文章编号：1674-5566(2017)01-0008-09 DOI:10.12024/jsou.20160401724

鲚属单核苷酸多态性位点(SNPs)标记开发及在物种界定中的应用初探

王 倩，程方圆，李晨虹

(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)

摘 要：单核苷酸多态性(SNP)分子标记结合二代测序技术是研究群体遗传学强有力的工具，同时也是界定物种的最佳方法之一。与以往获取 SNPs 数据的常用方法相比，靶基因富集的方法可以用来富集不同物种的同源片段，甚至可以用于部分降解的 DNA。本研究应用一套通用的脊椎动物单拷贝核基因标记，通过基因富集和 Illumina 测序获取鲚形目鲚属数千个 SNPs 位点。样本取自靠近长江入海口的沿海地区、长江干流和洞庭湖。用 STRUCTURE 和 Bayes factor delimitation(* with genomic data，BFD*：一种新的物种界定工具)分析 SNPs 数据，发现洞庭湖的短颌鲚显著不同于其他采样点的样本，而从沿海地区采集的刀鲚与长江干流的个体之间的差异不明显(Bayes factor = 11.3)。研究表明，基因富集可以用来获得非模式生物的 SNPs 数据，结合新的分析工具如 BFD* 可用于物种界定。

关键词：靶基因富集；基因标记；物种界定；SNP 位点；BFD*；鲚属

中图分类号：S 917 **文献标志码**：A

得益于二代测序技术的发展，单核苷酸多态性标记(SNPs)在种群遗传、生态保护、农作物杂交、适应性进化、系统发育以及物种鉴定等领域得到了广泛的应用[1-11]。目前有两种方法识别非模式物种中的 SNPs。第一种方法是转录组测序获取 SNPs[12-14]。由于用此方法得出的 SNPs 位于转录区域，因此可能会导致在估计种群参数比如遗传漂变，基因流时产生误差[12]。另外，除非是与已注释好的基因组做比较[13,15]，或者从头组装[16]，否则在内含子外显子边界附近的 SNPs 难以验证。第二种方法是对简化基因组测序，如：基因分型测序(GBS)[9]，或者限制性酶切位点测序(RAD)[1,17]。但是，这些方法需要高质量的 DNA 样本，并且在不同物种中的可比性很小且质量不高[9]。所以不能用于跨物种的系统学研究[1,18]。

目前，基因富集技术成为第三种获得 SNPs 的渠道[19-22]，甚至可以用于古 DNA 的实验[20,23]。基因富集的目标也是灵活多样的，比如叶绿体基因组[23]、基因组中的某一区段[20]、和疾病相关的基因[24]、或者是外显子组都可以成为富集的目标基因[25]，并且基因富集技术可富集不同物种的同源基因[26]，这意味着我们可以利用富集技术来研究种间问题[26-27]。利用分子数据进行物种鉴定的方法有许多[28-32]。利用贝叶斯因子进行物种界定是一种常用的方法[2,32]，大多方法只能处理较少的位点[2]，然而 Bayes factor delimitation(* with genomic data，BFD*)可以用来分析上千 SNP 数据，它的优点是通过共显性标记直接推算物种树，而不需要估算基因树[33]。

刀鲚(*Coilia*)是一种鲚形目的鱼类[34-35]，长江中已经发现了多个刀鲚的亚种及生态型[34,36-37]。在洞庭湖中发现与刀鲚(*C. nasus*)不同的短颌鲚(*C. brachygnathus*)[38]，后来上颌骨较短的生态型在长江中游、下游，以及与中下游相通的其他湖泊中相继被发现[34,36-38]。这些物种

收稿日期：2016-04-03 **修回日期**：2016-09-19

基金项目：上海市教育委员会科研创新项目资助(B2-5308-13-0455)

作者简介：王 倩(1987—)，女，硕士研究生，研究方向为分子系统发育。E-mail:fengkuangbaozha@126.com

通信作者：李晨虹，E-mail:chli@shou.edu.cn

以及生态型的分类合理性仍备受争议[32,39-41]。线粒体和 AFLP 数据显示短颌鲚和刀鲚难以区分,因此现在一致认为短颌鲚和刀鲚是同一物种[32,39-41]。因为它们可能有不同的产卵场[42],已有相关研究致力于辨别这些不同来源的个体[43-44]。

LI 等[26]开发了 1 265 个单拷贝核基因标记。本研究有两个目标:(1)根据 LI 等的方法,开发可用于鲱形目鱼类的 SNP 分子标记,结合基因富集技术以及 BFD* 的方法进行物种界定;(2)富集这些 SNP 位点来区分短颌鲚和沿海以及河流中的刀鲚,以此来判断它们之间的遗传差异,对鲚属进行物种界定。

