瑞氏红鲃鲴(*Satyrichthys rieffeli*)基因组 survey 分析及线粒体基因组注释*

廖贤晖[1] 王乙婷[1] 瞿印权[1] 刘 启[2] 高天翔[1①]

(1. 浙江海洋大学水产学院 浙江舟山 316022; 2. 武汉万摩科技有限公司 湖北武汉 430076)

摘要 为了解瑞氏红鲃鲴(*Satyrichthys rieffeli*)的基因组特征及线粒体基因组结构, 采用二代高通量测序技术对瑞氏红鲃鲴进行基因组 survey 分析, 为研究其分子学内容提供基础信息。利用 Illumina nova 测序平台进行测序, 组装得到的基因组大小为 813 Mb, 杂合率为 0.92%, 重复序列比例为 45.97%, Contigs 的 N50 大小为 712 bp。瑞氏红鲃鲴线粒体基因组长度为 16 527 bp, 共 38 个基因(22 个 tRNA、*12S rRNA*、*16S rRNA*、*ND1~ND6*、*COXI~COXIII*、*Cyt b*、*ATP8*、*ATP6*、*ND4L* 和 1 个 D-loop 区), 基因之间未发生重排, GC 含量 45.70%; 蛋白质编码基因中出现不完整的密码子 T--和 TA-; tRNA 中只有 *tRNA-Ser* (GCT)缺失了二氢尿苷臂(DHU 臂)的简单环, 其他都是正常二级结构。此外, 联合 NCBI 中 18 个鲇形目鱼类的线粒体基因组, 利用最大似然法和贝叶斯法构建了鲇形目鱼类的分子系统发育关系。结果表明, 瑞氏红鲃鲴与同为黄鲃鲴科的须叉吻鲃鲴(*Scalicus amiscus*)为姐妹分支, 支持瑞氏红鲃鲴是黄鲃鲴科鱼类的形态学分类结果, 为开展黄鲃鲴科鱼类系统发育深入研究奠定了基础。

关键词 瑞氏红鲃鲴; 基因组 survey; 线粒体基因组; 系统发育关系

中图分类号 Q789; S965; Q953 doi: 10.11693/hyhz20230200042

高通量测序已成为一种广泛用于动植物基因组测序技术(孟刚等, 2021; 王子寅等, 2023)。通过高通量测序技术可以进行基因组 survey 研究, 包括基因组大小估计、重复序列比例和 GC 含量计算, 以及对遗传多样性、基因组结构和遗传改良的了解(Barchi *et al*, 2011; Rowe *et al*, 2011; Xu *et al*, 2014; Shi *et al*, 2018), 其结果可用于调查基因组图谱和提取线粒体基因组。对于缺乏基因组数据的非模式物种来说, 针对基因组的 survey 分析是分子机理研究和基因资源开发的前提(Kim *et al*, 2011)。

线粒体是半自主细胞性细胞器, 普遍存在于真核细胞内(Boore, 1999)。脊椎动物的线粒体 DNA (Mitochondrial DNA, mtDNA)基因结构紧凑, 长度一般在 15~20 kb 之间, 并且编码的基因库非常保守(张方等, 1998)。脊椎动物的线粒体基因组中一般包括了 22 个 tRNA 基因、2 个 rRNA 基因、13 个蛋白质编码区(PCGs)、1 个非编码区包括 D-loop 区和 1 个轻链复制起始区(O_L) (Satoh *et al*, 2016)。线粒体基因组具有结构简单、编码区域高度保守、母系遗传、进化速率快、拷贝数高等鲜明特征(Ruan *et al*, 2020), 被广泛应用于群体遗传学、进化生物学和谱系遗传学等方面(Ko *et al*, 2018; Zeng *et al*, 2018; 匡卫民等, 2019; Chan *et al*, 2019)。线粒体物种特异性 DNA 片段, 如核糖体 RNA (12S rRNA 和 16S rRNA)、细胞色素 *b* (Cyt *b*)和细胞色素 *c* 氧化酶 I (COI)通常用于鱼类物种鉴定(Cutarelli *et al*, 2018)。目前有较多的学者用线粒体基因组全序列构建系统发育树来比较进化关系, 相比于其他的 COI、12S rRNA 和 16S rRNA 等单一基因序列构建的系统发育树可以减少因信息分布的不均一性和序列长度太短而造成的进化树置信度降低带

* 浙江省重点研发计划项目, 2021C2047 号。廖贤晖, 硕士研究生, E-mail: 741050944@qq.com

① 通信作者: 高天翔, 博士生导师, 教授, E-mail: gaotianxiang0611@163.com

收稿日期: 2023-02-28, 收修改稿日期: 2023-04-28

来的影响, 使得研究结果更精确、可靠(肖家光, 2015)。

瑞氏红鲂鲉(*Satyrichthys rieffeli*)隶属鲉形目(Scorpaeniformes)、黄鲂鲉科(Peristediidae)、红鲂鲉属(*Satyrichthys*) (中坊徹次, 2013; 陈大刚等, 2015; 赵盛龙等, 2016), 广泛分布于黄海南部、东海、台湾海域、海南岛、日本、加里曼丹和西北太平洋海域(伍汉霖等, 2021)。目前关于瑞氏红鲂鲉的研究报道较少, 主要是对其分类特征方面的描述(Bussing, 2010; Ono *et al*, 2014), 遗传学方面, 仅在 GenBank 中发现 2 条 COI 基因片段序列。本研究采用高通量测序技术, 对瑞氏红鲂鲉的全基因组大小、杂合度等指标进行测定和评估; 通过基因组 survey 结果挖掘出线粒体基因组序列并对其进行注释分析, 为制定全基因组测序方案提供理论依据; 结合 GenBank 中 18 种鲉形目鱼类的线粒体基因组全序列构建系统发育树, 为瑞氏红鲂鲉分类及其系统发育关系深入研究奠定基础。

1 材料与方法

1.1 样品采集与保存

本研究所用瑞氏红鲂鲉样品于 2021 年 12 月采自东海大陆坡 1955 海区至 1956 海区, 共采集 6 尾样品, 冷冻保存运至实验室, 常规生物学测定后, 取 1 g 背部肌肉于–80 °C 保存, 用于后续 DNA 提取; 选取其中一尾进行 survey 测序。

1.2 实验方法

1.2.1 基因组 survey 测序 用标准的苯酚-氯仿法提取肌肉组织的总基因组 DNA (Maniatis *et al*, 1982)。检测合格的 DNA 样本通过超声波破碎仪随机打断成长度为 300~350 bp 的片段, 经末端修复、加 A 尾、加测序接头、纯化、PCR 扩增等步骤完成整个文库制备。构建好的文库通过 Illumina nova 进行 PE150 测序。使用 fastp 软件对测序数据进行数据过滤与去冗余操作(Kajitani *et al*, 2014), 过滤后得到的高质量数据将用于后续的杂合度、GC 含量以及基因组大小等信息分析。

1.2.2 基因组大小预估和 K-mer 分析 利用过滤后的数据, 取 K=17, 参照 Liu 等(2013)的 K-mer 分析方法对瑞氏红鲂鲉基因组大小、杂合率和重复率等基因特征进行进一步分析。基于 K-mer 分析的结果, 获得了关于峰的深度和预测的最佳 K-mer 数量的信息, 并用于估计基因组的大小。

1.2.3 基因组组装 使用 SOAPdenovo 软件(Luo *et al*, 2012)对过滤去除低质量数据之后的高质量测序数据进行基因组预组装分析。采用 K=48 构建 Contig 和 Scaffold, 利用高质量数据进行全基因组组装, 使用 SOAPdenovo 软件将过滤后的 reads 比对拼接到该组装序列上, 获得原始基因组序列及碱基深度。

