水产学报, 2024, 48(12): 129103

JOURNAL OF FISHERIES OF CHINA

DOI: 10.11964/jfc.20230914153

Science Press

基于“宁芯 3 号”基因组育种芯片和机器学习的大黄鱼种质资源精准鉴定

赵　吉[1,2]，冯苗胜[1]，柯巧珍[1,2]，王家迎[1,2]，江汀森[1]，吴雄飞[3]，彭士明[4]，白玉麟[1]，沈伟良[3]，周　涛[1]，濮　菲[1]，徐　鹏[1*]

(1. 厦门大学海洋与地球学院，海水养殖生物育种全国重点实验室，福建 厦门　361101；
2. 厦门大学环境与生态学院，近海海洋环境科学国家重点实验室，福建 厦门　361101；
3. 宁波市海洋与渔业研究院，浙江 宁波　315048；
4. 中国水产科学研究院东海水产研究所，上海　200090)

摘要：为了高效保护、管理以及利用大黄鱼种质资源，迫切需要开发精准的大黄鱼遗传种质鉴定技术。基于前期开发的大黄鱼“宁芯 3 号”55K 液相基因分型芯片，本研究对中国沿海野生群体、闽浙养殖群体和多个选育系共计 21 个大黄鱼群体进行遗传种质鉴定。群体遗传学分析结果揭示大黄鱼群体可划分为南海群体、闽东群体和岱衢群体，其中南海群体遗传分化最为显著。基于机器学习的大黄鱼群体分类结果显示，未知大黄鱼个体所属地理群体鉴定准确率大于 99%。未知大黄鱼个体所属遗传改良选育系也具有极高的鉴定准确率，例如经过 3 代选育的抗刺激隐核虫新品系 GS3F3，基于神经网络的鉴定精确率可以达到 99%。研究表明，利用“宁芯 3 号”芯片和机器学习方法可快速实现大黄鱼种质的精准鉴定。本研究为大黄鱼种质资源精准鉴定和种质管理、育种材料和新品种知识产权保护等提供了有效的工具和解决方案，也可为其他水产生物种质资源鉴定提供借鉴。

关键词: 大黄鱼；宁芯 3 号；种质鉴定；机器学习

中图分类号: S 917.4；TP 181　　**文献标志码:** A

大黄鱼 (*Larimichthys crocea*) 隶属石首鱼科 (Sciaenidae) 黄鱼属 (*Larimichthys*)，主要分布于我国东南沿海，曾经是我国重要的海洋捕捞对象之一，20 世纪 70 年代的过度捕捞使得大黄鱼野生资源逐渐枯竭[1-2]。20 世纪 90 年代，大黄鱼人工繁殖和养殖技术的攻克使得大黄鱼养殖产业快速发展，2023 年养殖产量已达 28.1 万 t。种质指遗传物质相对一致的同种生物的个体群，种质资源指一类生物或一个地区所拥有的种质的统称[3]，也即某种生物在长期的自然选择或人工选育过程中形成的具有遗传差异的种质集合。目前，通过审定的 4 个大黄鱼新品种主要针对生长、耐低温性状，对抗虫、抗菌、耐低氧等重要经济性状的遗传选育也在进行中，这些大黄鱼新品种、选育品系和东南沿海的野生群体构成了极为丰富的大黄鱼种质资源。种质资源是现代种业发展的物质基础，因此建立科学有效的种质资源鉴定体系对于大黄鱼种业的健康可持续发展具有重大意义，同时也能对大黄鱼种质资源进行更好的利用和保护[2]。

收稿日期：2023-09-09　　**修回日期：**2023-10-24

资助项目：国家重点研发计划 (2022YFD2401002)；国家杰出青年科学基金 (32225049)；福建省种业创新与产业化项目 (2021FJSCZY01)

第一作者：赵吉 (照片)，从事大黄鱼遗传选育研究，E-mail：zhaoji@xmu.edu.cn

通信作者：徐鹏，从事鱼类遗传学与遗传育种研究，E-mail：xupeng77@xmu.edu.cn

种质鉴定实质是对种质资源进行分类，通过形态学观察进行种质资源鉴定较为普遍，例如鲤(*Cyprinus carpio*)的不同品种，荷包红鲤(*C. carpio* var. *wuyuanensis*)通体鲜红、腹部肥大，而黄河鲤(*C. carpio haematopterus* Temminck et Schlegel)通体金黄、体型修长。虽然形态学方法具有简单直观的优点，但是只有极大的种质遗传差异才可导致可观察的形态学变异，因此基于形态学方法无法进行不存在表型差异的种质区分。随着生物技术的发展，遗传标记逐渐被应用到种质鉴定领域，其中分子标记因其不受环境限制、不影响目标性状等诸多优点而广为应用[4-5]。例如扩增片段长度多态性(AFLP)曾被应用于鲤的种质鉴定[6]，微卫星(SSR)被应用于斑点叉尾鮰(*Ictalurus punctaus*)的种质鉴定[7]。早期的分子标记具有数量少、多态性低等特性，因此其主要应用于具有相似表型特征的不同地理种质资源的鉴定[8]。随着单核苷酸多态性分子标记(SNP)的出现，其因数量多、多态性高等特点被广泛应用于动植物的遗传研究，例如群体遗传结构解析、遗传基础解析、遗传选育、种质鉴定等[9-11]。因此，基于SNP分子标记的高分辨率特性，不仅可以鉴定同一物种的不同地理种质资源，同时也可以鉴定人工遗传改良过程中形成的不同品种或选育系。