1　材料与方法

1.1　样本

采集了 6 尾刀鲚(*C. nasus*),其中 3 尾来自于上海沿岸,3 尾来自于长江主干的靖江;3 尾"短颌鲚"(*C. brachygnathus*)来自于模式产地岳阳洞庭湖(图 1 和表 1)。另外,从上海采集了 4 尾凤鲚(*C. mystus*)做为本研究的外群。

图 1　样本采集地点
Fig. 1　The site of samples

表 1　采样地点及水体
Tab. 1　Sampling localities and water bodies

样本编号 Sample ID	物种 Species	地点 Locality	水体 Water body
DT1	*C. brachygnathus*	岳阳	洞庭湖
DT2	*C. brachygnathus*	岳阳	洞庭湖
DT3	*C. brachygnathus*	岳阳	洞庭湖
JJ1	*C. nasus*	靖江	长江
JJ2	*C. nasus*	靖江	长江
JJ3	*C. nasus*	靖江	长江
SH1	*C. nasus*	上海	东海沿岸
SH2	*C. nasus*	上海	东海沿岸
SH3	*C. nasus*	上海	东海沿岸
SH4	*C. mystus*	上海	东海沿岸
SH5	*C. mystus*	上海	东海沿岸
SH6	*C. mystus*	上海	东海沿岸
SH7	*C. mystus*	上海	东海沿岸

1.2　目标基因的筛选和探针的设计

根据 LI 等[26]的描述,选用单拷贝核基因标记作为富集的目标基因。因为没有刀鲚基因组,因此 RNA 探针是根据斑马鱼的 1 265 个核编码基因序列设计,总计 358 798 bp。目标区域从 142 bp 到 2 462 bp 不等,平均片段大小是 222 bp。根据这些目标序列合成 120 bp 的 RNA 探针(MYBaits 试剂盒; MYcroarray, Ann Arbor, Michigan, USA),每个相邻探针之间留有 60 bp 的重叠区域,以提高基因富集效率。

1.3　文库制备、基因富集及混合测序

文库制备和基因富集按照 LI 等[26]的实验步骤。每个样本富集两次,即第一次富集的产物做为第二次富集的模板。通过条形码引物扩增将 8 bp 的 DNA 条形码添加到每个样本上,然后将所有样本等摩尔混合用 MiSeq(Illumina, Inc, San Diego, CA, USA)平台进行 2 ×151 bp 测序。

1.4 生物信息学分析

根据 8 bp 的 DNA 条形码将属于每个样本的原始序列(reads)分开,DNA 条形码没有错配的 reads 才可以用作下一步分析。接头序列和质量分数低于 20 的 reads 用 Cutadapt1.1 软件去掉[45-46]。对每个样本用 ABySS 1.3.4 软件进行从头组装[47]。ABySS 的 kmer 值设置为 63。用自编 Perl 脚本以斑马鱼为参考序列与组装好的重叠群比对,获得组装好的每个样本每个基因的重叠群,然后将每个样本的重叠群用 Geneious Pro v 5.6.2 软件(Biomatters, Auckland, New Zealand available from http://www.geneious.com/)进一步组装,这时选择 Geneious 中从头组装的选项。对组装好的重叠群进行手动检查。挑选与斑马鱼序列最相似的重叠群,仔细检查以确保是直系同源基因。筛选目标基因符合以下 3 个条件:(1)每个个体有大于 2 个等位基因的重叠群会被去除,因为二倍体生物有超过两个等位基因表明可能旁系同源序列被错误地组装到了一起;(2)如果一个位点在某一群体的所有个体中都没有数据的话,那么这个位点也会被舍弃,也就是说每个位点在每个群体至少要有一个样本中有数据;(3)无法对齐的序列的两端会被切除。经过以上数据清理后,每个基因输出一条共有序列用作原始 reads 映射的一个参考序列。

1.5 序列(reads)映射和 SNPs 位点的获取

所有样本原始序列经过修整,然后用 BWA-0.7.10 软件映射到共有参考序列上[48],重复序列(PCR 扩增产生的重复)用 Picard 1.118 软件(http://picard.sourceforge.net)标记出来。用 GATK 3.2 软件对碱基测序质量重新校正、序列重新比对,根据标准过滤参数对 13 个样本同时进行基因分型,获得 SNP 基因型数据[49]。所有分析步骤按照 GATK 最优的执行方法进行操作[50-51]。用自编 Perl 脚本将 SNP vcf 文件转换成可以被 BEAST 2.1.3 识别的 Nexus 文件[52],和 Structure 2.3.4 识别的输入文件[53]。因为 BEAST 和 Structure 软件的分析需要假定位点是连锁平衡的,用自编脚本选择每个目标区域最好的 SNP 位点(最高的 SNP 映射得分,最少丢失数据)用于后面的数据分析。数据预处理、参考序列组装、序列映射及 SNPs 获取的过程如图 2 所示。