1.2.4 线粒体基因组组装和注释 借助 NOVOPlasty 2.6.3 软件从原始的基因组测序数据中提取线粒体基因组序列(Dierckxsens *et al*, 2017), 参数采用默认设置, 参考序列为 GenBank 瑞氏红鲂鲉 COI 序列(GenBank No.MK777566.1)。通过 MitoFish 将得到的数据进行注释和可视化(Iwasaki *et al*, 2013)。利用 MEGA 11 计算基因组碱基组成、氨基酸使用频率及同义密码子相对使用频率(匡卫民等, 2019)。

1.2.5 系统进化分析 GenBank 中下载 18 种鲉形目鱼类的线粒体基因组, 以豹鲂鲉(*Dactylopterus volitans*)为外群, 基于 12 种蛋白质编码基因(除 ND6)进行系统发育分析, 确定瑞氏红鲂鲉的分类地位。利用 MEGA X 和贝叶斯法(Bayesian)构建最大似然树(ML)和贝叶斯树(BI) (Ronquist *et al*, 2003), 采用 MEGA X 方法估计了最优核苷酸替换模型(Kumar *et al*, 2018), 基于 BIC 标准得到最优 GTR+G+I 模型, 确定构建系统发育树的最佳模型。利用 MEGA X 软件构建了 1 000 个重复的最大似然系统发育树, 探究系统进化关系。

2 结果

2.1 基因组 survey 结果

2.1.1 测序数据统计及质量评估 使用 Illumina nova 测序平台进行双末端测序, 使用瑞氏红鲂鲉基因组总 DNA 构建 300 bp 的文库, 测序得到 raw reads 57.59 Gb, 质控后得到约 49.13 Gb 的 clean reads。瑞氏红鲂鲉质控后其 Q_{20} 值为 96.73%, Q_{30} 值为 91.72% (表 1), 显示本次瑞氏红鲂鲉的高通量测序结果有较高的准确性, 结果可用于后续的分析。

表 1 瑞氏红鲂鲉**survey** 测序结果

Tab.1 Results of the survey sequencing of *S. rieffeli*

名字	碱基数	碱基大小/Gb	Q_{20}/%	Q_{30}/%	GC 含量/%
原始读数	383 941 058	57.59	96.57	91.64	41.29
有效读数	329 460 024	49.13	96.73	91.72	41.24

注: Q_{20}: 质量百分比≥20; Q_{30}: 质量百分比≥30

2.1.2 污染评估及处理 为了验证瑞氏红鲂鲉的 DNA 是否存在污染, 从 350 bp 文库中对过滤后的高质量数据随机抽取 10 000 对 reads 数据, 通过 Blast

软件比对 NCBI 核苷酸数据库(NT 库), 将前 80%比对上的物种展示出来(表 2)。其中比对结果最高的是石斑鱼属的鱼类, 占比 19.93%, 与瑞氏红鲀鲋差异较远。对比结果没有出现较高比例的物种, 比对近源物种的 DNA, 可以判断此次测序样品未被污染, 并做进一步分析。

表 2　瑞氏红鲀鲋高通量测序结果与 **NT** 库比对

Tab.2　The results of *S. rieffeli* comparision with the NT library

物种		比对数量/bp	比对总数/bp	占比%
Epinephelus	石斑鱼属	1 185	5 945	19.93
Cottoperca	杜父鲈鳚属	672	5 945	11.3
Lateolabrax	花鲈属	549	5 945	9.23
Plectropomus	鳃棘鲈属	530	5 945	8.92
Sparus	鲷属	276	5 945	4.64
Scophthalmus	菱鲆属	267	5 945	4.49
Sebastes	平鲉属	160	5 945	2.69
Nibea	黄姑鱼属	144	5 945	2.42
Anarrhichthys	狼鳗属	134	5 945	2.25
Trachurus	竹荚鱼属	122	5 945	2.05
Myripristis	锯鳞鱼属	121	5 945	2.04
Sphaeramia	圆竺鲷属	120	5 945	2.02
Larimichthys	黄鱼属	120	5 945	2.02
Pseudochaenichthys	拟冰鱼属	108	5 945	1.82
Acanthopagrus	棘鲷属	108	5 945	1.82
Cyprinus	鲤属	107	5 945	1.80
Perca	鲈属	100	5 945	1.68

2.1.3　K-mer 分析　　使用 K-mer 分析预测了瑞氏红鲀鲋的基因组特征及杂合率和重复比例。选取 *K*=17, 预测瑞氏红鲀鲋基因组大小为 827 Mb, 修正后的基因组大小为 813 Mb; 基因组杂合率为 0.92%, 重复序列比例为 45.97% (表 3)。其 K-mer 深度在 48 时出现一个期望深度, 且没有出现其他峰值(图 1), 说明瑞氏红鲀鲋基因组杂合度较低。

2.1.4　瑞氏红鲀鲋基因组组装结果　　通过 SOAPdenovo 初步组装得到 Contigs 的结果见表 4。N50 和 N90 的值可以说明 survey 结果组装的质量, *S. rieffeli* contig N50 和 N90 分别为 712 bp 和 134 bp, Contigs 大于 100 bp 的占总数的 98.42%; A、T、G、C 和 N 的占比分别为 29.64%、28.81%、20.52%、21.02%和 0, GC 含量为 41.55%。

2.2　线粒体基因组注释

2.2.1　线粒体基因组组成及其特征　　从 survey 结果里面提取并组装线粒体基因组, 结果显示瑞氏红鲀鲋的线粒体全序列长 16 527 bp, 其中有 38 个基因获得注释, 分别由 22 个 tRNA、13 个蛋白质编码基因(PCGs: *APT6*、*ATP8*、*Cyt-b*、*COXI~COXIII*、*ND1~ND6* 和 *ND4L*)、2 个 rRNA (*12S rRNA* 和 *16S rRNA*)和 1 个 D-loop 区组成(图 2, 表 5)。

除 8 个 tRNA 基因(*tRNA-Gln*、*Ala*、*Asn*、*Cys*、*Tyr*、*Ser*、*Glu*、*Pro*)和 1 个 PCGs 基因(ND6)在线粒体基因组的 L 链上, 其他 29 个基因和 D-loop 区均在 H 链上, 基因间都未发生重排。其中 11 个基因发生了间隔, 共 99 bp, *tRNA-Cys* 基因间隔最长, 共 39 bp, 其次是基因 *ND6*, 共 32 bp; 有 4 个基因发生了重叠, 共 8 bp, 重叠的基因是 *tRNA-Gln*、*tRNA-Met*、*ND4* 和 *tRNA-Pro*, 其中基因 ND4 重叠最多, 有 5 bp。瑞氏红鲀鲋线粒体全基因碱基组成: A (27.4%)、T (26.9%)、G (16.8%)、C (28.9%) (表 6), 各基因的 GC

表 3　瑞氏红鲀鲋基因组特征统计

Tab.3　Statistics of *S. rieffeli* genome characteristics

样品	K-mer 数	K-mer 深度	基因组大小/Mb	修正后基因组大小/Mb	杂合率/%	重复率/%
S. rieffeli	43 850 343 050	48	827	813	0.92	45.97