机器学习广义上可以理解为赋予机器学习能力进而实现某些功能，像人类一样通过学习思考总结规律来指导行动，从实践上可以将机器学习理解为一种基于大量已知数据进行模型训练和预测的方法。机器学习一般分为监督学习和无监督学习，分类问题即为监督学习范畴，核心思想是对已知标注的大量样本进行学习，建立解释模型，进而对未知样本的所属类别进行预测。愈来愈多的遗传学科研工作者使用机器学习处理高通量的基因型数据，例如在全基因组选择领域，使用机器学习预测候选选育群体的基因组育种值[12-13]。机器学习也被应用到品系鉴定领域，例如Schiavo等[14]使用机器学习方法寻找能鉴定7个意大利猪品系的显著SNP标记。Miao等[15]开发了一个在线网站，以鉴定分布于全球的猪的不同种质，其核心算法基于偏最小二乘法(PLS)分类器。机器学习将在种质资源鉴定领域发挥越来越重要的作用。

大黄鱼“宁芯3号”55K液相基因分型芯片是继2019年“宁芯1号”[16]、2020年“宁芯2号”[17]固相基因分型芯片之后，本研究团队对“宁芯”系列大黄鱼基因组育种芯片进行优化升级的液相基因分型芯片，它采用了靶向测序基因分型(genotyping by targeted sequencing，GBTS)技术，可实现高通量、低成本、高精度、高灵活度的SNP基因分型[18]。目前，“宁芯3号”已成功应用于大黄鱼遗传解析和育种研究，已被多个大黄鱼育种团队应用[19-22]。为了探究“宁芯3号”对大黄鱼种质资源精准鉴定的可行性和效果，本研究收集不同地理群体的重测序数据和基于“宁芯3号”基因分型选育系的基因型数据，利用机器学习方法对不同地理群体和选育系进行种质资源鉴定。

1　材料与方法

1.1　大黄鱼群体样本收集

本研究共使用21个大黄鱼群体，其中9个为野生群体，12个为养殖群体。所有涉及大黄鱼群体采样实验流程严格按照《厦门大学实验动物管理办法》执行，采集时间为2018—2023年。野生群体的采样点从南到北分别为徐闻(XW)、湛江(ZJ)、惠州(HZ)、东山(DS)、厦门(XM)、平潭(PT)、福鼎(FD)、大嵛山(DYS)、舟山(ZS)。养殖群体的采样点为宁德、霞浦、福鼎、宁波，其中宁德包含5个群体，分别为蕉城区的养殖群体(JC)、GS3F3[抗刺激隐核虫(*Cryptocaryon irritans*)选育系F_3]、GS3RF3(抗刺激隐核虫平行选育系F_3)、GS6F2[抗变形假单胞菌(*Pseudomonas plecoglossicida*)选育系F_2]、GS8F1[抗溶藻弧菌(*Vibrio alginolyticus*)选育系F_1]以及GS10F1(强游泳能力选育系F_1)，霞浦包含3个来源于不同养殖场的群体(XP1、XP2、XP3)，福鼎包含2个来源于不同养殖场的群体(FD1、FD2)，宁波包含1个当地的养殖品系(NB)。来自宁德的5个GS选育系是厦门大学联合宁德富发水产有限公司基于全基因组选择技术(genomic selection，GS)培育的新品系。宁波群体样本来自宁波市海洋与渔业研究院。以上群体的简介见表1。采样过程中操作人员严格遵守厦门大学动物实验伦理规范，并按照厦门大学海洋与地球学院动物实验伦理委员会制定的规章制度执行。

1.2　不同群体基因型数据处理

本研究中野生群体所有个体的基因型数据通

表 1　本研究中所使用的大黄鱼群体种质资源简介

Tab. 1　Details of population of *L. crocea*

地理位置 site	群体缩写 abbr.	是否野生 wild or not	数量/尾 number of individuals
徐闻 Xuwen	XW	是	12
湛江 Zhanjiang	ZJ	是	25
惠州 Huizhou	HZ	是	20
东山 Dongshan	DS	是	13
厦门 Xiamen	XM	是	16
平潭 Pingtan	PT	是	15
宁德 Ningde	JC	否	20
宁德 Ningde	GS3F3	否	579
宁德 Ningde	GS3RF3	否	30
宁德 Ningde	GS6F2	否	490
宁德 Ningde	GS8F1	否	291
宁德 Ningde	GS10F1	否	242
霞浦 Xiapu	XP1	否	8
霞浦 Xiapu	XP2	否	8
霞浦 Xiapu	XP3	否	8
大嵛山 Dayushan	DYS	是	5
福鼎 Fuding	FD	是	28
福鼎 Fuding	FD1	否	8
福鼎 Fuding	FD2	否	8
宁波 Ningbo	NB	否	830
舟山 Zhoushan	ZS	是	29

注：GS3F3缩写代表抗刺激隐核虫GS选育系 F_3，GS3RF3缩写代表抗刺激隐核虫平行GS选育系 F_3，GS6F2缩写代表抗变形假单胞菌GS选育系 F_2，GS8F1 缩写代表抗溶藻弧菌GS选育系 F_1，GS10F1缩写代表强游泳能力GS选育系 F_1。

Notes: GS3F3 represents the genomic selection breeding line (F_3 generation) against *C. irritans*. GS3RF3 represents the repeated genomic selection breeding line (F_3 generation) against *C. irritans*. GS6F2 represents the genomic selection breeding line (F_2 generation) against *P. plecoglossicida*. GS8F1 represents the genomic selection breeding line (F_1 generation) against *V. alginolyticus*. GS10F1 represents the genomic selection breeding line (F_1 generation) having strong swimming performance.