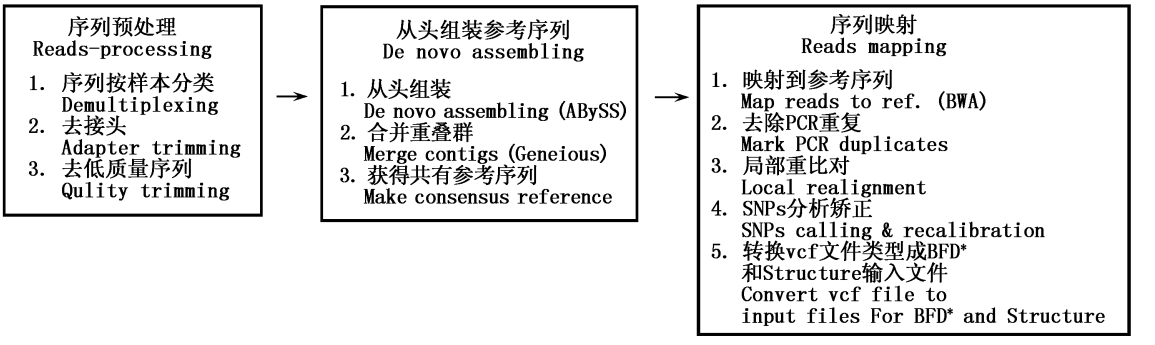

图 2 根据参考序列组装原始数据以及获取 SNPs 数据流程图

Fig. 2 The workflow of the raw data preprocessing, reference assembling and SNPs calling

1.6 种群结构分析

用 STRUCTURE 分析来检测样本可以分为几个不同的遗传群体。测试值从 1 到 4(K = 1 到 4),每个 K 值的运行次数、迭代次数均采用默认设置。用 Structure Harvester 软件(http://taylor0.biology.ucla.edu/structureHarvester/)来评估增加 K 值时,Delta K 的变化[54],根据此值来选择最佳的 K 值[60]。

1.7 物种界定

BFD* 与其他物种界定的方法相比优点是可以利用全基因组共显性等位基因数据[2]。用 BEAST 2 的插件 SNAPP 的改进版进行 BFD* 物种界定[52]。根据 BFD* 维基操作说明(http://www.beast2.org/wiki/index.php/BFD*, 12/22/2014)安装程序、建立 XML 文件及运行,详细分析原理请见“SNAPP 分析如何更简单地处理丢失

数据和路径抽样”(http://blog.beast2.org/tag/snapp/,12/22/2014)。共运行了 24 次路径抽样(每次包括 100 000 MCMC 步骤,burnin 设置为 10 000代)来估计边际似然值。

2　结果

2.1　测序结果和 SNPs 获取

经过去接头和测序质量分数校正,共获得了 600 万条序列,每个样本获得 301 046 ~ 1 366 562 条序列不等,不同群体间的样本数据量无明显差异(表 2)。用 ABySS 组装所有的序列,通过和斑马鱼的目标序列比较,获得刀鲚每个样本每个基因的重叠群,用 Geneious 合并重叠群,去除组装错误的重叠群和在某一种群所有样本均有缺失的数据,我们最终得到 698 条共有目标序列作为参考序列。使用 BWA 将每个样品的序列映射到参考序列上。每个样本映射到目标区域的序列从 57 350 到 675 411 条不等,目标序列平均覆盖深度从 17 × 到 197 × 不等。

表 2　测序结果统计
Tab. 2　Summary of sequencing results

样本编号 Sample ID	原始序列数目 No. of raw data	比对到目标基因上的序列数目 No. of data mapped on targets	平均覆盖深度 Average coverage depth
DT1	653 946	321 867	94
DT2	402 830	181 611	53
DT3	1 366 562	675 411	197
JJ1	332 820	153 747	45
JJ2	401 498	183 033	54
JJ3	448 904	208 993	61
SH1	301 046	139 085	41
SH2	142 968	57 350	17
SH3	475 904	212 485	62
SH4	370 458	155 744	46
SH5	327 984	129 199	38
SH6	385 642	160 670	47
SH7	490 492	234 896	69

对所有 13 个样本进行了序列分值校正、局部重比对、SNP 和 INDEL 的发掘与基因分型。经过应用默认参数的挑选,我们从 682 个重叠群中提取了 8 011 个 SNPs。最后,我们选择了每一个重叠群中映射分值最高的和缺失数据最少的,共 682 个 SNP 位点用于后续的分析。我们将 vcf 文件格式转换成了 nexus 格式,以及 STRUCTURE 分析的输入文件格式。

2.2　种群遗传结构分析

在 STRUCTURE 分析中,我们的目标是评估刀鲚和“短颌鲚”的遗传结构,因此没有使用 4 个凤鲚的样本,造成一些 SNP 位点变成了无变异的位点,所以我们删除了这些位点,从而使 SNPs 总数从 682 减少到 338。STRUCTURE 分析结果支持将 9 个个体分为两组:洞庭湖区的“短颌鲚”样本为一组,靖江和上海的刀鲚样本作为另一组(图 3b),而对靖江和上海的样本进一步分组得不到数据的支持(图 3a)。我们也对所有 13 个样本的 682 个 SNPs 位点进行了分析,这些样本中包括了凤鲚样本,结果相同,即刀鲚和“短颌鲚”分为截然不同的两组(图 4)。