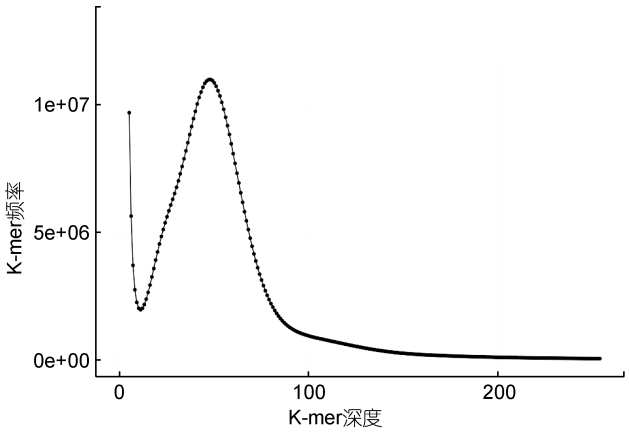

图 1 瑞氏红鲀鲋K-mer 深度分布图

Fig. 1 Distribution of the K-mer depth of *S. rieffeli*

含量在 44.4%~51.4%之间, 其中 A+T 的含量(54.3%)大于 G+C 的含量(45.7%)。

2.2.2 蛋白质编码基因(PCGs) 瑞氏红鲀鲋的蛋白质编码基因(PCGs)由 *ATP6*、*ATP8*、*Cyt-b*、*COXI-III*、*ND1-6* 和 *ND4L* 组成, 其大小为 11 400 bp, 占线粒体基因组的 68.98%。12 个 PCGs 在 H 链上编码, 只有 *ND6* 在线粒体基因组的 L 链编码。蛋白质编码基因的起始密码子除 *COI* 是 GTG 外, 其余均是 ATG; 终止密码子除 *ND1* 和 *ND6* 是 TAG, 其他都是 TAA, 其中有不完整终止密码子 TA-或 T--; 有 6 个完整的终止密码子(*ND1*、*COXI*、*ATP8*、*ND4L*、*ND5*、*ND6*), 7 个不完整的终止密码子(*Cyt b*、*ND4*、*ND3*、*COXIII*、*ATP6*、*COXII*、*ND2*)。当蛋白质编码基因(如 *ND2*、

表 4 Contigs 组装结果统计

Tab.4 Statistics of the contigs assembly

DNA 片段	总长/bp	总数(≥1 kb)	最长/bp	N50/bp	N90/bp	GC 含量	Contigs>100 bp
Contigs	752 061 430	2 083 983	31 368	712	134	41.55%	2 051 006

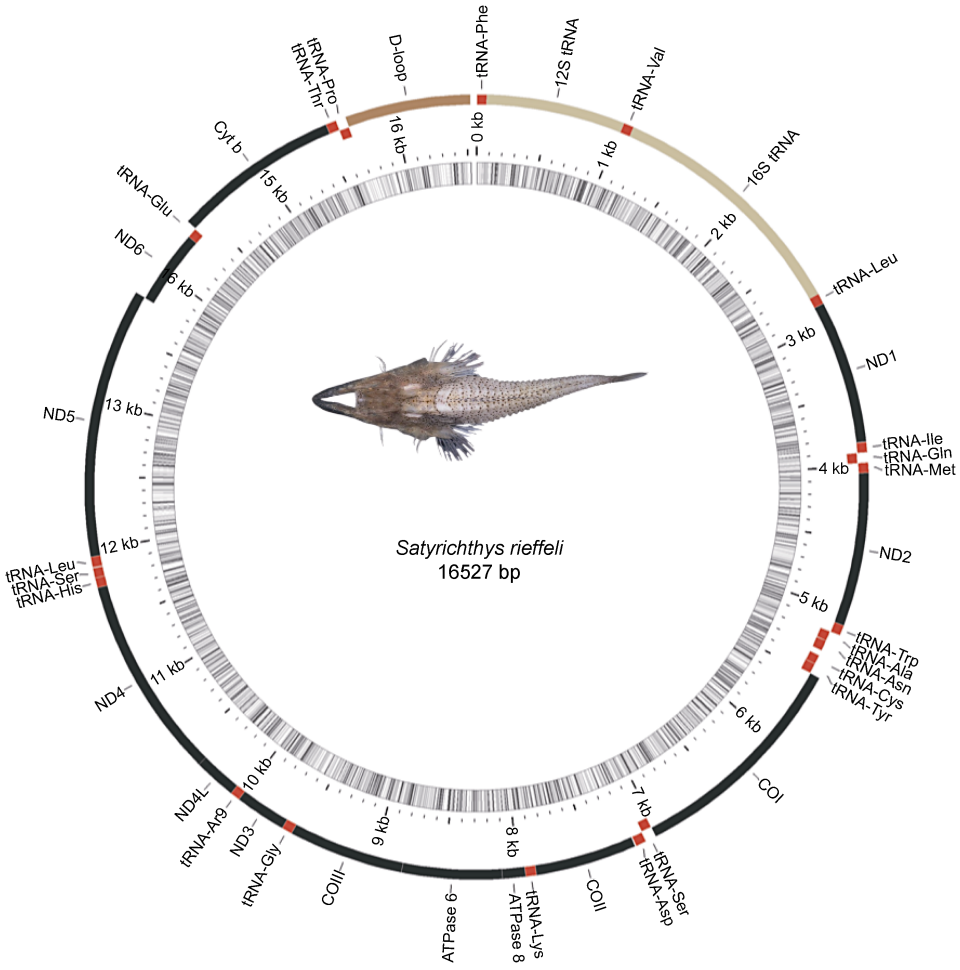

图 2 瑞氏红鲀鲋线粒体基因组的圆形图谱

Fig.2 The circular map of the mitogenome of *S. rieffeli*

表 5　瑞氏红鲔鳚的线粒体基因组结构特征

Tab.5　Structural characteristics of the mitochondrial genome of *S. rieffeli*

基因	起始位点	终止位点	长度/bp	氨基酸	起始密码子	终止密码子	反密码子	基因间隔/bp	链
tRNA-Phe	1	68	68				GAA	0	H
12S rRNA	69	1 014	946					0	H
tRNA-Val	1 015	1 086	72				TAC	0	H
16S rRNA	1 087	2 785	1 699					0	H
tRNA-Leu	2 786	2 859	74				TAA	0	H
ND1	2 860	3 834	975	324	ATG	TAG		0	H
tRNA-Ile	3 839	3 908	70				GAT	4	H
tRNA-Gln	3 908	3 978	71				TTG	–1	L
tRNA-Met	3 978	4 046	69				GAT	–1	H
ND2	4 047	5 092	1 046	348	ATG	TA		0	H
tRNA-Trp	5 093	5 163	71				TCA	0	H
tRNA-Ala	5 165	5 233	69				TGC	1	L
tRNA-Asn	5 235	5 307	73				GTT	1	L
tRNA-Cys	5 347	5 412	66				GCA	39	L
tRNA-Tyr	5 413	5 482	70				GTA	0	L
COI	5 484	7 034	1 551	516	GTG	TAA		1	H
tRNA-Ser	7 035	7 105	71				GCT	0	L
tRNA-Asp	7 109	7 181	73				GTC	3	H
COII	7 189	7 879	691	230	ATG	T		7	H
tRNA-Lys	7 880	7 953	74				TTT	0	H
ATPase 8	7 955	8 122	168	55	ATG	TAA		1	H
ATPase 6	8 113	8 795	683	227	ATG	TA		0	H
COIII	8 796	9 580	785	261	ATG	TA		0	H
tRNA-Gly	9 581	9 652	72				TCC	0	H
ND3	9 653	10 001	349	116	ATG	T		0	H
tRNA-Arg	10 002	10 070	69				TCG	0	H
ND4L	10 071	10 367	297	98	ATG	TAA		0	H
ND4	10 361	11 741	1 381	460	ATG	T		–5	H
tRNA-His	11 742	11 810	69				GTG	0	H
tRNA-Ser	11 811	11 878	68				TGA	0	H
tRNA-Leu	11 884	11 956	73				TAG	5	H
ND5	11 957	13 759	1 803	600	ATG	TAA		0	H
ND6	13 792	14 313	522	173	ATG	TAG		32	L
tRNA-Glu	14 314	14 382	69				TTC	0	L
Cyt b	14 388	15 528	1 141	380	ATG	T		5	H
tRNA-Thr	15 529	15 600	72				TGT	0	H
tRNA-Pro	15 600	15 669	70				TGG	–1	L
D-loop	15 670	16 527	858					0	H