过基因组重测序获得，养殖群体的所有个体的基因型数据通过“宁芯 3 号”获得，所有基因型数据均以标准格式存储于海水养殖生物育种全国重点实验室建立的大黄鱼育种数据库中。通过调取所需实验大黄鱼群体的基因型数据，利用 bcftools 软件获取各群体基因型文件的共有 SNP 位点，并使用该软件将所有群体的基因型数据合并在一起。使用 plink 软件对合并基因型数据集进行共有 SNP 位点提取和质量控制。质量控制条件为 geno<0.01(过滤掉 SNP 分型率低于 99% 的 SNP 位点)，mind<0.1(过滤掉个体分型率低于 90% 的个体)，maf<0.05(过滤掉最小等位基因频率低于 0.05 的 SNP 位点)[23]。质量控制后的基因型数据用于后续分析。

1.3　群体遗传学分析

使用 R 包 hierfstat 计算观测杂合度 (observed heterozygosity，H_o)、群体内遗传多样性 (gene diversity，H_s)[24]。使用 plink 软件计算自交系数 (F)。使用 vcftools 软件计算以整个染色体为窗口的核苷酸多样性 (π)。

使用 3 种方法进行群体遗传结构解析，分别为主成分分析 (principal component analysis，PCA)、遗传混合分析 (admixture analysis) 和系统发育树分析 (phylogenetic tree analysis)。主成分分析使用 plink 软件执行，遗传混合分析使用 admixture 软件执行[25]，系统发育树分析使用 python3 的 vcf-kit 库执行 (邻接法)。主成分分析和遗传混合分析结果使用 R 包 ggolot2 进行可视化，系统发育树 Newick 格式文件使用 iTOL 在线工具进行可视化。

1.4　大黄鱼种质鉴定

使用 4 种机器学习方法进行大黄鱼不同群体的种质鉴定，分别为卷积神经网络 (Neural Network，NN)、K 近邻 (K Nearest Neighbors，KNN)、逻辑斯蒂回归 (Logistic Regression，LR) 以及随机森林 (Random Forest，RF)。使用 python 库 keras 实现 NN 算法，具体网络搭建参数简介如下，神经网络由 3 个全连接层组成，第 1 层有 64 个节点，第 2 层有 32 个节点，第 3 层的节点数等于类别数，每一个全连接层后使用一个非线性激活函数，最后一层使用 softmax 激活函数，前两层使用线性分段激活函数 relu。训练过程中采用二值交叉熵损失函数，优化器采用 Adam，学习率设置为 0.000 1，训练迭代次数设置为 20 次。使用 python 库 sklearn 实现 KNN、LR 和 RF 算法，参数设置均为默认。

采用 5 折交叉验证分析不同机器学习方法的分类能力，即将数据集随机划分为 5 份，依次将其中 1 份作为测试集，剩余的为训练集；利用训练集建立模型，并对测试集进行预测，整个过程重复 4 次。评价指标采用基于混淆矩阵的 4 个统计量，针对整个模型的准确率 (Accuracy)，以及针对每个分类的精确率 (Precision)、召回率 (Recall)

和 F1-score。准确率等于所有判断正确的结果占总观察值的比重，精确率等于模型预测 Positive 的个体中预测正确的个体所占比重，召回率等于真实值是 Positive 的个体中预测正确的个体所占的比重，F1-score 同时考虑精确率和召回率，等于 2×Precision×Recall / (Precision + Recall)。同时计算 3 个指标 (Precision、Recall、F1-score) 的宏平均 (Macro average，等于每个类别的加权平均)，以及加权平均 (Weighted average)。计算上述指标的平均值和方差，使用 Python 库 matplotlib 进行混淆矩阵的可视化。

2　结果

2.1　群体遗传参数

所有群体的基因型文件经过合并质控等步骤后，产生一个包含 2 685 个个体和 20 917 个 SNP 的数据集，SNP 检出率 (call rate) 为 99.65%。对不同群体的 H_e (观测杂合度)、H_s (群体内遗传多样性)、F (自交系数)、π (核苷酸多样性) 进行计算，结果见表 2。不同群体 H_e 指标的范围为 0.274 4~0.333 5，H_s 指标为 0.293 4~0.336 8，F 指标为 −0.0190~0.1620，π 指标为0.00000630~0.00000722。由表 2 还可发现，与野生大黄鱼群体相比，养殖群体具有更高的观测杂合度，更高的种内遗传多样性，更低的自交系数，以及更高的核苷酸多样性 (图 1)。通常野生群体具有较高的遗传多样性，但是本研究却得到了相反的结果，这可能由以下原因造成：首先是计算上述指标使用的 SNP 个数较少，而且过滤掉一些稀有等位基因，导致野生群体特有的等位基因并未被考虑在内；其次是野生个体的数量较少，在统计计算遗传多样性时，会存在较大的误差；大黄鱼野生资源在 20 世纪面临枯竭，过度捕捞导致种群规模降低，有效群体减少，这也可能导致野生大黄鱼遗传多样性降低。

