍体鲟鱼基因组的平均杂合度(0.80%~1.06%)明显高于四倍体(0.32%~0.39%),但存在例外,如:八倍体的施氏鲟基因组的杂合度仅为0.37%,接近四倍体的水平;而四倍体小体鲟基因组的杂合度达0.7%,与多数八倍体基因组的杂合度相当。杂合度估计的结果与先前发表的组装结果接近<sup>[12,26]</sup>。在 $k$ -mer频率的分布图中出现的双峰和曲线拐点,可能是由于多倍体不同亚基因组之间的差异和染色体较高的杂合度造成的。在八倍体中会观察到一个拐点和一个主次峰,或两个拐点的情况,这暗示着八倍体基因组中至少发生过两次染色体的加倍过程。

根据组装的基因组大小与流式细胞仪结果的一致性,并综合评估N50、N90、测序深度(Coverage)和GC含量等指标,最终确定各survey基因组的组装结果(表2)。从组装结果来看,八倍体鲟鱼的survey基因组平均大小为5 Gb,而四倍体基因组平均大小为2.5 Gb,明显高于流式细胞仪测定的值,这可能是因为二代组装质量受鲟鱼基因组复杂结构的影响,整体上碎片化,这些未组装成contig或scaffold的片段使得组装后的结果远高于其基因组的实际值。由于多倍体基因组的复杂性和基因组随机测序短读长的影响,这些survey基因组的scaffold N50长度相对较短。欧鳇的scaffold N50长度

最长(657 bp),其次是俄罗斯鲟(375 bp),其他物种约为 180 bp。序列的覆盖度为 58~137×。其中,欧鲤基因组的 N50 长度和覆盖度较好,而欧鲤基因组的大小最接近流式细胞仪测定的基因组大小实验值。除了已经发表过的小体鲟基因组,其余 6 种鲟鱼 survey 基因组组装序列已经上传至国家生物信息中心(<https://ngdc.cncb.ac.cn/gsub/submit/gsa/>,数据登录号:GWHFWG00000000.1)。

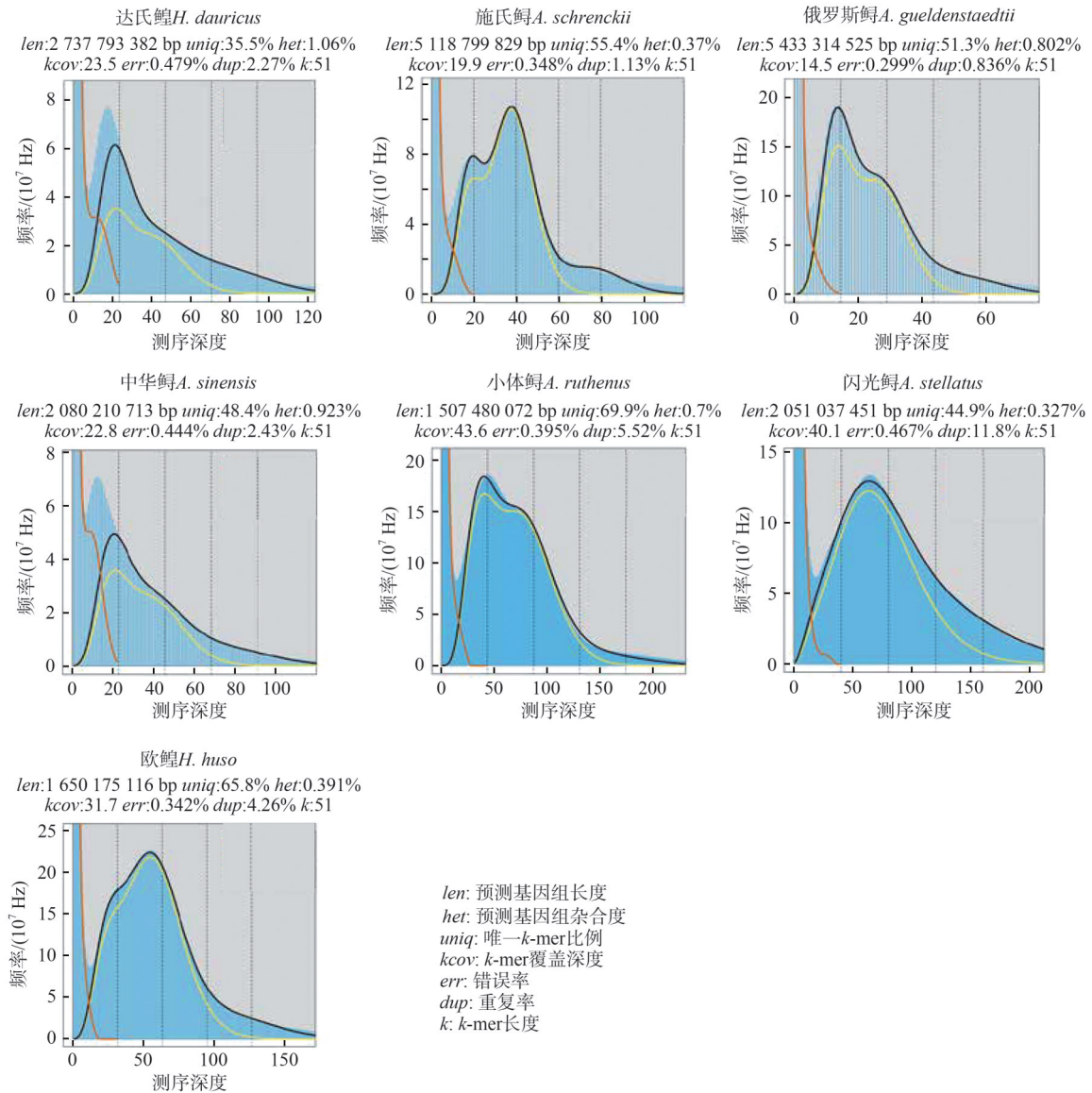


图 1 鲟鱼基因组  $k$ -mer ( $k=51$ ) 的频率分布

Fig. 1 Frequency distribution of  $k$ -mer ( $k=51$ ) occurrence in sturgeon genomes