COXII、*ATP6*、*COXIII*、*ND3*、*ND4*、*Cyt b*)的 3′端后面是编码一个 tRNA 基因在同一链上时, 主要使用不完全终止密码子 TA-、T-和 AG-。通过比较瑞氏红鲔鳚的 13 个蛋白质编码基因的 AT-skew 和 GC-skew 值(图 3), 可以看出除 AT-skew 的 COI 与 ATP8 和 GC-skew 的 ND6 为正值以外, 其余均为负值。

2.2.3　RNA 基因　　瑞氏红鲔鳚线粒体基因组中包含 22 个 tRNA, 各 tRNA 的序列长度分别为 66~74 bp,

表 6 线粒体基因组核苷酸组成比例

Tab.6 The composition ratio of mitochondrial genome nucleotide

基因	T/%	C/%	A/%	G/%	A+T/%	C+G/%	AT-Skew	GC-Skew
ND1	30.8	31	22.4	15.8	53.2	46.8	–0.158	–0.325
ND2	28	33.1	26.3	12.5	54.3	45.6	–0.031	–0.452
COI	30.5	27.3	24.2	18	54.7	45.3	–0.115	–0.205
COII	27.4	28	28.3	16.4	55.7	44.4	0.016	–0.261
ATPase 8	21.2	36.4	30.9	11.5	52.1	47.9	0.186	–0.520
ATPase 6	28.6	32.9	25.3	13.2	53.9	46.1	–0.061	–0.427
COIII	28.4	29.8	24.3	17.6	52.7	47.4	–0.078	–0.257
ND3	32.8	31.3	18.7	17.2	51.5	48.5	–0.274	–0.291
ND4L	28.6	34.7	20.1	16.7	48.7	51.4	–0.175	–0.350
ND4	28.3	30.9	25.8	14.9	54.1	45.8	–0.046	–0.349
ND5	27.8	31.0	27.0	14.1	54.8	45.1	–0.015	–0.375
ND6	37.6	13.7	15.2	33.5	52.8	47.2	–0.424	0.419
Cyt b	29.1	31.8	23.8	15.3	52.9	47.1	–0.100	–0.350
PCGs	29.2	30.0	24.6	16.2	53.8	46.2	–0.086	–0.299
tRNA	26.7	21.4	27.6	24.2	54.3	45.6	0.017	0.061
rRNA	21.9	25.3	31.9	20.9	53.8	46.2	0.186	–0.095
CR	30.0	21.4	32.3	16.3	62.3	37.7	0.037	–0.135
Mitogenome	26.9	28.9	27.4	16.8	54.3	45.7	0.009	–0.265

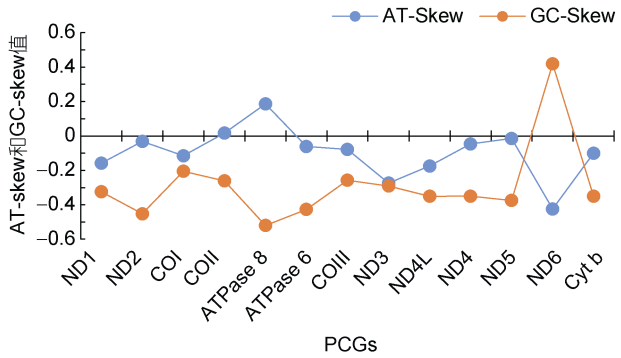

图 3 瑞氏红鲌鲴13 个蛋白质编码基因的 AT-skew 和 GC-skew 值

Fig.3 The AT-skew and GC-skew values of 13 protein-coding genes in *S. rieffeli*

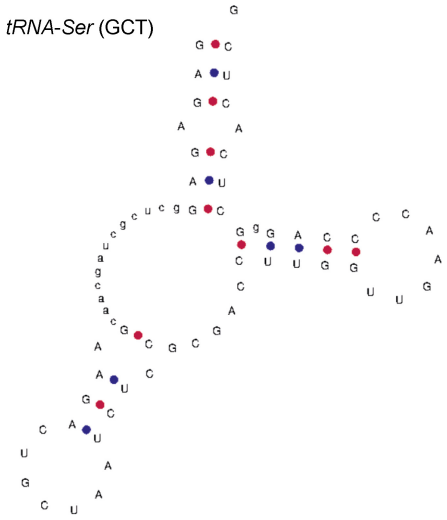

图 4 瑞氏红鲌鲴*tRNA-Ser* (GCT)二级结构预测图

Fig.4 Prediction of secondary structure of *tRNA-Ser* (GCT) in *S. rieffeli*

其中 *tRNA-Leu* (TAA、TAG)和 *tRNA-Ser* (GCT、TGA)由两种反密码子编码，其他的都是一种反密码子编码。此外，21 个 tRNA 呈典型的三叶草二级结构，只有 *tRNA-Ser* (GCT)缺失了二氢尿苷臂(DHU 臂)的简单环(图 4)。

瑞氏红鲌鲴线粒体基因组中发现 2 个 rRNA 基因(*12S rRNA* 和 *16S rRNA*)，其长度分别为 946 bp 和 1 699 bp，rRNA 的碱基含量为分别为 A (31.9%)、T (21.9%)、C (25.3%)、G (20.9%) (表 6)，没有表现出明显的反 G 偏倚。

2.2.4 线粒体基因组密码子使用情况 瑞氏红鲌鲴线粒体基因组中密码子使用最常见的氨基酸为异亮氨酸(Leu)、脯氨酸(Pro)、苏氨酸(Thr)和丝氨酸(Ser-TGA)，使用最少的是半胱氨酸(Cys)，仅由两个密码子编码(图 5)，氨基酸脯氨酸(Pro)的使用量最高，可以被 4 种不同密码子编码，异亮氨酸(Leu)可以被 6 种不同的密码子编码。而相对同义密码子使用频率

(relative synonymous codon usage, RSCU)也被用于评估线粒体基因密码子的使用情况(图 6)。瑞氏红鲃鲃的同义密码子使用中 NNC 和 NNT (N 代表 A、T、C、G)的使用频率更高。