2.3　Bayes factor 物种界定

采用新的 Bayes factor 种类划分方法——BFD* 检测了长江和沿海采集样本间是否存在遗传差异。用 BEAST 软件包运行 SNAPP 分析两种不同的情况,将来自靖江的样本和上海的样本作为一个物种,或将它们作为独立的物种。然后,我们用路径抽样计算了边际似然率。合并物种和区分独立物种的边缘似然值分别是 -10 123.5 和 -10 112.2。两种情况的贝叶斯因子是 11.3 (Bayes factor = 11.3)。根据 RAFTERY[55-56],当 3 < Bayes factor < 20 时,对替代方案的支持是有利的,因此,BFD* 分析结果表明长江和沿海地区采集的刀鲚之间存在着少量的遗传差异,但差异不显著。

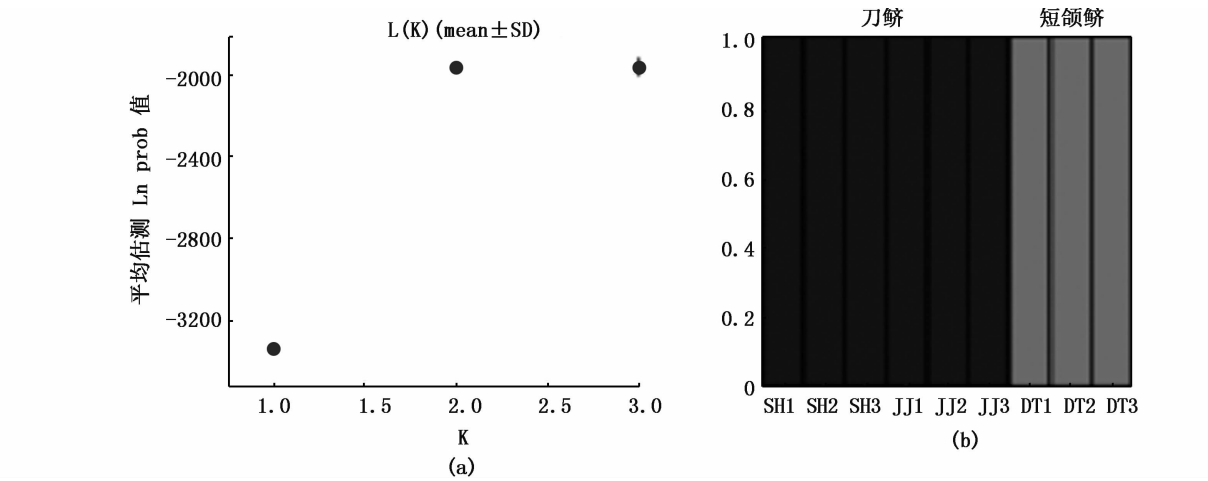

图 3 遗传结构分析

Fig. 3 Structure analyses

当 K 从 1 变为 2 时,估算的似然值增加了 1 378 单位,但是将样本分为更多组不被支持(a);9 个鲚属样本被分为 2 组:洞庭湖的一组,靖江和上海的样本为另一组(b)

Showed that the estimated likelihood of the data increased by 1378 units when K changed from one to two, while dividing the fish into more groups was not supported (a). The nine individuals of *Coilia* were clustered into two groups: fish from Lake Dongting as one group and samples from Jingjiang and Shanghai as the other group(b)