## 2.2 基因组重复序列的预测及注释

重复序列注释结果显示,7种鲟鱼检测到的重复序列的总长度为 216.9(闪光鲟)~1424.2 Mb(俄罗斯鲟),占整个基因组的比例为 27.90%~39.94%;转座元件占全部重复序列的 40% 以上,其中,DNA 转座子的长度占比最高,为 17.26%(中华鲟)~26.67%(俄罗斯鲟)(表 3)。鲟亚科基因组中最丰富的 DNA 转座子家族是 Tc1/Mariner 超家族,平均有 265 896 个拷贝,占有所有已识别 DNA 转座子的 5.7%。在各种逆转座子中,LINE/L2、LTR/DIRS、LINE/CR1、DNA/TMar-Tc1 和 DNA/hAT Charlie 是鲟鱼共有的 5 种长度占比最高的转座元件。

表2 鲟鱼 survey 基因组的组装结果

Table 2 Assembly results of the sturgeon survey genomes

物种	组装用的 <i>k</i> -mer/bp	基因组 大小/Gb	最长 scaffold/bp	scaffold N50 长度/bp	≥500 bp 的 scaffold 数 量/个	≥500 bp 的 scaffold 比例/%	平均测序 深度/×	GC 含量/%	重复 序列 比例/%
达氏鳊	61	5.33	64935	183	1367403	5.43	61.9	42.01	38.26
施氏鲟	61	4.33	19964	145	929317	3.74	93.7	40.87	36.78
俄罗斯鲟	61	7.67	106471	375	2472633	8.81	124.4	40.15	39.94
中华鲟	71	3.95	29214	196	1113493	5.85	59.5	39.60	31.50
闪光鲟	51	2.71	9361	160	734083	4.88	134.1	40.40	27.90
小体鲟	51	2.66	8027	147	666721	4.32	58.4	39.78	28.52
欧鳊	61	2.22	63980	657	543096	7.53	90.9	39.87	35.44

表3 各种转座元件的长度比例

Table 3 Length proportions of various transposable elements

物种	DNA 转座子/%	逆转座子/%				未知转座子/%
		LTR	SINE	LINE	其他	
欧鳊	19.57	7.26	6.46	9.63	8.24	48.84
闪光鲟	17.94	6.62	6.76	9.50	8.77	50.41
小体鲟	17.93	6.77	6.57	9.53	8.51	50.70
俄罗斯鲟	26.67	5.12	2.13	9.66	5.84	50.58
中华鲟	17.26	7.70	6.70	10.99	8.69	48.65
施氏鲟	18.18	8.66	5.68	9.65	7.76	50.07
达氏鳊	20.46	6.95	6.81	11.92	7.66	46.20

注:LINE,长散布核元件逆转座子(Long interspersed nuclear element retrotransposon);SINE,短散布核元件逆转座子(Short interspersed nuclear element retrotransposon);LTR,长末端重复反转录座子(Long terminal repeat retrotransposons)。

### 2.3 系统发育分析及物种分化时间推测

基因家族聚类的结果显示,共7330个基因家族(占有所有基因家族的17.2%)为所有鲟鱼共有(图2),仅有26个一对一单拷贝基因家族。这反映了鲟鱼基因组多倍体的特点,也表明目前普遍采用的通过一对一单拷贝基因家族构建物种进化关系的方法,对鲟鱼这种多倍体基因组的物种来说是不适用的。我们选取了304个低拷贝的、各物种共有的同源基因(具体方法详见1.4节)串联成基因矩阵(长度为14822 bp),构建了7种鲟鱼的物种系统发育树。系统发育树的分析结果显示,鲟科祖先起源于3.03亿年前,在约1.55亿年前分化为两个地理分支——大西洋分支和太平洋分支,与鲟鱼的地理起源一致(图3)。物种进化关系也说明鲟属并非单系起源<sup>[30]</sup>,鲟属的物种分别定位于上述两个不同的地理分支中。在太平洋分支中,中华鲟与施氏鲟的亲缘关系较近,与达氏鳊较远;在大西洋分支中,小体鲟、欧鳊、闪光鲟的亲缘关系相近,与俄罗斯鲟较远。

在进行了系统进化和各基因家族中基因拷贝数的比较后,共鉴定出627个快速膨胀的基因家族。使用KEGG代谢途径数据库对这些基因进行功能注释和富集分析后发现,如图4所示,这些基因主要富集于如下通路:1)与免疫相关的代谢通路:Toll受体信号通路(Toll-like receptor signaling pathway)、补体和凝血级联反应(Complement and coagulation cascades)、中性粒细胞外陷阱形成(Neutrophil extracellular

trap formation)、造血细胞谱系(Hematopoietic cell lineage)、病毒蛋白与细胞因子和细胞因子受体的相互作用(Viral protein interaction with cytokine and cytokine receptor);2)细胞代谢相关通路:嘌呤代谢(Purine metabolism)和嘧啶代谢(Pyrimidine metabolism);3)信号转导相关通路:cAMP信号通路(cAMP signaling pathway)和cGMP-PKG信号通路(cGMP-PKG signaling pathway);4)代谢和生理活动相关通路:胰腺分泌(Pancreatic secretion)、血管平滑肌收缩(Vascular smooth muscle contraction)、破骨细胞分化(Osteoclast differentiation)、肾素分泌(Renin secretion)和Apelin信号通路(Apelin signaling pathway)等。在这些有功能注释信息的基因中,补体与凝血级联反应途径涉及的基因最多,涉及了21个小体鲟基因、21个达氏鳇基因和81个俄罗斯鲟基因,说明免疫基因的大量复制在鲟鱼环境适应性过程中起着重要作用。