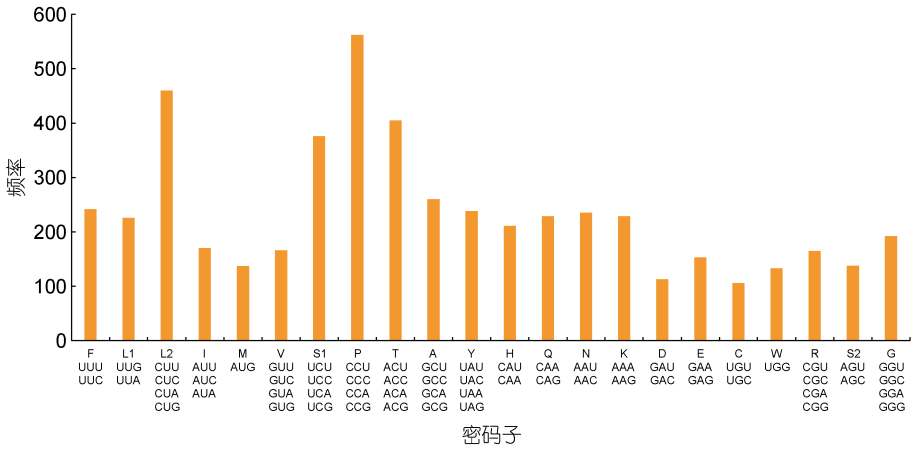

图 5　瑞氏红鲃鲃线粒体基因密码子使用频率

Fig.5　Frequency of usage of mitochondrial gene codon of *S. rieffeli*

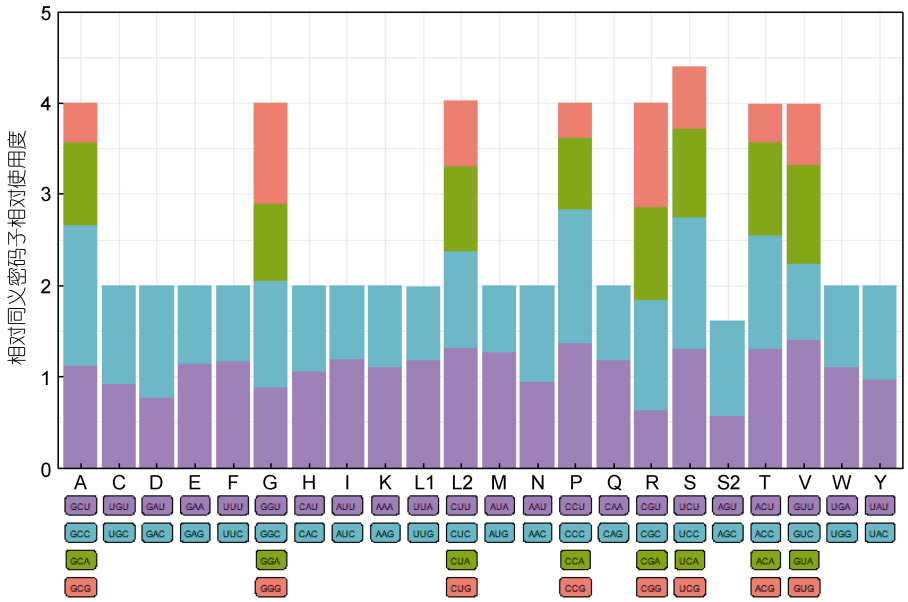

图 6　瑞氏红鲃鲃线粒体基因相对同义密码子使用频率

Fig.6　Frequency of relative synonymous codon usage (RSCU) in *S. rieffeli*

2.2.5　非编码区　　非编码区又称之为 D-环区(displacement-loop region), 位于 *tRNA-Pro* 和 *tRNA-Phe* 基因之间, 是整个线粒体基因组序列和长度变异最大的区域, 但其中也包含保守片段。瑞氏红鲃鲃的非编码区各碱基含量分别为 T (30.0%)、C (21.4%)、A (32.3%)、G (16.3%), 表现出明显的 AT 偏好性, A+T 的含量达到了 62.3%, G+C 的含量仅为 37.7%。

2.3　系统发育关系

为了分析瑞氏红鲃鲃的系统发育关系, 从 NCBI 数据库里面下载 18 个鲉形目(7 科 18 种)鱼类的线粒体基因组, 选择在 H 链上的蛋白质编码基因(除 ND6 以外的所有蛋白质编码基因)使用最大似然法和贝叶斯法构建 ML 树和 BI 树(图 7, 图 8)。以豹鲃鲃为外群, 构建的系统发育树显示, BI 树中鲉科鱼类首先与豹鲃鲃科鱼类先分离出来, 鲃鲃科和黄鲃鲃科的亲缘关系更近, 作为姐妹类群分为一支; 而狮子鱼科、六线鱼科和杜父鱼科鱼类作为一支分离出来。ML 树与之不同的是鲉科与其他 5 科鱼类形成姐妹类群分开, 与 BI 树相同的是黄鲃鲃科与鲃鲃科鱼类亲缘关系较近, 互为姐妹类群, 瑞氏红鲃鲃与须叉吻鲃鲃亲