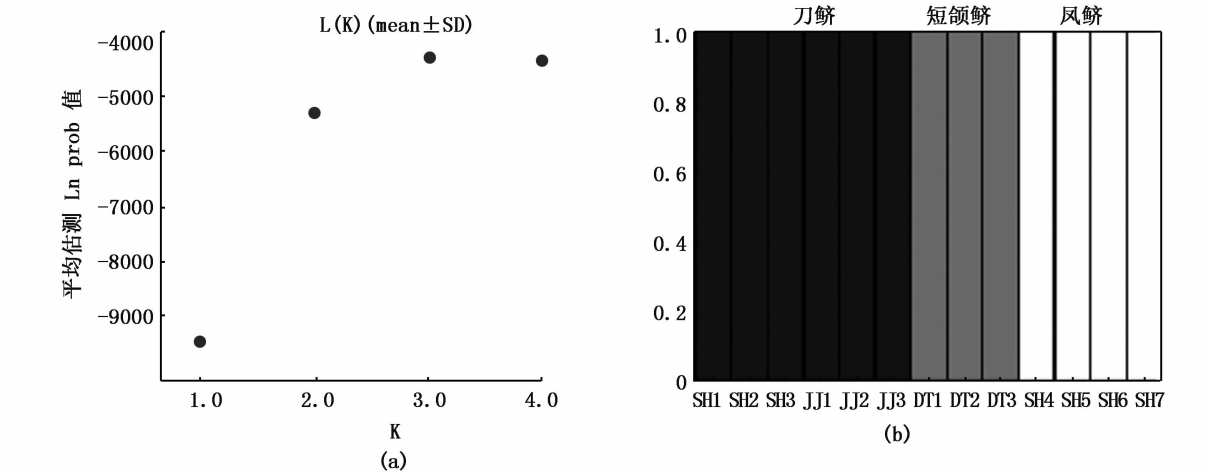

图 4 13 个个体的遗传结构分析,包括鲚属的两种鱼

Fig. 4 Structure analysis using all 13 fish, including both species of *Coilia*

3 讨论

本研究所用的目标基因是根据所有脊椎动物,包括鲨鱼、辐鳍鱼类、两栖类、爬行类、鸟类和哺乳动物开发的单拷贝核基因标记[26,57]。我们在鲚属中富集到了大部分标记基因。另外,我们还成功地将这些标记基因应用于鱼类(虾虎鱼,鲈鱼)、短吻鳄、鸟类和软骨鱼类的研究中,因此本研究所用的分子标记可以运用于富集大部分种类的脊椎动物。使用通用标记富集相比匿名标记有两大优点:首先,可以将富集到的序列与参考序列进行比较。比如在这次研究中,没有鲚属的参考基因组,可以用斑马鱼的序列作为参考序列来获得鲚属中的序列。其次,由于这些单拷贝基因在大多数的脊椎动物中比较保守,借此,可以获得不同物种的同源序列并重建系统发育关系,用来研究不同分类水平的系统发育问题。值得注意的是,通用标记并不意味着可以仅凭一组 RNA 探针可以富集到所有的脊椎动物的基因。应该用与目标物种亲缘关系最近的参考序

列来设计探针。比如,用斑马鱼的探针来富集鲚属的目标基因,用鸡的基因组序列来富集鸟类中的目标基因。

大多数物种界定的方法必须先建立基因树,因此只能利用少量的位点进行分析[57]。BFD* 通过将多物种溯祖的方法结合到 SNAPP 软件中[2],从而完全避免了对基因树的依赖,达到了真正意义上用基因组 SNPs 数据进行分析。BFD* 唯一的限制是需要预先设定样本属于哪一个类群,所以我们先用 STRUCTURE 将样本分类,这一策略也见于其他的研究中[58]。

STRUCTURE 个体之间的聚类分析,结果显示从洞庭湖中采得的"短颌鲚"样本与长江下游和沿海地区采得的刀鲚样本不同。洞庭湖的"短颌鲚"由于其具有较短的未覆盖到鳃边缘的上颌骨,因此被定为 *C. brachygnathus*,而长江中"刀鲚"的上颌骨可以覆盖过鳃的边缘[34,38]。后来在长江的其他区段以及相连接的湖中相继发现了具有上颌骨较短的生态型,因此有人认为 *C. brachygnathus* 已经扩散至这些水体中[36-37]。近年来分子研究多支持 *C. brachygnathus* 和 *C. nasus* 无法区分,因此,*C. brachygnathus* 和 *C. nasus* 被认为是同物异名[32,40-41]。然而,这些研究所用样本采自非典型的水域,比如太湖[32],或者只用了单个线粒体的位点,因此,这无法反映洞庭湖中的 *Coilia* 是 *C. brachygnathus* 还是 *C. nasus*。而我们根据基因组水平数据的研究结果表明洞庭湖的"短颌鲚"样本和其他地区的刀鲚样本在 STRUCTURE 聚类分析和 BFD* 模型选择的数据支持下,应该属于不同的物种。因此今后有必要从洞庭湖和其他区域采集更多样本进行基因组水平的分析,结合物种界定的方法,来进一步验证 *C. brachygnathus* 是否是有效的种。我们认为上颌骨的长度不应该作为鉴别物种的依据,因为上颌骨长度相似的样本在遗传上有时相距甚远[32,39,41]。

在春季,*C. nasus* 迁移到长江产卵,其中有些个体会更早地迁徙到河流中去产卵,有些个体则更多地停留在河口地区,因此它们可能有不同的产卵场,它们之间的基因交流可能会受到限制[42]。有研究表明栖居于河流和沿海地区的 *C. nasus* 的氨基酸组成不同[43-44],意味着它们之间可能存在着基因分化,但本研究 BFD* 分析显示将河流与沿海地区个体分成不同独立群体的支持度很低,这和 STRUCTURE 的分析结果是相同的。因此我们认为尽管长江中的刀鲚与沿海区域的刀鲚在空间与时间分布上存在差异,但根据我们基因组水平的遗传研究,认为应归为一个种群。本研究仅从 3 个采样点各选取了 3 个个体来研究它们之间的遗传差异,因此样本的数量是不够的,若要解决有关 *C. nasus* 复杂的系统分类问题需要从更广的区域采集更多的样本,比如陆封种群和其他地区的种群来研究它们之间的关系。不过,本研究的主要目标是结合基因富集、脊椎动物的通用标记以及 BFD* 方法来获得 SNPs 数据并进行物种界定研究,结果表明这一方法是可行的,为今后进一步研究刀鲚的分子系统学奠定了基础。