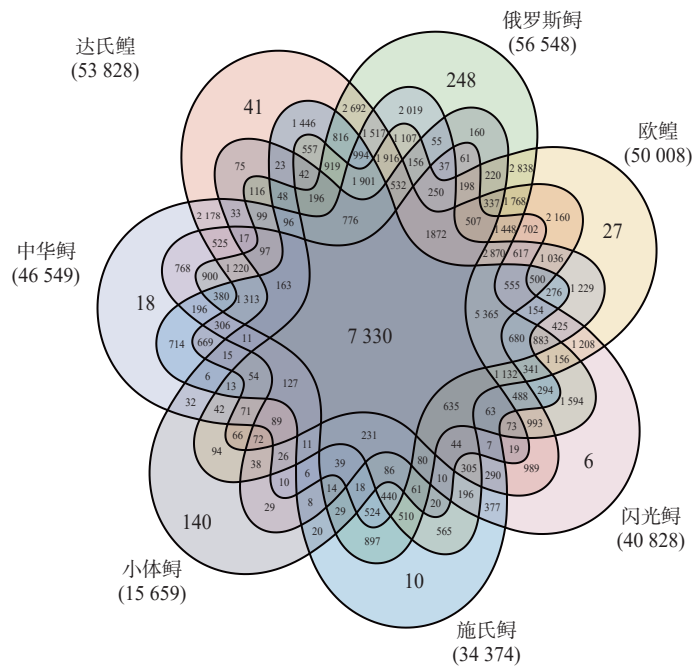


图 2 共有和特有基因家族数量的韦恩图

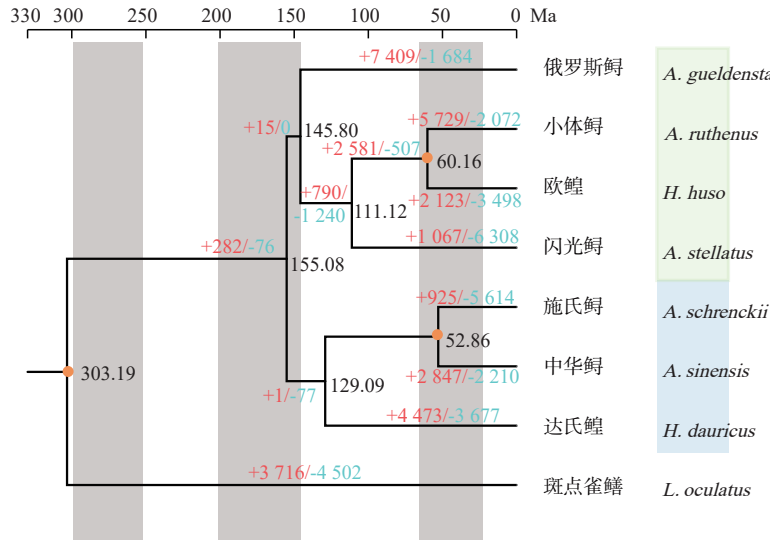
Fig. 2 Venn diagram of the number of shared and specific gene families

### 2.4 全基因组复制发生时间的计算

先前的研究基本明确了鲟科具有额外的第三轮WGD,即As3R<sup>[25]</sup>。但对As3R发生的时间还有较大争论,特别是对于八倍体基因组复制发生的时间,因缺少八倍体基因组的数据,相关研究尚未系统性开展。本研究采用两种方法计算基因间的同义替换率( $K_s$ ),估计基因组复制发生的时间。

首先,挑选鲟鱼基因家族中的双拷贝基因家族,即每个物种中有且仅有两个拷贝的基因家族,计算这两个拷贝之间的同义替换率( $K_s$ )来估计WGD发生的时间。之所以挑选双拷贝的基因家族,是因为在这些基因家族中只发生过一轮确定的基因复制事件。而具有3个或3个以上拷贝的基因家族,基因复制事件的发生过程比较复杂,计算过程中会涉及部分发生复制基因的祖先基因确定问题,因而会引入较大的误差。从获得的各种鲟鱼物种中双拷贝基因家族 $K_s$ 值的分布曲线上看,四倍体和八倍体都仅发现一个峰,主要集中于0.024至0.074,中位数为0.045(图5)。

同时,将7个鲟鱼基因组测序的数据比对到各自的survey基因组上,鉴定出位于基因编码区的杂合SNP位点,然后根据SNP的信息,通过替换碱基的办法,获得携带杂合SNP的基因信息,与基因组上的原基因编码序列组合,计算出它们之间的 $K_s$ 值。发现峰值出现在小体鲟的0.06~0.07、闪光鲟的0.06~0.07、欧鳇的0.04~0.05、中华鲟的0.06~0.07、施氏鲟的0.06~0.07、达氏鳇的0.06~0.07和俄罗斯鲟的0.06~0.07等区域。



橙色节点表示分化时间经过化石校正过的节点。每个节点边的黑色数字表示物种分化时间。树枝上的正数(红色)表示该物种发生快速扩张的基因数量,负数(蓝色)表示快速收缩的基因数量。