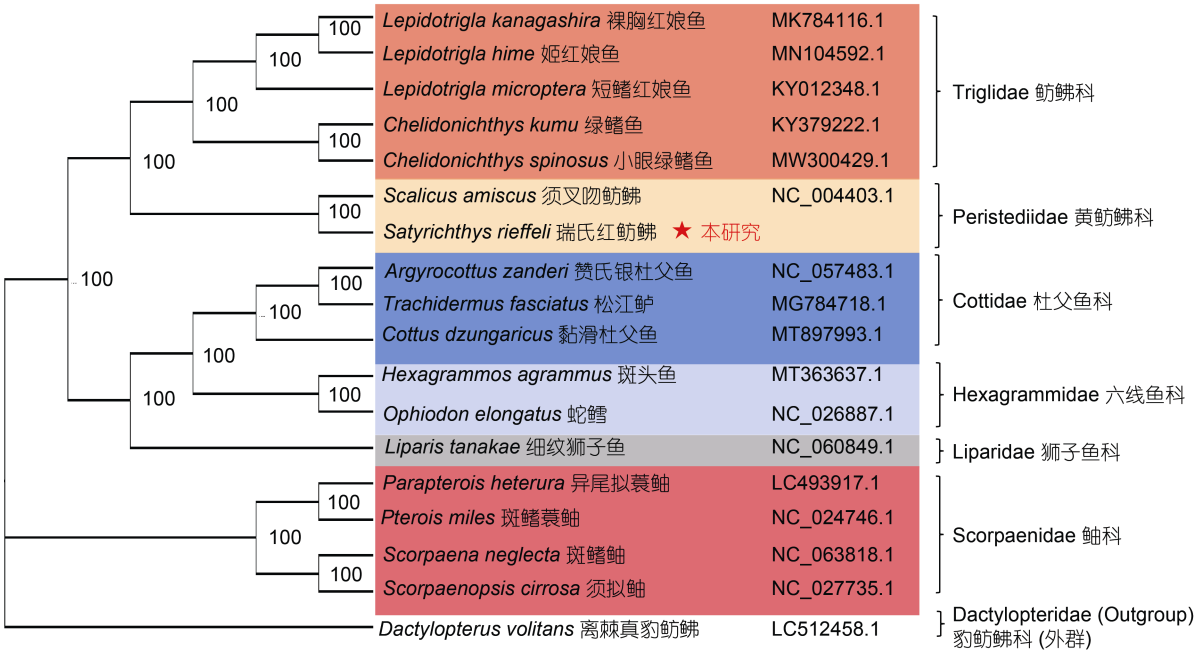

图 7 基于瑞氏红鲂鲉12 个 PCGs 构建的 BI 树

Fig.7 The BI tree based on 12 PCGs of *S. rieffeli*

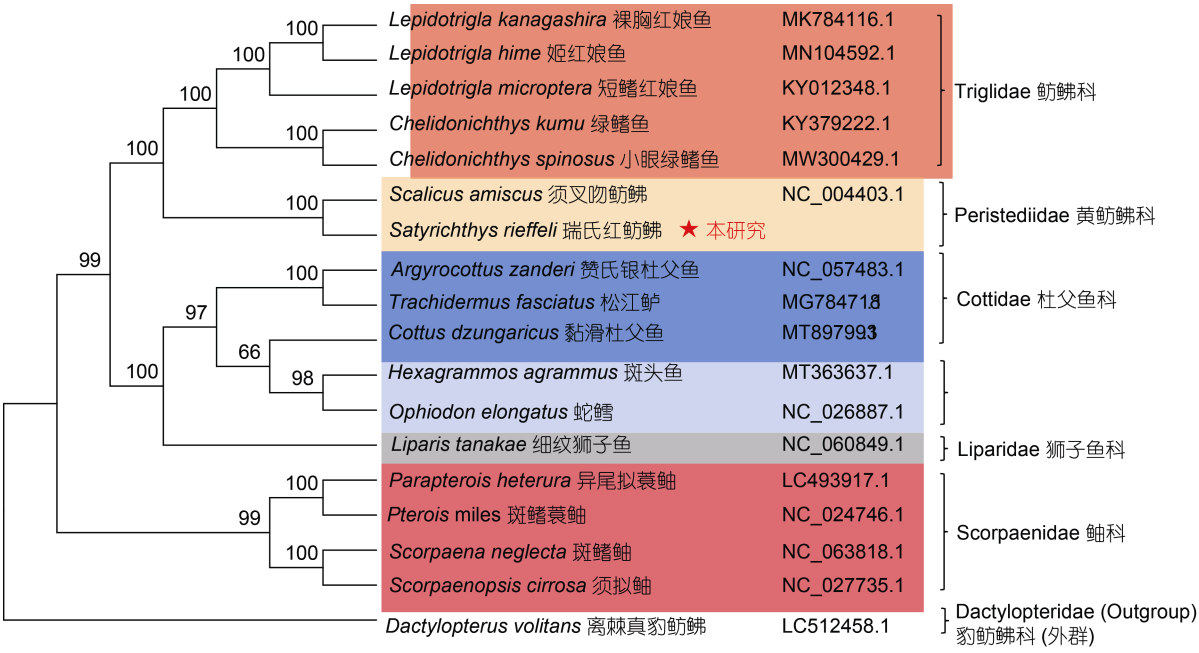

图 8 基于瑞氏红鲂鲉12 个 PCGs 构建的 ML 树

Fig.8 The ML tree based on 12 PCGs of *S. rieffeli*

缘关系最近，同属黄鲂鲉科鱼类。

3 讨论

近年来，高通量测序技术的迅速发展为解决非模式海洋鱼类的基因组问题提供了有效的技术手段。利用 K-mer 方法可以估计非模型物种的基因组大小，无须任何先验知识(Li *et al*, 2019b)，其方法已被应用于绒杜父鱼(*Hemitripterus villosus*) (赵蕊蕊等, 2022)、褐菖鲉 (*Sebastiscus marmoratus*) (Xu *et al*, 2020)、菖鲉属(*Sebastiscus*) (Jia *et al*, 2021)等鲉形目鱼类的研究。目前瑞氏红鲂鲉的分子信息还不完善，本研究 survey 结果显示 GC 含量在 41%左右出现峰值(双端

测序结果都出现在 41%左右), 且没有出现明显的偏差, 表明测序结果没有偏向性; 并获得 49.13 Gb 的 clean reads, Q_{20} 大于 90%, 基因组大小为 827 Mb, 修正后的基因组大小为 813 Mb, 其杂合率为 0.92%, 重复率为 45.97%。瑞氏红鲃鲃与大多数的海洋鱼类相似, 其基因组大小稍大于绒杜父鱼 713.18 Mb (赵蕊蕊等, 2022) 、褐菖鲉 812.86 Mb (Xu *et al*, 2020)、艾氏蛇鳗(*Ophichthus lithinus*) 755.24 Mb (宁子君等, 2022), 稍小于许氏平鲉(*Sebastes schlegelii*) 846.36 Mb、朝鲜平鲉(*Sebastes koreanus*) 832.53 Mb、金斑平鲉(*Sebastes nudus*) 813.12 Mb (Xu *et al*, 2019)、双带缟鲕虎(*Tridentiger bifasciatus*) 887.60 Mb (Zhao *et al*, 2022), 导致物种间基因组大小的差异可能是由于重复序列含量差异所导致。瑞氏红鲃鲃的重复率为 45.97%, 较绒杜父鱼 38.61%、褐菖鲉 39.65%、艾氏蛇鳗 43.30%、许氏平鲉 44.10%、朝鲜平鲉 43.65%、金斑平鲉 41.40%、双带缟鲕虎 32.60%均较大。此外, 瑞氏红鲃鲃的杂合率为 0.92%, 较绒杜父鱼 0.26%、褐菖鲉 0.17%、艾氏蛇鳗 0.70%、许氏平鲉 0.22%、朝鲜平鲉 0.20%、金斑平鲉 0.31%、双带缟鲕虎 0.47%均较大, 一般来说, 杂合率大于 0.8%, 重复率大于 60%就称为复杂基因组(高胜寒等, 2018), 由于瑞氏红鲃鲃的杂合率高于 0.8%, 重复率低于 60%, 其基因组不属于复杂基因组类型。在本研究中, 组装的初步结果显示 Contigs 的 N50 为 712 bp, N50 数量为 254 317, 其结果满足基因组组装要求。通过过滤后的 clean reads 组装瑞氏红鲃鲃的基因组, 这是目前第一个基于二代测序技术组装瑞氏红鲃鲃的基因组, 该基因组为瑞氏红鲃鲃的进化生物学研究提供了基础数据, 也为进一步探索该物种的基因组特征提供了参考。由于二代测序技术存在读长短、组装结果的连续性无法保证以及由于基因组的杂合导致过度组装或组装不彻底等问题, 因此本研究缺少对 GC-depth 和含量的相关分析, 后续开展三代测序以获得瑞氏黄鲃鲃高质量的基因组。

本研究组装出来的线粒体基因组大小为 16 527 bp, 线粒体结构未出现基因重排现象, GC 含量为 45.7%, 出现明显的 AT 偏倚, 其结果符合硬骨鱼类线粒体碱基组成偏好于碱基 A 和 T 的特点(Broughton *et al*, 2001; 蒙子宁等, 2004; Consuegra *et al*, 2015), 碱基 G 的含量为 16.8%, 表现出明显的反 G 偏倚(孟刚等, 2021), 与大多数鱼类研究结果一致(丁少雄等, 2006)。13 个 PCGs 中, *COI* 是以 GTG 作为起始密码子, 其余均为 ATG 起始密码子, 其中密码子 ATG 是翻译效率最高的一个(Consuegra *et al*, 2015)。PCGs 终止密码子中出现 T--和 TA-这类不完整的终止密码子, 是因为这些基因之后是一个编码在同一条链上的基因, 允许转录在没有完整密码子的情况下终止(Hecht *et al*, 2017)。不完全终止密码子的存在在鱼类的线粒体基因组中很常见, 可以通过在 mRNA 加工过程中添加一个 poly A 尾巴来完成(Ojala *et al*, 1981; Liu *et al*, 2009)。PCGs 的 AT-skew 和 GC-skew 为负值(表 6), 一般来说, AT 倾斜的幅度小于 GC 倾斜, 而且在许多情况下, 没有统计学意义, 在这里, AT-skew 低于 GC-skews (绝对值), 这符合传统的偏好(Yu *et al*, 2019)。AT-skew 的最大值和 GC-skew 的最小值均出现在 ND6 中, 且该基因的 AT/GC-skew 值波动较大, 核苷酸偏斜可能是由于在复制和转录过程中突变压力和选择压力之间的平衡, 为基因复制提供了一个潜在的方向(McLean *et al*, 1998; Touchon *et al*, 2008)。tRNA 基因结构中仅 *tRNA-Ser* (TCG)基因缺失 DHU 臂的结构, 这是鱼类线粒体基因组的共同特征, DHU 臂的缺失改变了 *tRNA-Ser* (TCG)在线粒体基因中的识别潜力, 使其更容易被识别(Hardt *et al*, 1993)。瑞氏红鲃鲃线粒体基因组的 *12S rRNA* 和 *16S rRNA* 基因定位于 H 链 *tRNA-Phe* 和 *tRNA-Leu* (UUR)基因之间, 中间以 *tRNA-Val* 基因为间隔, *12S rRNA* 基因比 *16S rRNA* 基因更保守(Satoh *et al*, 2016)。当 RSCU 值=1 时, 表示密码子的使用频率与其他简并密码子没有差异; 当氨基酸的 RSCU 值均不等于 1, 说明每个氨基酸的使用都有不同程度的偏倚(周丹等, 2013; 孟乾等, 2020)。D-loop 区在 *tRNA-Pro* 和 *tRNA-Phe* 之间, 与其他大多数脊椎动物的排列一样(Wei *et al*, 2010; Li *et al*, 2019a)。