表 2　各群体遗传结构参数

Tab. 2　Basic population genetic parameters of different populations

群体名称 population	观测杂合度 H_o	群体内遗传多样性 H_s	自交系数 F	核苷酸多样性 π
徐闻　XW	0.291 8	0.293 4	0.108 2	0.000 006 38
湛江　ZJ	0.274 4	0.294 2	0.161 5	0.000 006 30
惠州　HZ	0.294 9	0.296 2	0.098 8	0.000 006 43
东山　DS	0.293 1	0.295 5	0.104 2	0.000 006 43
厦门　XM	0.297 3	0.323 1	0.087 0	0.000 006 43
平潭　PT	0.288 1	0.294 1	0.119 4	0.000 006 39
蕉城　JC	0.329 4	0.327 8	−0.006 6	0.000 007 13
抗刺激隐核虫GS选育系F_3　GS3F3	0.326 0	0.305 4	0.003 8	0.000 006 65
抗刺激隐核虫GS平行选育系F_3　GS3RF3	0.325 6	0.314 6	0.005 3	0.000 006 84
抗内脏白点病GS选育系F_3　GS6F2	0.326 5	0.317 5	0.013 4	0.000 006 73
抗弧菌GS选育系F_3　GS8F1	0.326 9	0.315 9	0.001 3	0.000 006 87
强游泳能力GS选育系F_1　GS10F1	0.327 4	0.321 5	−0.000 4	0.000 007 00
霞浦1　XP1	0.332 1	0.324 1	−0.014 8	0.000 007 07
霞浦2　XP2	0.327 5	0.325 5	−0.000 8	0.000 007 09
霞浦3　XP3	0.333 5	0.331 1	−0.019 0	0.000 007 22
福鼎　FD	0.307 6	0.336 8	0.058 5	0.000 007 13
福鼎1　FD1	0.300 8	0.317 4	0.080 8	0.000 006 89
福鼎2　FD2	0.328 6	0.322 6	−0.004 2	0.000 007 04
大嵛山　DYS	0.327 3	0.330 7	−0.000 2	0.000 007 19
舟山　ZS	0.274 5	0.311 9	0.162 0	0.000 006 58
宁波　NB	0.288 5	0.298 8	0.118 4	0.000 006 50

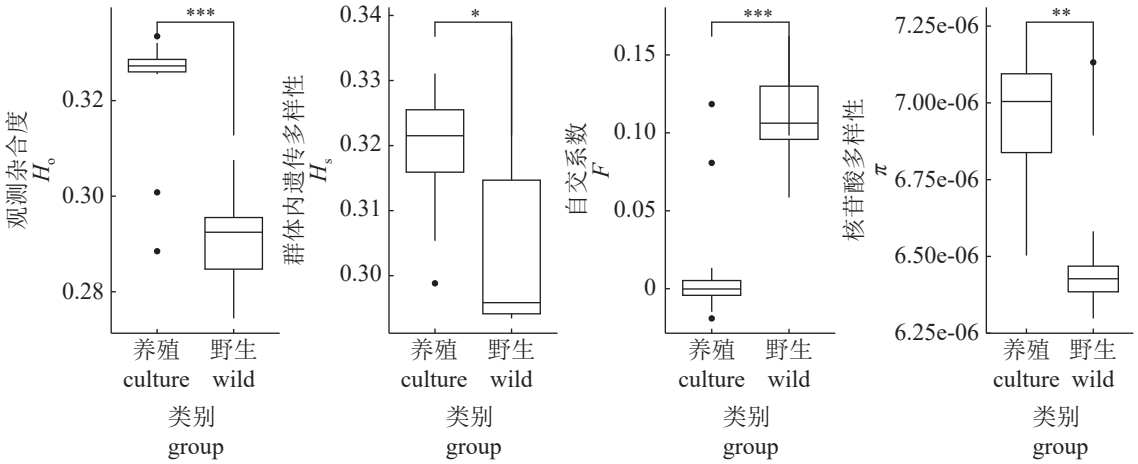

图 1　养殖群体和野生群体的群体遗传多样性

“*”表示 $P<0.05$，“**”表示 $P<0.01$，“***”表示 $P<0.001$。图中的黑色圆点代表离群点。

Fig. 1　Population genetic diversity between farmed and wild populations

"*" represents that the *P* value of *t*-test is lower than 0.05, "**"represents that the *P* value of *t*-test is lower than 0.01, "***" represents that the *P* value of *t*-test is lower than 0.001. The black dots represent outliers.

2.2　群体遗传结构

使用主成分分析、遗传混合分析和系统发育树 3 种方法揭示不同大黄鱼群体的遗传结构。主成分分析结果揭示，本研究使用的所有群体可以划分为 3 个大的地理类群：南海类群、闽东类群和岱衢类群，其中南海类群包含 XP、ZJ、HZ、DS、XM、PT，闽东类群包含 JC、GS3F3、GS3RF3、GS6F2、GS8F1、GS10F1、XP1、XP2、XP3、FD、FD1、FD2、DYS，岱衢类群包含 ZS 和 NB 群体(图 2)。南海类群与其他两个类群具有明显的隔离，岱衢和闽东两个类群的 95% 置信椭圆存在小部分重叠。遗传混合分析同样证实了以上结果，当样本所包含的祖先数设置为 2 时 (k=2)，南海类群与其他类群相比具有显著不同的祖先成分构成，