图3 鲟鱼物种的系统发育树  
Fig. 3 Phylogenetic tree of the sturgeon species

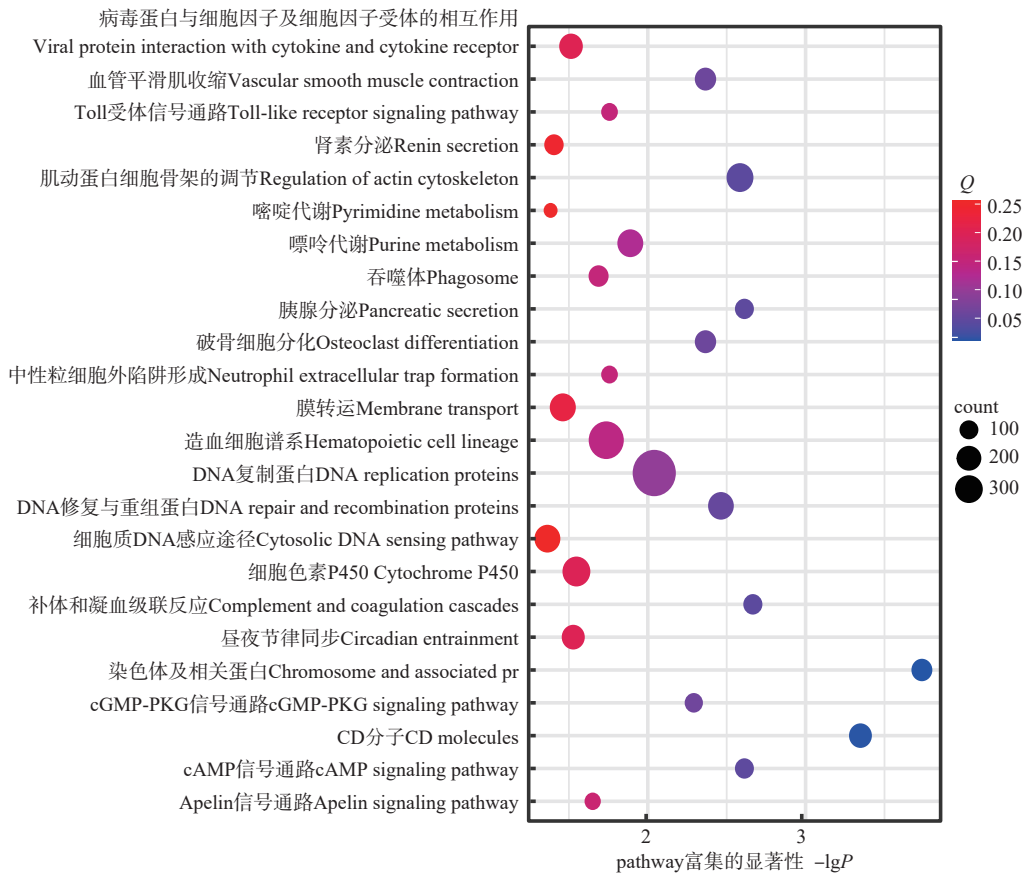
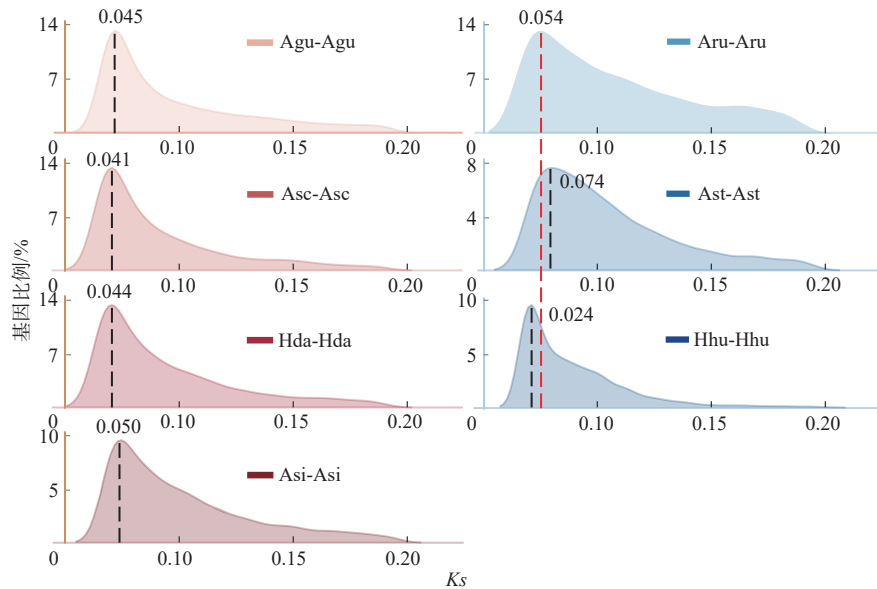


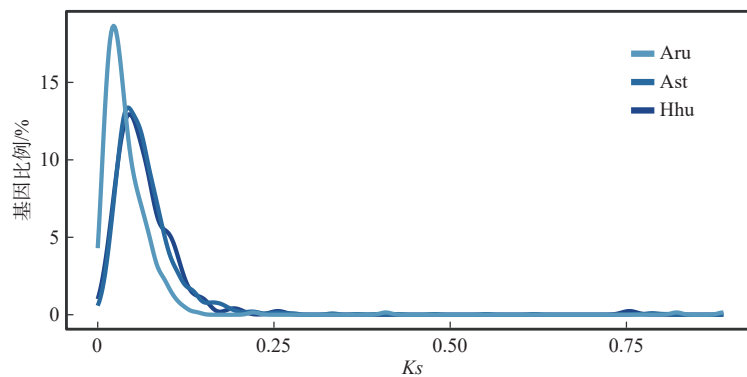
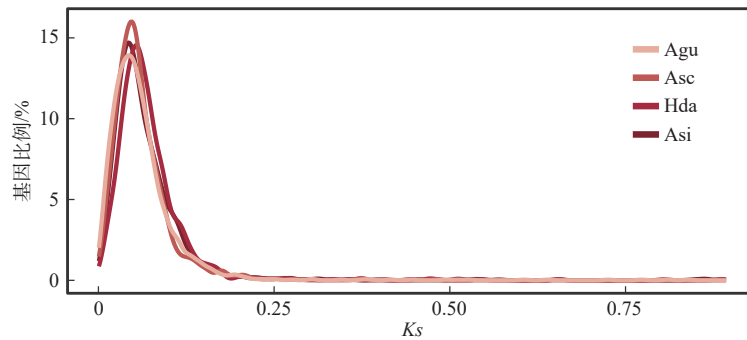
图4 发生迅速扩张的基因家族的KEGG代谢途径富集分析  
Fig. 4 Enrichment analysis of KEGG metabolic pathways in rapidly expanded gene families