参考文献:

[1] EMERSON K J, MERZ C R, CATCHEN J M, et al. Resolving postglacial phylogeography using high-throughput sequencing [J]. Proceedings of the National Academy of Sciences of the United States of America, 2010, 107(37): 16196-16200.

[2] LEACHÉ A D, FUJITA M K, MININ V N, et al. Species delimitation using genome-wide SNP data [J]. Systemtic Biology, 2014, 63(4): 534-542.

[3] MILANO I, BABBUCCI M, CARIANI A, et al. Outlier SNP markers reveal fine-scale genetic structuring across European hake populations (*Merluccius merluccius*) [J]. Molecular Ecology, 2014, 23(1): 118-135.

[4] SATO T, NAKAGOME S, WATANABE C, et al. Genome-wide SNP analysis reveals population structure and demographic history of the ryukyu islanders in the southern part of the Japanese archipelago[J]. Molecular Biology and Evolution, 2014, 31(11): 2929-2940.

[5] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3 000 shared controls[J]. Nature, 2007, 447(7145): 661-678.

[6] VIQUEZ-ZAMORA M, VOSMAN B, VAN DE GEEST H, et al. Tomato breeding in the genomics era: insights from a SNP array[J]. BMC Genomics, 2013, 14: 354.

[7] WILLIAMS L M, MA X, BOYKO A R, et al. SNP identification, verification, and utility for population genetics in a non-model genus[J]. BMC Genetics, 2010, 11: 32.

[8] GHOLAMI M, ERBE M, GÄRKE C, et al. Population genomic analyses based on 1 million SNPs in commercial egg layers[J]. PLoS One, 2014, 9(4): e94509.

[9] GOHLI J, LEDER E H, GARCIA-DEL-REY E, et al. The evolutionary history of Afrocanarian blue tits inferred from

genomewide SNPs[J]. Molecular Ecology, 2015, 24(1): 180-191.

[10] OGDEN R, GHARBI K, MUGUE N, et al. Sturgeon conservation genomics: SNP discovery and validation using RAD sequencing[J]. Molecular Ecology, 2013, 22(11): 3112-3123.

[11] YU H H, XIE W B, LI J, et al. A whole-genome SNP array (RICE6K) for genomic breeding in rice [J]. Plant Biotechnology Journal, 2014, 12(1): 28-37.

[12] HELYAR S J, HEMMER-HANSEN J, BEKKEVOLD D, et al. Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges[J]. Molecular Ecology Resources, 2011, 11 (s1): 123-136.

[13] XIA J H, WAN Z Y, NG Z L, et al. Genome-wide discovery and in silico mapping of gene-associated SNPs in Nile tilapia [J]. Aquaculture, 2014, 432: 67-73.

[14] YU Y, WEI J K, ZHANG X J, et al. SNP discovery in the transcriptome of white Pacific shrimp *Litopenaeus vannamei* by next generation sequencing[J]. PLoS One, 2014, 9(1): e87218.

[15] HELYAR S J, LIMBORG M T, BEKKEVOLD D, et al. SNP discovery using next generation Transcriptomic sequencing in Atlantic herring (*Clupea harengus*)[J]. PLoS One, 2012, 7(8): e42089.

[16] MONTES I, CONKLIN D, ALBAINA A, et al. SNP discovery in European anchovy (*Engraulis encrasicolus*, L) by high-throughput transcriptome and genome sequencing[J]. PLoS One, 2013, 8(8): e70051.

[17] BAIRD N A, ETTER P D, ATWOOD T S, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers [J]. PLoS One, 2008, 3(10): e3376.

[18] WAGNER C E, KELLER I, WITTWER S, et al. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation[J]. Molecular Ecology, 2013, 22(3): 787-798.

[19] BI K, LINDEROTH T, VANDERPOOL D, et al. Unlocking the vault: next-generation museum population genomics[J]. Molecular Ecology, 2013, 22(24): 6018-6032.

[20] CLARKE W E, PARKIN I A, GAJARDO H A, et al. Genomic DNA enrichment using sequence capture microarrays: a novel approach to discover sequence nucleotide polymorphisms (SNP) in *Brassica napus* L[J]. PLoS One, 2013, 8(12): e81992.

[21] COSART T, BEJA-PEREIRA A, CHEN S Y, et al. Exome-wide DNA capture and next generation sequencing in domestic and wild species[J]. BMC Genomics, 2011, 12: 347.