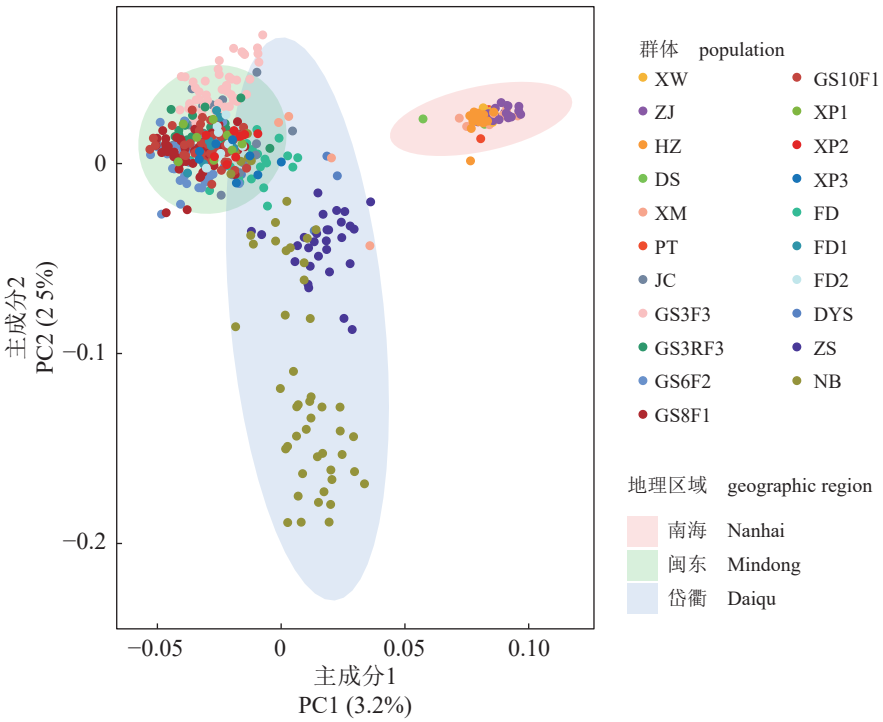

图 2　主成分分析结果

Fig. 2　Results of PCA analysis

k=3 时，岱衢类群也显示出与其他类群不同的祖先成分构成，当 *k*=4 时，闽东类群下的 GS3F3 也凸显出差别 (图 3)。系统发育树分析结果也与上述结果相一致，即所有群体可以被划分为 3 个大的地理类群 (图 4)。虽然平潭以南的 6 个群体可被划分为南海地理类群，但是并没有独立的聚类拓扑结构。闽东地理类群介于蕉城和大嵛山之间，除了 GS3F3 以及 GS3RF3 群体外，其他群体均没有明显独立的拓扑结构，而是分为几部分夹杂在不同群体之间。岱衢地理群体内部的 ZS 和 NB 群体具有分离的拓扑结构。以上结果表明，可以利用“宁芯 3 号”进行大黄鱼群体遗传结构解析。

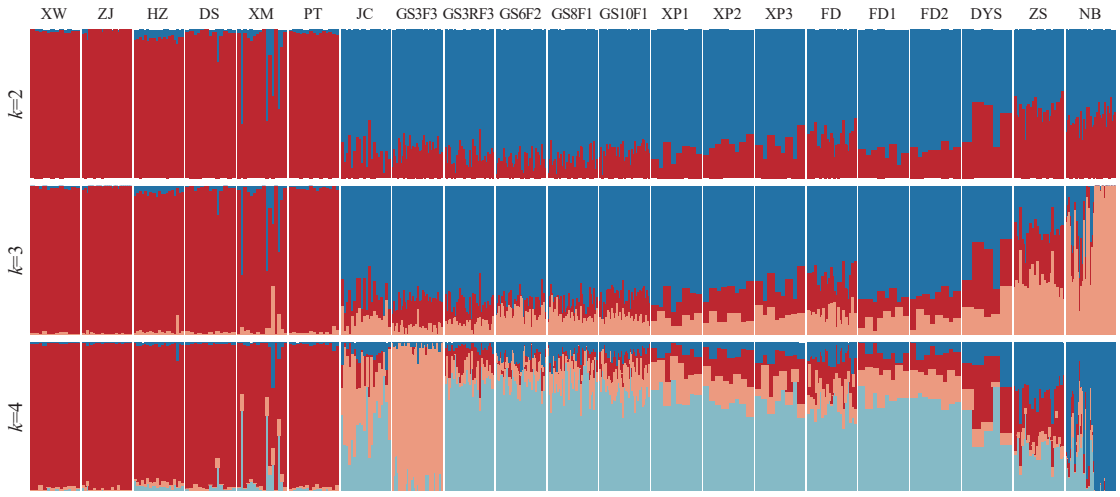

图 3　遗传混合分析结果

图中不同颜色代表不同的祖先成分。

Fig. 3　Results of admixture analysis

The different colors represent different ancestral components.

2.3　种质鉴定

基于合并基因型数据集，使用 NN、KNN、LR 和 RF 这 4 种机器学习方法对 21 个群体进行种质预测，各模型的预测准确率 (accuracy) 分别为 0.939±0.007、 0.925±0.005、 0.950±0.007、 0.914±0.006。但是单个群体在种质分类上的精确率、召回率和 F1-score 存在较大差异 (表 3)。训练集数量对各模型在群体种质鉴定上的准确性具有显著影响，当训练群体数量较多时 (GS3F3、GS6F2、GS8F1、GS10F1、NB)，鉴定准确性较高 (NN 算法的精确率大于 0.99)，但是对于经过几代抗刺激隐核虫遗传改良的 GS3RF3 选育系，即使训练集数量较少 (*n*=30)，各机器学习模型在种质鉴定上仍具有较高的准确性 (NN 算法的精确率为 0.932)。以上结果表明，在样本量充足的情况下，使用机器学习方法可以实现大黄鱼不同种质资源的精准鉴定。