这两组预测数值比较接近,在去除离群值后,主要分布于0.045~0.070的范围。根据鲟形目物种的进化速率(每百万年每个位点发生  $1.9 \times 10^{-4}$  个突变)<sup>[29]</sup>,估计As3R发生的时间介于118~196 Ma

之间。我们计算得出的时间十分接近文献[25-26]中报道的 As3R 发生时间,值得一提的是,该时间点与发生在 2 亿年前的二叠纪至三叠纪灭绝事件(Permian-Triassic extinction event)时间点十分接近,暗示了大型地质灾害事件对生物进化的重要影响。



(a) 双拷贝基因家族中, 具有不同同义替换率 ( $K_s$ ) 的基因数量比例



(b) 携带杂合 SNP 基因的同义替换率 ( $K_s$ ) 分布

Agu, 俄罗斯鲟; Asc, 施氏鲟; Hda, 达氏鲤; Asi, 中华鲟; Aru, 小体鲟; Ast, 闪光鲟; Hhu, 欧鲤。

图 5 基于基因同义替换率分布估计的全基因组复制发生时间

Fig. 5 Estimation of the WGD occurrence time based on the distribution of genes' synonymous substitution rates

### 2.5 鲟鱼基因组的种间变异分析

以小体鲟基因组<sup>[25]</sup>为参考基因组,使用 BWA 软件将 7 种鲟鱼物种的原始测序数据映射到参考基因组上,然后使用 GATK 软件完成 SNP 的鉴定。从测序序列与参考基因组比对的结果来看,发现

6种鲟鱼序列与小体鲟基因之间的比对率都达到了90%以上,不同种属个体之间的基因组相似性较高,表明鲟科基因组具有相近的起源。具体来看,太平洋鲟鱼物种的施氏鲟、中华鲟和达氏鳊与小体鲟的基因组比对率分别为98.16%、97.53%和95.30%,低于大西洋鲟鱼物种的闪光鲟(98.85%)、欧鳊(98.87%)和俄罗斯鲟(98.46%),说明大西洋分支和太平洋分支在分化后各自发生不同的进化过程,这与物种系统进化分析的结果基本一致。从鉴定出来的SNP结果来看,分别在达氏鳊、中华鲟、俄罗斯鲟、施氏鲟、闪光鲟、欧鳊、小体鲟中检测到23972516、33230741、44171485、34268694、15713058、12458254、7410054个SNP,这7个物种的SNP密度见图6。四倍体鲟鱼中检测到的平均SNP频率(如欧鳊的0.0068和闪光鲟的0.0086)明显低于八倍体(如俄罗斯鲟的0.024、中华鲟的0.018和施氏鲟的0.019),表明在染色体多次加倍以后,基因组中由于不同亚基因组的存在,在染色体的同源区域中累积了大量SNPs。用SNPeff软件对变异区域的位置进行注释后显示,36.12%的变异位于基因组的内含子区域(Intron),12.19%的变异位于基因间隔区(Intergenic),1.26%的变异在外显子区(Exon),38.50%的变异发生在不翻译的转录本区(Transcript),10.18%的变异位于基因的上下游区域,仅1.75%的变异发生在非翻译区域(UTR)以及其剪切位点区(Splicing sites)。从外显子区域发生变异的基因注释结果来看,这些基因主要参与了免疫调控、基因表达调控和细胞结构与信号传导等生物学过程,与发生快速膨胀基因家族的注释结果相似。说明这些变异的发生伴随着基因的扩张,而鲟鱼基因扩张的主要原因是基因组加倍,这从侧面反映了鲟鱼基因组的多倍化在基因进化过程中的关键作用。