[22] CARPENTER M L, BUENROSTRO J D, VALDIOSERA C, et al. Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries[J]. The American Journal of Human Genetics, 2013, 93(5): 852-864.

[23] MARIAC C, SCARCELLI N, POUZADOU J, et al. Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies [J]. Molecular Ecology Resources, 2014, 14(6): 1103-1113.

[24] LIU G, WEI X M, CHEN R, et al. A novel mutation of the *SLC25A13* gene in a Chinese patient with citrin deficiency detected by target next-generation sequencing[J]. Gene, 2014, 533(2): 547-553.

[25] COSART T, BEJA-PEREIRA A, CHEN S Y, et al. Exome-wide DNA capture and next generation sequencing in domestic and wild species[J]. BMC Genomics, 2011, 12: 347.

[26] LI C H, HOFREITER M, STRAUBE N, et al. Capturing protein-coding genes across highly divergent species [J]. BioTechniques, 2013, 54(6): 321-326.

[27] MANDEL J R, DIKOW R B, FUNK V A, et al. A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae[J]. Applications in Plant Sciences, 2014, 2(2): 1300085.

[28] ENCE D D, CARSTENS B C. SpedeSTEM: a rapid and accurate method for species delimitation [J]. Molecular Ecology Resources, 2011, 11(3): 473-480.

[29] GRUMMER J A, BRYSON Jr R J, REEDER T W. Species delimitation using Bayes factors: simulations and application to the *Sceloporus scalaris* species group (Squamata: Phrynosomatidae)[J]. Systematic Biology, 2014, 63(2): 119-133.

[30] O'MEARA B C. New heuristic methods for joint species delimitation and species tree inference [J]. Systematic Biology, 2010, 59(1): 59-73.

[31] PONS J, BARRACLOUGH T G, GOMEZ-ZURITA J, et al. Sequence-based species delimitation for the DNA taxonomy of undescribed insects[J]. Systematic Biology, 2006, 55(4): 595-609.

[32] YANG Z H, RANNALA B. Bayesian species delimitation using multilocus sequence data [J]. Proceedings of the National Academy of Sciences of the United States of America, 2010, 107(20): 9264-9269.

[33] BRYANT D, BOUCKAERT R, FELSENSTEIN J, et al. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis [J]. Molecular Biology and Evolution, 2012, 29(8): 1917-1932.

[34] WHITEHEAD P J P. FAO species catalogue: vol. 7. Clupeoid fishes of the world: an annotated and illustrated catalogue of the herrings, sardines, pilchards, sprats, shads, anchovies and wolf-herrings [R]. London, UK: FAO Species Catalogue, 1985: 125-127.

[35] WONGRATANA T. Systematics of clupeoid fishes of the indo-pacific region [M]. London: University of London,

1980.

[36] 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.
YUAN C M, LIN J B, QIN A L, et al. On the classification history and status quo of genus Colia in China[J]. Nanjing University (Natural Sciences), 1976(2): 1-12.

[37] 袁传宓, 秦安黔, 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): 67-82.
YUAN C M, QIN A Q, LIU R H, et al. On the clasification of the anchovies, Coilia, from the lower Yangtze River and the southeast coast of China[J]. Nanjing University (Natural Sciences), 1980(3): 67-82.

[38] KREYENBERG W, PAPPENHEIM P. Ein beitrag zur kenntnis der fische der jangtze und seiner zuflüsse [J]. Sitzungsberichte der Gesellschaft Naturforschender Freunde zu Berlin, 1908, 19: 95-109.

[39] CHENG 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(4/6): 544-552.

[40] 程万秀, 唐文乔. 长江刀鲚不同生态型间的某些形态差异[J]. 动物学杂志, 2011(5): 33-40.
CHENG W X, TANG W Q. Some phenotypic varieties between different ecotypes of *Coilia nasus* in Yangtze River [J]. Chinese Journal of Zoology, 2011(5): 33-40.

[41] 唐文乔, 胡雪莲, 杨金权. 从线粒体控制区全序列变异看短颌鲚和湖鲚的物种有效性[J]. 生物多样性, 2007, 15(3): 224-231.
TANG W Q, HU X L, YANG J Q. Species validities of *Coilia brachygnathus* and *C. nasus* taihuensis based on sequence variations of complete mtDNA control region[J]. Biodiversity Science, 2007, 15(3): 224-231.

[42] 姜涛. 长江及其周边海域刀鲚矢耳石的形态和微化学研究[D]. 上海: 上海海洋大学, 2011.
JIANG T. Morphometric and microchemical studies on sagittal otoliths of *Coilia nasus* collected from the Yangtze River and adjacent sea areas [J]. Shanghai: Shanghai Ocean University, 2011.