目前通过线粒体基因组构建系统发育树来确定物种的进化关系已经成为系统发育分析的常用方法, 比如周志雄等(2020)通过线粒体基因组构建系统发育树得出东方鲀属内系统发育关系较为复杂, 且存在野生环境下种间自然杂交现象; Miya 等(2015)通过过去 15 年线粒体基因组发展应用做出的总结, 认为线粒体基因组分析和高通量测序技术会成为物种分类重要方法。本研究基于 12 个 PCGs 构建系统发育树, 以豹鲃鲃作为外群, 分析瑞氏红鲃鲃的系统发育关系。ML 树和 BI 树结果表明鲉科在 ML 树中与外群遗传距离较近, 聚在一起, 与在 BI 树中结果不一致, 但是两个树结果都显示鲉科与其他科遗传距离都较

远。瑞氏红鲂鲷与须叉吻鲂鲷聚为一支，作为黄鲂鲷科鱼类，与形态学研究结果一致(Kawai, 2008)。

4 结论

瑞氏红鲂鲷隶属于鲉形目黄鲂鲷科，目前仅见瑞氏红鲂鲷形态分类描述，尚未见遗传学相关研究报道。本研究基于基因组 survey 结果分析得到瑞氏红鲂鲷的基因组大小 827 Mb、杂合率 0.92%和重复序列比列 45.97%；线粒体基因组大小为 16 527 bp，由 38 个基因组构成，未出现基因重排现象。结合 GenBank 中 18 种鲉形目鱼类线粒体基因组，构建两种系统发育树比较分析，结果显示瑞氏红鲂鲷为黄鲂鲷科鱼类，支持形态学分类结果，为开展黄鲂鲷科鱼类系统发育深入研究奠定了基础。

致谢 感谢赵宸枫在采集样品和鉴定方面的帮助及杨天燕老师的修改意见，谨致谢忱。

参 考 文 献

丁少雄, 王颖汇, 王军, 等, 2006. 基于 16S rDNA 部分序列探讨中国近海 30 种石斑鱼类的分子系统进化关系[J]. 动物学报, 52(3): 504-513.

王子寅, 刘秉儒, 李子豪, 等, 2023. 基于 Survey 分析的濒危植物四合木基因组研究[J]. 植物研究, 43(2): 186-193.

宁子君, 刘玉萍, 张书飞, 等, 2022. 艾氏蛇鳗线粒体基因组全序列结构分析和系统发育关系探讨[J]. 中国水产科学, 29(9): 1264-1276.

匡卫民, 于黎, 2019. 基因组时代线粒体基因组拼装策略及软件应用现状[J]. 遗传, 41(11): 979-993.

伍汉霖, 钟俊生, 2021. 中国海洋及河口鱼类系统检索[M]. 北京: 中国农业出版社: 546.

肖家光, 2015. 基于线粒体基因组全序列的鳢属鱼类系统发育研究[D]. 青岛: 中国海洋大学: 24-25.

张方, 米志勇, 1998. 动物线粒体 DNA 的分子生物学研究进展[J]. 中国生物工程杂志, 18(3): 25-31, 6.

陈大刚, 张美昭, 2015. 中国海洋鱼类[M]. 青岛: 中国海洋大学出版社: 811.

周丹, 薛仁余, 张晓峰, 等, 2013. 鲤和斑马鱼 *HOX* 基因家族同义密码子使用偏性的分析[J]. 水产学杂志, 26(2): 19-25.

周志雄, 刘波, 宫杰, 等, 2020. 基于线粒体基因组的东方鲀属系统发育学及群体遗传学[J]. 水产学报, 44(11): 1792-1803.

孟刚, 王瑞娴, 楚渠, 2021. 野桑蚕基因组 Survey 分析及线粒体基因组注释[J]. 基因组学与应用生物学, 40(S2): 2505-2512.

孟乾, 张志勇, 张志伟, 等, 2020. 斑石鲷和条石鲷线粒体基因组密码子使用分析[J]. 水产科学, 39(5): 702-709.

赵盛龙, 徐汉祥, 钟俊生, 等, 2016. 浙江海洋鱼类志[M]. 杭州: 浙江科学技术出版社: 523-524.

赵蕊蕊, 徐胜勇, 2022. 绒杜父鱼全基因组 survey 分析及微卫星分布特征[J]. 中国水产科学, 29(7): 994-1001.

高胜寒, 禹海英, 吴双阳, 等, 2018. 复杂基因组测序技术研究进展[J]. 遗传, 40(11): 944-963.

蒙子宁, 庄志猛, 丁少雄, 等, 2004. 中国近海 8 种石首鱼类的线粒体 16S rRNA 基因序列变异及其分子系统进化[J]. 自然科学进展, 14(5): 514-521.

中坊徹次, 2013. 日本産魚類検索全種の同定[M]. 3 版. 東海大学出版会: 727-730.

BARCHI L, LANTERI S, PORTIS E, *et al*, 2011. Identification of SNP and SSR markers in eggplant using RAD tag sequencing [J]. BMC Genomics, 12: 304.

BOORE J L, 1999. Animal mitochondrial genomes [J]. Nucleic Acids Research, 27(8): 1767-1780.

BROUGHTON R E, MILAM J E, ROE B A, 2001. The complete sequence of the zebrafish (*Danio rerio*) mitochondrial genome and evolutionary patterns in vertebrate mitochondrial DNA [J]. Genome Research, 11(11): 1958-1967.

BUSSING W A, 2010. A new fish, *Peristedion nesium* (Scorpaeniformes: Peristediidae) from Isla del Coco, Costa Rica [J]. Revista de Biologia Tropical, 58(4): 1149-1156.

CHAN E K F, TIMMERMANN A, BALDI B F, *et al*, 2019. Human origins in a southern African palaeo-wetland and first migrations [J]. Nature, 575(7781): 185-189.

CONSUEGRA S, JOHN E, VERSPOOR E, *et al*, 2015. Patterns of natural selection acting on the mitochondrial genome of a locally adapted fish species [J]. Genetics Selection Evolution, 47(1): 58.

CUTARELLI A, GALIERO G, CAPUANO F, *et al*, 2018. Species identification by means of mitochondrial cytochrome *b* DNA sequencing in processed anchovy, sardine and tuna products [J]. Food and Nutrition Sciences, 9(4): 369-375.

DIERCKXSENS N, MARDULYN P, SMITS G, 2017. NOVOPlasty: *de novo* assembly of organelle genomes from whole genome data [J]. Nucleic Acids Research, 45(4): e18.

HARDT W D, SCHLEGL J, ERDMANN V A, *et al*, 1993. Role of the D arm and the anticodon arm in tRNA recognition by eubacterial and eukaryotic RNase P enzymes [J]. Biochemistry, 32(48): 13046-13053.

HECHT A, GLASGOW J, JASCHKE P R, *et al*, 2017. Measurements of translation initiation from all 64 codons in *E. coli* [J]. Nucleic Acids Research, 45(7): 3615-3626.

IWASAKI W, FUKUNAGA T, ISAGOZAWA R, *et al*, 2013. MitoFish and mitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline [J]. Molecular Biology and Evolution, 30(11): 2531-2540.

JIA C H, YANG T Y, YANAGIMOTO T, *et al*, 2021. Comprehensive draft genome analyses of three rockfishes (Scorpaeniformes, *Sebastiscus*) via genome survey sequencing [J]. Current Issues in Molecular Biology, 43(3): 2048-2058.