对于数量较少的野生和养殖群体，可以实现地理类群的精准分类。将融合基因型文件的群体标签按照进化树分析结果修改为地理类群标签，重新使用同样的机器学习方法进行地理类群鉴定分类工作。结果揭示 4 种机器学习方法均能对大黄鱼不同地理类群进行准确判别，各项评价指标均达到 0.99 以上，仅 RF 在南海类群上表现出较差的分辨能力 (表 4)。以上结果表明，基于机器学习和“宁芯 3 号”可以实现大黄鱼不同地理类群的精准鉴定，同时也证实大黄鱼群体可以划分为南海、闽东和岱衢三个地理类群。

3　讨论

种质资源鉴定是进行种质资源保护、研究和利用的基础。某个特定物种的种质资源鉴定大致可分为两部分，大尺度的地理类群鉴定，小尺度的品系鉴定。对于地理类群鉴定，除了根据外部形态特征可以明显区分之外，目前主要方式是通过基于高通量基因型数据的群体遗传学手段鉴定，在此基础上，也可以开发一些分子标记进行辅助鉴定。一项关于大西洋鲑 (*Salmo salar*) 的群体遗传学研究揭示，基于 12 个微卫星 DNA 标记，29 个采集的野生群体可划分为北美洲和欧洲地理类群[8]。Magris 等[26] 使用 ddRAD 测序技术对分布于意大利的褐鳟 (*Salmo trutta*) 开展了群体遗传学研究，结果揭示褐鳟可划分为 5 个地理类群 (马尔莫

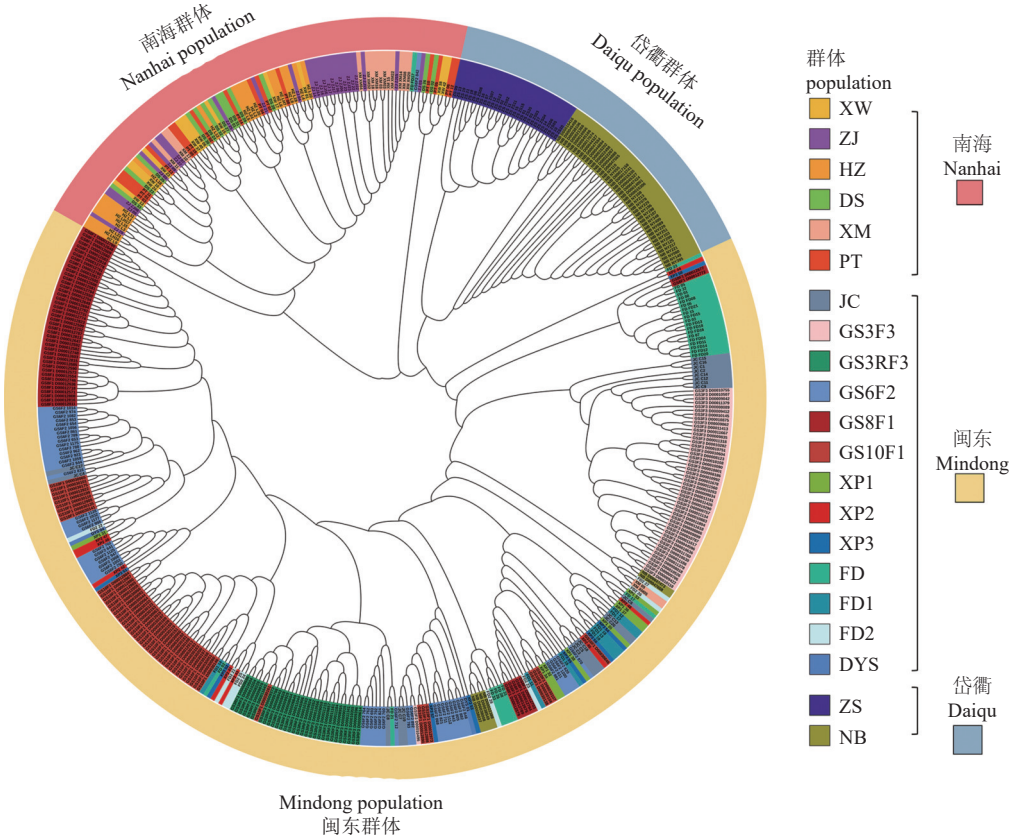

图 4　进化树分析结果

Fig. 4　Results of tree analysis

拉图斯、卡皮奥内、大西洋、地中海内陆、地中海岛屿)。Chen 等[27] 基于重测序技术揭示中国沿海花鲈 (*Lateolabrax japonicus*) 野生群体可划分为渤海湾、北部湾以及介于二者之间的 3 个地理类群，并挖掘到两个 inDel 分子标记可快速鉴定南北方花鲈群体[28]。大黄鱼是我国重要的海水养殖经济鱼类，除了具有丰富的野生遗传资源外，目前还存在众多的养殖品种和品系，因此精准鉴定大黄鱼种质资源对未来大黄鱼养殖产业的发展至关重要。长期以来，关于大黄鱼地理类群的划分备受关注。20 世纪 60 年代有关学者根据形态学将大黄鱼群体分为岱衢、闽粤东、硇洲三个地理族群[29]，也得到有关分子遗传学证据的支撑[30-31]。但是后续的许多学者认为岱衢族和闽粤东族不存在明显的遗传分化[32-33]。因此研究焦点是闽粤东族与岱衢族是否是两个独立的地理类群。本研究基于“宁芯 3 号”芯片进行的群体遗传学分析揭示大黄鱼存在 3 个遗传分化明显的族群，但是岱衢类群与闽东类群之间并不存在相互独立的遗传背景，推测大黄鱼不同地理群体的演化历史时间不同，闽东类群与岱衢类群的遗传分化时间较晚。本研究使用的野生群体的大部分个体来源于 Chen 等[34] 基于重测序的大黄鱼群体遗传学研究，在增加大量养殖群体后，得到的结果与 Chen 等[27] 的研究结果相同，这表明将大黄鱼分为 3 个地理类群是合适的。同时，本研究利用机器学习方法可以实现对大黄鱼未知个体所属地理类群的 (南海、闽东、岱衢) 精准鉴定。