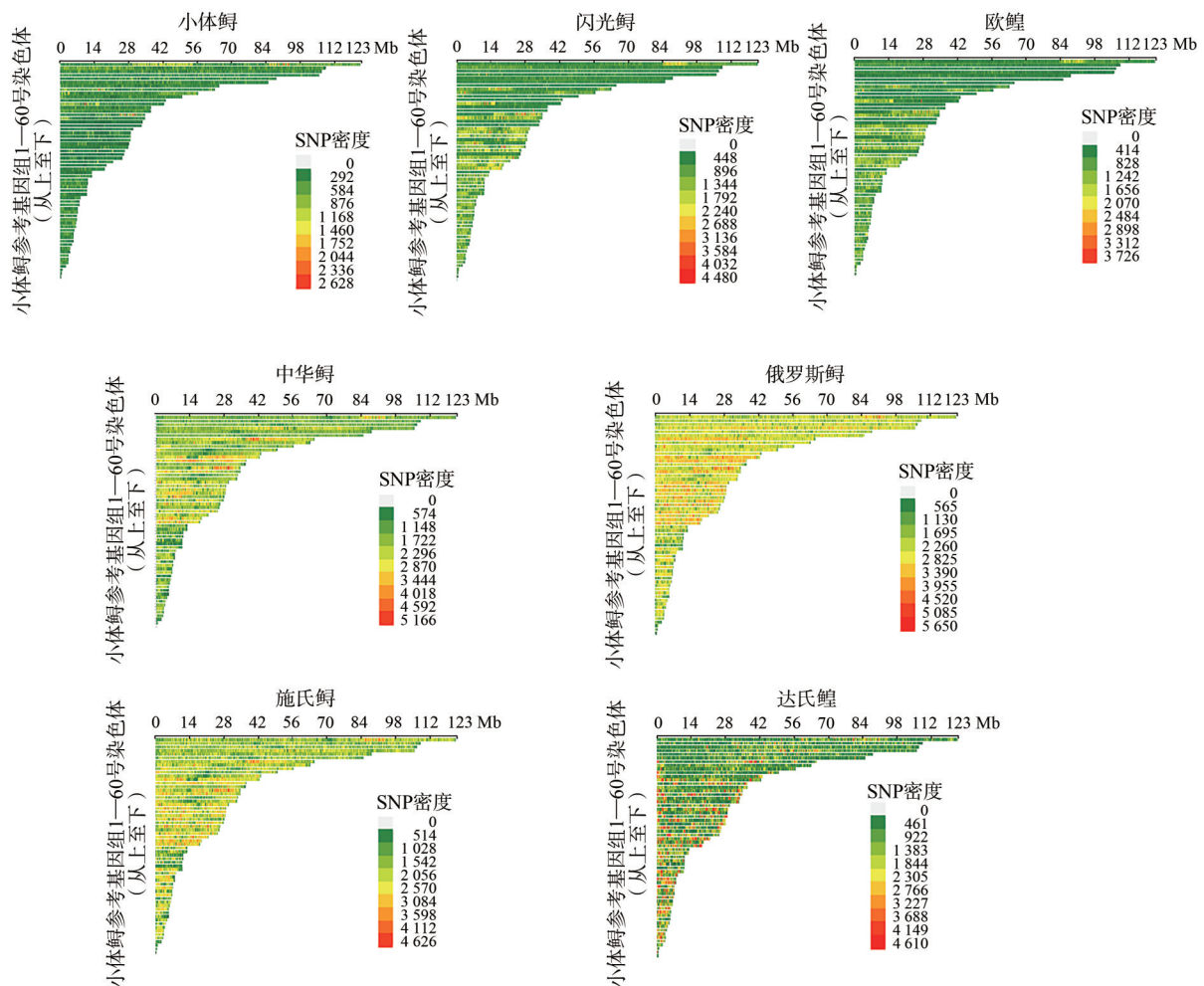


图6 SNP密度

Fig. 6 Density of SNPs

### 3 讨论

#### 3.1 鲟鱼基因组的结构反映了不同的适应性进化过程

本研究采用二代 Illumina 测序平台测序和组装了7种鲟鱼的基因组。在组装过程中,发现八倍体鲟鱼基因组平均大小为5 Gb,四倍体基因组平均大小为2.5 Gb,与流式细胞仪的测定结果相似。鲟鱼基因组具有较高的杂合度(0.37%~1.06%)和较高的GC含量(39.60%~42.01%)。鲟鱼的整体重复序列比例为27.90%~39.94%,7种鲟鱼的重复序列比例略低于先前已发表的鲟鱼基因组,与七鳃鳗(*Petromyzon marinus*, 39.1%)和象鲨(*Callorhinchus milii*, 42.7%)<sup>[31]</sup>等原始鱼类相似。用二代测序组装的 survey 基因,含有大量碎小的片段(<500 bp),容易丢失较多的重复序列信息,并在基因组中引入较多的组装空隙,这也是组装后的基因组中重复序列的比例被低估的原因。此外,重复序列比例与基因组的倍性也成正比,四倍体的平均重复序列比例(31%)低于八倍体(36%)。除俄罗斯鲟外,其余6种鲟鱼转座子的比例模式也和已发表的鲟鱼基因组中的对应比例一致<sup>[12,32]</sup>,并且与一些其他古老的鱼类的重复序列分布相似<sup>[33]</sup>。这些结果表明,作为“活化石”的鲟鱼与其他古老的鱼类具有相似的基因组组成。鲟亚科基因组中最丰富的DNA转座子家族是Tc1/Mariner超家族,平均有265896个拷贝,占有已识别DNA转座子的5.7%。Tc1/Mariner的高含量似乎与鲟鱼的洄游特性相关,这类转座子的高含量也在美洲鳗鲡(*Anguilla rostrata*)和大西洋鲑(*Salmon salar*)的基因组中被发现<sup>[33]</sup>。

#### 3.2 核基因组分析对于揭示复杂基因组的进化机制起着关键作用

依赖线粒体基因组或分子标记构建的物种系统进化树,对于亲缘关系接近的近缘物种,存在分辨率不足的问题。本研究从全基因组层面,利用更广泛的核基因组的同源基因重新构建了鲟形目物种系统发育树。在此过程中,针对多倍体物种几乎不存在单拷贝基因的情况,创新性地采用各个低拷贝基因家族构建了物种的进化树,提高了物种进化树构建的准确性,而这个进化树的拓扑结构与先前大部分的研究结果一致。鲟鱼起源于约3亿年前,约在1.55亿年前开始分化成两个分支——大西洋分支和太平洋分支。根据生物地理学分析的结果,这可能与晚三叠世到晚侏罗世(约1.5亿~2亿年前)古大西洋和古太平洋逐渐分离,鲟鱼迁移至太平洋地区<sup>[34]</sup>有关。此外,来自我国北方的史前鲟鱼——潘氏北票鲟(*Peipiaosteus*)化石<sup>[35]</sup>证实了鲟形目可能在1.3亿年前从里海地区传播到东亚,这可以作为分化时间的证据。鲟鱼基因家族的膨胀和收缩分析表明,基因家族的收缩多发生于四倍体鲟鱼中,这暗示着四倍体有区别于八倍体的再二倍化进程。对快速扩张的基因进行KEGG富集分析结果表明,这些基因与鱼类的免疫、生长代谢、生理功能和细胞信号转导等功能有关。