KAJITANI R, TOSHIMOTO K, NOGUCHI H, *et al*, 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads [J]. Genome Research, 24(8): 1384-1395.

KAWAI T, 2008. Phylogenetic systematics of the family Peristediidae (Teleostei: Actinopterygii) [J]. Species Diversity,

13(1): 1-34.

KIM E B, FANG X D, FUSHAN A A, *et al*, 2011. Genome sequencing reveals insights into physiology and longevity of the naked mole rat [J]. Nature, 479(7372): 223-227.

KO A M S, ZHANG Y Q, YANG M A, *et al*, 2018. Mitochondrial genome of a 22,000-year-old giant panda from southern China reveals a new panda lineage [J]. Current Biology, 28(12): R693-R694.

KUMAR S, STECHER G, LI M, *et al*, 2018. MEGA X: Molecular evolutionary genetics analysis across computing platforms [J]. Molecular Biology and Evolution, 35(6): 1547-1549.

LI Q, WANG Q F, JIN X, *et al*, 2019a. Characterization and comparison of the mitochondrial genomes from two *Lyophyllum* fungal species and insights into phylogeny of *Agaricomycetes* [J]. International Journal of Biological Macromolecules, 121: 364-372.

LI Z Y, TIAN C X, HUANG Y, *et al*, 2019b. A first insight into a draft genome of silver sillago (*Sillago sihama*) via genome survey sequencing [J]. Animals (Basel), 9(10): 756.

LIU B H, SHI Y J, YUAN J Y, *et al*, 2013. Estimation of genomic characteristics by analyzing k-mer frequency in *de novo* genome projects [J]. arXiv: 1308.2012.

LIU Y, CUI Z X, 2009. The complete mitochondrial genome sequence of the cutlassfish *Trichiurus japonicus* (Perciformes: Trichiuridae): Genome characterization and phylogenetic considerations [J]. Marine Genomics, 2(2): 133-142.

LUO R B, LIU B H, XIE Y L, *et al*, 2012. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler [J]. GigaScience, 1(1): 18.

MANIATIS T, FRITSCH E F, SAMBROOK J, 1982. Molecular Cloning [M]. Cold Spring Harbor, USA: Cold Spring Harbor Laboratory.

MCLEAN M J, WOLFE K H, DEVINE K M, 1998. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes [J]. Journal of Molecular Evolution, 47(6): 691-696.

MIYA M, NISHIDA M, 2015. The mitogenomic contributions to molecular phylogenetics and evolution of fishes: a 15-year retrospect [J]. Ichthyological Research, 62(1): 29-71.

OJALA D, MONTOYA J, ATTARDI G, 1981. tRNA punctuation model of RNA processing in human mitochondria [J]. Nature, 290(5806): 470-474.

ONO M, KAWAI T, 2014. Review of Armored searobins of the genus *Peristedion* (Teleostei: Peristediidae) in Japanese waters [J]. Species Diversity, 19(2): 117-131.

RONQUIST F, HUELSENBECK J P, 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models [J]. Bioinformatics, 19(12): 1572-1574.

ROWE H C, RENAUT S, GUGGISBERG A, 2011. RAD in the realm of next-generation sequencing technologies [J]. Molecular Ecology, 20(17): 3499-3502.

RUAN H T, LI M, LI Z H, *et al*, 2020. Comparative analysis of complete mitochondrial genomes of three gerres fishes (Perciformes: Gerreidae) and primary exploration of their evolution history [J]. International Journal of Molecular Sciences, 21(5): 1874.

SATOH T P, MIYA M, MABUCHI K, *et al*, 2016. Structure and variation of the mitochondrial genome of fishes [J]. BMC Genomics, 17(1): 719.

SHI L L, YI S K, LI Y H, 2018. Genome survey sequencing of red swamp crayfish *Procambarus clarkii* [J]. Molecular Biology Reports, 45(5): 799-806.

TOUCHON M, ROCHA E P C, 2008. From GC skews to wavelets: a gentle guide to the analysis of compositional asymmetries in genomic data [J]. Biochimie, 90(4): 648-659.

WEI S J, SHI M, CHEN X X, *et al*, 2010. New views on strand asymmetry in insect mitochondrial genomes [J]. PLoS One, 5(9): e12708.

XU P, XU S Z, WU X H, *et al*, 2014. Population genomic analyses from low-coverage RAD-Seq data: a case study on the non-model cucurbit bottle gourd [J]. The Plant Journal, 77(3): 430-442.

XU S Y, ZHANG H, GAO T X, 2020. Comprehensive whole genome survey analyses of male and female brown-spotted flathead fish *Platycephalus* sp.1 [J]. Genomics, 112(6): 4742-4748.

XU S Y, ZHAO L L, XIAO S J, *et al*, 2019. Whole genome resequencing data for three rockfish species of *Sebastes* [J]. Scientific Data, 6(1): 97.YU P, ZHOU L, ZHOU X Y, *et al*, 2019. Unusual AT-skew of *Sinorhodeus microlepis* mitogenome provides new insights into mitogenome features and phylogenetic implications of bitterling fishes [J]. International Journal of Biological Macromolecules, 129: 339-350.

ZENG L, WEN J, FAN S G, *et al*, 2018. Species identification of fish maw (*Porcupinefish*) products sold on the market using DNA sequencing of 16S rRNA and COI genes [J]. Food Control, 86: 159-162.

ZHAO X, LIU Y X, DU X Q, *et al*, 2022. Whole-Genome survey analyses provide a new perspective for the evolutionary biology of shimofuri goby, *Tridentiger bifasciatus* [J]. Animals (Basel), 12(15): 1914.

THE GENOME SURVEY ANALYSIS AND MITOCHONDRIAL GENOME ANNOTATION OF *SATYRICHTHYS RIEFFELI*

LIAO Xian-Hui[1], WANG Yi-Ting[1], QU Yin-Quan[1], LIU Qi[2], GAO Tian-Xiang[1]
(1. *Fishery College*, *Zhejiang Ocean University*, *Zhoushan* 316022, *China*; 2. *Wuhan Wanmo Technology Company Limited*, *Wuhan* 430076, *China*)

Abstract In order to understand the genome characteristics and mitochondrial genome structure of *Satyrichthys rieffeli*, the second-generation high-throughput sequencing technology was used to conduct genome survey analysis of *S. rieffeli*, to provide basic information for studying its molecular content on the Illumina nova sequencing platform. Results show that the size of the assembled genome was 813 Mb, the heterozygosity was 0.92%, the proportion of repetitive sequences was 45.97%, and the N50 of Contigs was 712 bp. The mitochondrial genome of *S. rieffeli* is 16 527 bp long, with 38 genes (22 tRNAs, *12S rRNAs*, *16S rRNAs*, *ND1~ND6*, *COXI~COXIII*, *Cyt b*, *ATP8*, *ATP6*, *ND4L*, and one D-loop region). No rearrangement occurred among genes, and the GC content was 45.70%. Incomplete codons T-- and TA- appeared in protein coding genes. In tRNA, only *tRNA Ser* (GCT) lacked the simple loop of dihydrouridine arm (DHU arm), and the others had normal secondary structures. In addition, combined with the mitochondrial genomes of 18 Scorpaeniformes in NCBI, the molecular phylogenetic relationship of Scorpaeniformes was constructed using the maximum likelihood method and Bayesian method. Moreover, *S. rieffeli* and *Scalicus amiscus* form a clade as sisters as they belong to the same family Peristediidae, which is consistent with the previously reported morphological classification of *S. rieffeli*. The study laid the foundation for in-depth study on the phylogenic relationship of Peristediidae.

Key words Satyrichthys rieffeli; genome survey; mitochondrial genome; phylogenetic relationship