对于小尺度的品系鉴定，其困难程度远高于大尺度的地理类群鉴定。品系是在某种地理遗传资源基础之上经过人工选育形成的具有某种特性的同种生物个体集合，其与经过长期自然选择形成的地理种质资源相比，遗传分化程度较低，因此品系鉴定通常具有较高的技术瓶颈[35]。基于群体遗传学的聚类分析手段 (主成分分析、祖先成分分析、进化树分析等) 已很难进行品系之间的鉴定工作，需要额外的统计检验分析实现小尺度的品系鉴定。Troutlodge 公司是世界上最大的商业虹鳟 (*Oncorhynchus mykiss*) 受精卵生产商之一，其拥有 8 个育种群体，Liu 等[36] 使用包含 95 个 SNP 的芯片对 1 732 个育种个体进行基因分型，基于似然法，97.1% 的育种个体被正确归类到 8 个群

表 3　不同机器学习方法对不同群体的分类准确率

Tab. 3　Classification accuracy of different machine learning methods for different populations

群体 population	精准率 precision				召回率 recall				F1得分 F1 score			
	神经网 NN	K近邻 KNN	逻辑斯蒂回归 LR	随机森林 RF	神经网络 NN	K近邻 KNN	逻辑斯蒂回归 LR	随机森林 RF	神经网络 NN	K近邻 KNN	逻辑斯蒂回归 LR	随机森林 RF
XW	0.050±0.224	0.101±0.124	0.000±0.000	0.000±0.000	0.013±0.056	0.375±0.425	0.000±0.000	0.000±0.000	0.020±0.089	0.157±0.188	0.000±0.000	0.000±0.000
ZJ	0.630±0.254	0.200±0.410	0.578±0.105	1.000±0.000	0.617±0.274	0.070±0.149	0.910±0.121	0.510±0.165	0.538±0.160	0.102±0.215	0.704±0.097	0.660±0.145
HZ	0.362±0.376	0.412±0.111	0.446±0.098	0.000±0.000	0.244±0.256	0.550±0.224	0.900±0.126	0.000±0.000	0.265±0.242	0.445±0.121	0.593±0.105	0.000±0.000
DS	0.108±0.255	0.09±0.134	0.000±0.000	0.000±0.000	0.071±0.151	0.225±0.343	0.000±0.000	0.000±0.000	0.083±0.183	0.128±0.191	0.000±0.000	0.000±0.000
XM	0.487±0.25	0.700±0.470	0.95±0.224	0.800±0.410	0.880±0.269	0.282±0.248	0.567±0.269	0.349±0.254	0.598±0.222	0.390±0.290	0.685±0.243	0.470±0.281
PT	0.216±0.347	0.225±0.413	0.275±0.444	0.000±0.000	0.173±0.311	0.082±0.147	0.116±0.195	0.000±0.000	0.152±0.216	0.120±0.214	0.160±0.260	0.000±0.000
JC	0.200±0.174	0.200±0.410	0.792±0.370	0.000±0.000	0.550±0.456	0.050±0.103	0.300±0.174	0.000±0.000	0.282±0.231	0.080±0.164	0.421±0.217	0.000±0.000
GS3F3	0.990±0.008	0.974±0.009	0.984±0.010	0.978±0.011	0.976±0.012	0.993±0.008	0.992±0.009	0.992±0.009	0.986±0.008	0.984±0.006	0.989±0.008	0.986±0.007
GS3RF3	0.932±0.100	0.914±0.083	0.931±0.080	0.65±0.489	0.951±0.080	0.958±0.076	0.907±0.114	0.126±0.107	0.936±0.062	0.930±0.054	0.913±0.072	0.210±0.170
GS6F2	0.994±0.007	0.879±0.015	0.935±0.017	0.864±0.027	0.946±0.024	0.995±0.006	0.998±0.004	0.994±0.007	0.970±0.010	0.932±0.010	0.968±0.008	0.924±0.016
GS8F1	0.990±0.015	0.945±0.025	0.958±0.021	0.922±0.027	0.946±0.033	0.993±0.010	0.994±0.009	0.990±0.011	0.966±0.019	0.968±0.012	0.974±0.010	0.956±0.014
GS10F1	0.991±0.012	0.982±0.017	0.922±0.029	0.936±0.035	0.926±0.045	0.987±0.016	0.990±0.012	0.970±0.022	0.957±0.027	0.984±0.013	0.955±0.018	0.953±0.021
XP1	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000
XP2	0.100±0.308	0.000±0.000	0.250±0.444	0.000±0.000	0.100±0.308	0.000±0.000	0.150±0.286	0.000±0.000	0.100±0.308	0.000±0.000	0.184±0.334	0.000±0.000
XP3	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000
FD	0.465±0.195	0.400±0.503	0.964±0.092	0.900±0.308	0.905±0.178	0.068±0.085	0.499±0.181	0.243±0.166	0.593±0.172	0.116±0.146	0.637±0.144	0.366±0.202
FD1	0.100±0.262	0.050±0.224	0.200±0.410	0.000±0.000	0.125±0.319	0.025±0.112	0.125±0.275	0.000±0.000	0.100±0.245	0.034±0.150	0.151±0.316	0.000±0.000
FD2	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000
DYS	0.050±0.224	0.000±0.000	0.050±0.224	0.000±0.000	0.050±0.224	0.000±0.000	0.050±0.224	0.000±0.000	0.050±0.224	0.000±0.000	0.050±0.224	0.000±0.000
ZS	0.835±0.153	0.900±0.308	1.000±0.000	0.950±0.224	0.925±0.187	0.309±0.189	0.874±0.130	0.400±0.184	0.866±0.161	0.442±0.230	0.928±0.078	0.548±0.201
NB	0.998±0.004	0.972±0.009	0.984±0.007	0.889±0.016	0.980±0.015	0.998±0.004	0.999±0.003	0.998±0.004	0.989±0.008	0.986±0.006	0.99±0.003	0.942±0.009
宏平均 macro average	0.452±0.032	0.426±0.060	0.534±0.042	0.425±0.040	0.495±0.035	0.380±0.032	0.494±0.034	0.312±0.018	0.450±0.029	0.371±0.029	0.491±0.039	0.334±0.023
加权平均 weighted average	0.976±0.009	0.897±0.013	0.929±0.009	0.870±0.011	0.939±0.007	0.925±0.005	0.950±0.007	0.913±0.006	0.954±0.007	0.905±0.008	0.934±0.007	0.882±0.008