本研究中估算鲟科特有的第三轮全基因组复制事件发生在118~196 Ma前,这个As3R事件的时间与2亿年前第三次大规模生物灭绝事件(P-Tr)时间相近。然而,由于survey基因组数据还比较碎片化,本研究没有发现能够代表As4R发生的Ks频率曲线峰值。Wang等<sup>[26]</sup>通过计算编码基因和单一假基因中的Ks值得到中华鲟As4R约发生在3500万年前,这可能是由于发生时间距离现代较近,在我们的survey基因组中没能检测到足够的、近期发生的同义突变来估计As4R发生的时间。

### 参考文献:

- [1] HILTON E J, GRANDE L, BEMIS W E. Skeletal anatomy of the shortnose sturgeon, *Acipenser brevirostrum* Lesueur, 1818, and the systematics of sturgeons (Acipenseriformes, Acipenseridae) [J]. Fieldiana Life and Earth Sciences, 2011, 3(1): 1-168.
- [2] POURHOSEIN S S, FALAHATKAR B. Tracking scientific gaps in the conservation of endangered sturgeon: A bibliometric approach and content analyses [J]. Aquaculture Research, 2024, 2024(1): 4496931.
- [3] HUNG S S O. Recent advances in sturgeon nutrition [J]. Animal Nutrition, 2017, 3(3): 191-204.
- [4] BEAMESDERFER R C P, FARR R A. Alternatives for the protection and restoration of sturgeons and their habitat

- [J]. Environmental Biology of Fishes, 1997, 48(1):407-417.
- [5] WEI Q W, ZOU Y, LI P, et al. Sturgeon aquaculture in China: Progress, strategies and prospects assessed on the basis of nation-wide surveys (2007—2009)[J]. Journal of Applied Ichthyology, 2011, 27(2):162-168.
- [6] LUDWIG A, BELFIORE N M, PITRA C, et al. Genome duplication events and functional reduction of ploidy levels in sturgeon (*Acipenser*, *Huso* and *Scaphirhynchus*)[J]. Genetics, 2001, 158(3):1203-1215.
- [7] VASIL'EV V P, VASIL'EVA E D, SHEDKO S V, et al. Ploidy levels in the Kaluga, *Huso dauricus* and Sakhalin sturgeon *Acipenser mikadoi*(*Acipenseridae*, *Pisces*)[J]. Doklady Biological Sciences, 2009, 426(1):228-231.
- [8] SYMONOVÁ R, FLAJŠHANS M, SEMBER A, et al. Molecular cytogenetics in artificial hybrid and highly polyploid sturgeons: An evolutionary story narrated by repetitive sequences[J]. Cytogenetic and Genome Research, 2013, 141(2/3):153-162.
- [9] ZHOU H, FUJIMOTO T, ADACHI S, et al. Genome size variation estimated by flow cytometry in *Acipenser mikadoi*, *Huso dauricus* in relation to other species of *Acipenseriformes*[J]. Journal of Applied Ichthyology, 2011, 27(2):484-491.
- [10] DRAUCH SCHREIER A, GILLE D, MAHARDJA B, et al. Neutral markers confirm the octoploid origin and reveal spontaneous autopolyploidy in white sturgeon, *Acipenser transmontanus*[J]. Journal of Applied Ichthyology, 2011, 27(S2):24-33.
- [11] BILTUEVA L S, PROKOPOV D Y, ROMANENKO S A, et al. Chromosome distribution of highly conserved tandemly arranged repetitive DNAs in the Siberian sturgeon(*Acipenser baerii*)[J]. Genes, 2020, 11(11):1375.
- [12] CHENG P L, HUANG Y, DU H, et al. Draft genome and complete *Hox*-cluster characterization of the sterlet (*Acipenser ruthenus*)[J]. Frontiers in Genetics, 2019, 10:776.
- [13] CROW K D, STADLER P F, LYNCH V J, et al. The "fish-specific" *Hox* cluster duplication is coincident with the origin of teleosts[J]. Molecular Biology and Evolution, 2005, 23(1):121-136.
- [14] BIRSTEIN V J, POLETAEV A I, GONCHAROV B F. DNA content in Eurasian sturgeon species determined by flow cytometry[J]. Cytometry, 1993, 14(4):377-383.
- [15] 张四明, 晏勇, 邓怀, 等. 几种鲟鱼基因组大小、倍体的特性及鲟形目细胞进化的探讨[J]. 动物学报, 1999, 45(2):200-206.
- [16] KIM D S, NAM Y K, NOH J K, et al. Karyotype of North American shortnose sturgeon *Acipenser brevirostrum* with the highest chromosome number in the *Acipenseriformes*[J]. Ichthyological Research, 2005, 52(1):94-97.
- [17] FONTANA F, LANFREDI M, CHICCA M, et al. Fluorescent in situ hybridization with rDNA probes on chromosomes of *Acipenser ruthenus* and *Acipenser naccarii*(*Osteichthyes Acipenseriformes*)[J]. Genome, 2011, 42(5):1008-1012.
- [18] BIRSTEIN V J, VASILIEV V P. Tetraploid-octoploid relationships and karyological evolution in the order *Acipenseriformes*(*pisces*) karyotypes, nucleoli, and nucleolus-organizer regions in four acipenserid species[J]. Genetica, 1987, 72(1):3-12.
- [19] ZHOU G Z, GUI L, LI Z Q, et al. Establishment of a Chinese sturgeon *Acipenser sinensis* tail-fin cell line and its susceptibility to frog iridovirus[J]. Journal of Fish Biology, 2008, 73(8):2058-2067.
- [20] 陈金平, 孙大江, 张树义, 等. 鲟鱼类系统进化的研究现状[J]. 四川动物, 2004, 23(4):374-379.
- [21] OHNO S. Evolution by gene duplication[M]. Berlin, New York: Springer, 1970.
- [22] BERTHELOT C, BRUNET F, CHALOPIN D, et al. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates[J]. Nature Communications, 2014, 5(1):3657.
- [23] XU P, XU J, LIU G J, et al. The allotetraploid origin and asymmetrical genome evolution of the common carp *Cyprinus carpio*[J]. Nature Communications, 2019, 10(1):4625.
- [24] LI J T, WANG Q, HUANG YANG M D, et al. Parallel subgenome structure and divergent expression evolution of allo-tetraploid common carp and goldfish[J]. Nature Genetics, 2021, 53(10):1493-1503.
- [25] DU K, STÖCK M, KNEITZ S, et al. The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization[J]. Nature Ecology & Evolution, 2020, 4(6):841-852.
- [26] WANG B Z, WU B, LIU X Q, et al. Whole-genome sequencing reveals Autooctoploidy in Chinese sturgeon and its evolutionary trajectories[J]. Genomics, Proteomics & Bioinformatics, 2023, 22(1):qzad002.
- [27] WANG Y P, LU Y, ZHANG Y, et al. The draft genome of the grass carp(*Ctenopharyngodon idellus*)provides insights into its evolution and vegetarian adaptation[J]. Nature Genetics, 2015, 47(6):625-631.
- [28] YANG Y Z, LI Y, CHEN Q, et al. WGDdetector: A pipeline for detecting whole genome duplication events using the genome or transcriptome annotations[J]. BMC Bioinformatics, 2019, 20(1):75.