[43] 徐钢春, 顾若波, 张呈祥, 等. 刀鲚两种生态类群"江刀"和"海刀"鱼肉营养组成的比较及品质的评价[J]. 海洋渔业, 2009, 31(4): 401-409.
XU G C, GU R B, ZHANG C X, et al. Comparison and evaluation of nutrient composition of two ecological groups of Japanese grenadier anchovy-river-anchovy and sea-anchovy [J]. Marine Fisheries, 2009, 31(4): 401-409.

[44] 张呈祥, 徐钢春, 顾若波, 等. 2 种生态类群刀鲚(江刀和海刀)肌肉游离氨基酸的差异[J]. 长江大学学报(自科版), 2010, 7(2): 32-35.
ZHANG C X, XU G C, GU R B, et al. Comparative study on free amino acid composition and content between two ecological groups of Japanese grenadier anchovy, *Coilia nasus* (river-anchovy and sea-anchovy) [J]. Journal of Yangtze University (Natural Science Edition), 2010, 7(2): 32-35.

[45] FALUSH D, STEPHENS M, PRITCHARD J K. Inference of population structure using multilocus genotype data: dominant markers and null alleles[J]. Molecular Ecology Notes, 2007, 7(4): 574-578.

[46] MARTIN M. Cutadapt removes adapter sequences from high-throughput sequencing reads[J]. EMBnet Journal, 2011, 17: 10-12.

[47] SIMPSON J T, WONG K, JACKMAN S D, et al. ABySS: a parallel assembler for short read sequence data[J]. Genome Research, 2009, 19(6): 1117-1123.

[48] LI H, DURBIN R. Fast and accurate short read alignment with burrows-wheeler transform[J]. Bioinformatics, 2009, 25(14): 1754-1760.

[49] MCKENNA A, HANNA M, BANKS E, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data[J]. Genome Research, 2010, 20(9): 1297-1303.

[50] DEPRISTO M A, BANKS E, POPLIN R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data[J]. Nature Genetics, 2011, 43(5): 491-498.

[51] VAN DER AUWERA G A, CARNEIRO M O, HARTL C, et al. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline[J]. Current Protocols in Bioinformatics, 2013, 43: 10-13.

[52] BOUCKAERT R, HELED J, KÜHNERT D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis[J]. PLoS Computational Biology, 2014, 10(4): e1003537.

[53] PRITCHARD J K, STEPHENS M, DONNELLY P. Inference of population structure using multilocus genotype data[J]. Genetics, 2000, 155(2): 945-959.

[54] EARL D A, VONHOLDT B. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method[J]. Conservation Genetics Resources, 2012, 4(2): 359-361.

[55] EVANNO G, REGNAUT S, GOUDET J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study[J]. Molecular Ecology, 2005, 14(8): 2611-2620.

[56] RAFTERY A E. Hypothesis testing and model selection [M]//GILKS W R, RICHARDSON S, SPIEGELHALTER D. Markov Chain Monte Carlo in Practice. London: Chapman and Hall, 1996: 163-188.

[57] CARSTENS B C, PELLETIER T A, REID N M, et al. How to fail at species delimitation[J]. Molecular Ecology, 2013, 22(17): 4369-4383.

[58] SATLER J D, CARSTENS B C, HEDIN M. Multilocus species delimitation in a complex of morphologically

conserved trapdoor spiders (Mygalomorphae, Antrodiaetidae, Aliatypus)[J]. Systematic Biology, 2013, 62(6): 805-823.

Developing SNP markers for *Coilia* and its application in species delimitation

WANG Qian, CHENG Fangyuan, LI Chenhong

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

Abstract: Single nucleotide polymorphism (SNP) markers coupled with the next-generation sequencing technology are powerful tools for studying population genetics, as well as for species delimitation. The popular ways of obtaining SNP data include RNA sequencing and restriction site associated DNA (RAD) sequencing. However, both methods have limitations. For example, RNA sequencing requires fresh tissue samples, whereas RAD sequencing targets anonymous loci usually not transferable across species. In comparison with these methods, target gene capture could be used to enrich homologous fragments across divergent species and even from degraded DNA. We showed that by targeting a set of universal single-copy nuclear gene markers of vertebrates, we could retrieve thousands of SNPs from the Japanese grenadier anchovy (*Coilia nasus*) sampled from three locations: coastal region close to the estuary of the Yangtze River, the mainstream of the river and Lake Dongting, a lake connected to the middle reaches of the Yangtze River. We analyzed hundreds of representative SNPs from the data using STRUCTURE and Bayes factor delimitation (* with genomic data; BFD*), a new Bayesian species delimitation tool. We found that the fish from Lake Dongting are genetically different from the fish of other sample locations. We also observed marginal difference between fish collected from the coastal region and the mainstream of the Yangtze River (Bayes factor = 11.3). Our study demonstrated that gene capture could be used to generate SNP data for species delimitation applying new analytic tools such as BFD*.

Key words: target enrichment; universal gene markers; species delimitation; SNPs; Bayes factor delimitation (* with genomic data, BFD*); *Coilia*