- [29] BROWNSTEIN C D, MACGUIGAN D J, KIM D, et al. The genomic signatures of evolutionary stasis[J]. *Evolution*, 2024, 78(5):821-834.
- [30] CHENG P L, YU D, LIU H Z, et al. Molecular phylogeny of Acipenseriformes based on complete mitochondrial genome sequence[J]. *Acta Hydrobiologica Sinica*, 2021, 45(3):487-494.
- [31] CHALOPIN D, NAVILLE M, PLARD F, et al. Comparative analysis of transposable elements highlights mobile diversity and evolution in vertebrates[J]. *Genome Biology and Evolution*, 2015, 7(2):567-580.
- [32] CHENG P L, HUANG Y, LV Y Y, et al. The American paddlefish genome provides novel insights into chromosomal evolution and bone mineralization in early vertebrates[J]. *Molecular Biology and Evolution*, 2020, 38(4):1595-1607.
- [33] YUAN Z H, LIU S K, ZHOU T, et al. Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments[J]. *BMC Genomics*, 2018, 19(1):141.
- [34] SHEN Y, YANG N, LIU Z, et al. Phylogenetic perspective on the relationships and evolutionary history of the Acipenseriformes[J]. *Genomics*, 2020, 112(5):3511-3517.
- [35] GRANDE L, JIN F, YABUMOTO Y, et al. *Protopsephurus liui*, a well-preserved primitive paddlefish (Acipenseriformes: Polyodontidae) from the Lower Cretaceous of China[J]. *Journal of Vertebrate Paleontology*, 2002, 22(2):209-237.

(责任编辑 那顺布和)

## Survey Sequencing and Evolutionary Insights from the Genomes of Seven Sturgeon Species

HUANG Qinyi<sup>1,2</sup>, LIU Jie<sup>1,2</sup>, KONG Lingyu<sup>1,2</sup>, ZHOU Yan<sup>1,2</sup>, QI Qian<sup>3</sup>, LU Ying<sup>1,2</sup>

(1. *Key Laboratory of Exploration and Utilization of Aquatic Genetic Resources, Ministry of Education, Shanghai Ocean University, Shanghai 201306, China;*

2. *Key Laboratory of Freshwater Aquatic Genetic Resources, Ministry of Agriculture and Rural Affairs, Shanghai Ocean University, Shanghai 201306, China;*

3. *College of Animal Science and Technology, Henan University of Science and Technology, Luoyang 471023, China*)

**Abstract:** Sturgeons represent one of the most ancient species among the existing fish species. Multiple whole genome duplication (WGD) events have occurred in the ancestors of the Acipenseriformes, resulting in the widespread polyploidy and complex genomic architectures across sturgeon species. These characteristics make sturgeons an exceptional model for investigating genome polyploidization and vertebrate genome evolution. Flow cytometry was used to estimate the genome sizes of three tetraploid species, *Acipenser stellatus*, *Acipenser ruthenus*, and *Huso huso*, at 2.30, 2.16, and 2.15 Gb, as well as four octoploid species, *Huso dauricus*, *Acipenser schrenckii*, *Acipenser gueldenstaedtii*, and *Acipenser sinensis*, at 4.36, 4.56, 4.51, and 4.33 Gb. The genome survey sequencing and assembly were constructed on seven sturgeon species using Illumina high-throughput technology. Phylogenetic analysis based on the survey genomes suggests that the Acipenseridae originates approximately 300 million years ago, and subsequently diverges into Atlantic and Pacific lineages around 155 million years ago. Notably, a lineage-specific expansion of immune-related genes is detected across geographically distinct species, implying adaptive evolution. Furthermore, synonymous substitution rates ( $K_s$ ) are calculated both among orthologous genes across species and within heterozygous alleles of the same genes, supporting a lineage-specific third round of WGD in sturgeons occurring between 118 and 196 million years ago. This study provides fundamental genomic resources for the sturgeon family and sheds new light on the evolutionary history and polyploidization events of this ancient vertebrate lineage.

**Key words:** sturgeon; genome sequencing; comparative genomics; whole genome duplication; polyploid