表 4 不同机器学习方法对不同地理类群的分类准确率

Tab. 4 Classification accuracy of different machine learning methods for geographic groups

群体 population	精准率 precision				召回率 recall				F1得分 F1 score			
	神经网络 NN	K近邻 KNN	逻辑斯蒂回归 LR	随机森林 RF	神经网络 NN	K近邻 KNN	逻辑斯蒂回归 LR	随机森林 RF	神经网络 NN	K近邻 KNN	逻辑斯蒂回归 LR	随机森林 RF
南海 Nanhai	0.980±0.025	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	0.955±0.039	0.975±0.038	0.547±0.129	0.989±0.014	0.976±0.020	0.986±0.020	0.700±0.109
闽东 Mindong	1.000±0.002	0.996±0.005	0.998±0.004	0.965±0.012	0.997±0.005	0.998±0.004	1.000±0.000	1.000±0.000	1.000±0.002	0.998±0.004	1.000±0.002	0.981±0.006
岱衢 Daiqu	0.996±0.005	0.994±0.006	0.996±0.005	0.997±0.005	0.996±0.005	0.994±0.005	0.994±0.005	0.977±0.013	0.998±0.004	0.995±0.005	0.996±0.005	0.988±0.007
宏平均 macro average	0.992±0.010	0.998±0.004	1.000±0.000	0.989±0.003	0.999±0.003	0.983±0.015	0.991±0.013	0.841±0.044	0.996±0.005	0.990±0.008	0.995±0.008	0.889±0.038
加权平均 weighted average	0.998±0.004	0.996±0.005	0.998±0.004	0.976±0.007	0.998±0.004	0.996±0.005	0.998±0.004	0.976±0.007	0.998±0.004	0.996±0.005	0.998±0.004	0.972±0.008

体中。另一项关于尼罗罗非鱼 (*Oreochromis niloticus*) 的研究揭示，基于 ddRAD 文库 (double-digest restriction site-associated DNA libraries) 挖掘的 2 180 个 SNP 标记，使用主成分判别分析能以 77% 的准确率鉴定分布于坦桑尼亚的 7 个群体[37]。Hamilton 等[38] 搜集了来源于孟加拉国和菲律宾的 205 个孵化场的 2 057 尾尼罗罗非鱼个体，基于一个低密度 SNP 标记集合，使用主成分判别分析可以分别以 74.1% 和 80.6% 的准确率鉴定来自孟加拉国和菲律宾的不同孵化场的养殖群体。本研究同样证实，使用大量 SNP 标记，可以进行大黄鱼小尺度的品系鉴定工作。基于“宁芯 3 号”芯片和机器学习方法，一些经过数代遗传选育的新品系可以被精准鉴定，例如本课题组基于全基因组选择培育的 5 个新品系 (GS3F3、GS3RF3、GS6F2、GS8F1、GS10F1) 的鉴定准确率大于 93%。但是一些来自不同采样地理位置的野生群体或者来自不同养殖场的养殖群体具有低的鉴定准确率，这可能是由以下原因造成：首先，这些群体具有较低的样本量，机器学习无法对其进行足够的训练；其次，不同大的地理区域的野生个体具有相同的生境，不同的采样点不能代表不同的群体；最后，来源于不同养殖场的群体由于没有清晰的系谱信息，亲鱼或者苗种往往来自于采购，因此不同养殖场不能代表不同群体。总之，在样本数量充足的情况下，基于 SNP 和成熟的统计检验方法能实现某个物种的种质资源精准鉴定，即使种质之间遗传分化程度较低。这将为实现种质资源保护、高效种质管理、育种材料和新品种知识产权保护等提供有力保障。

本研究汇集融合了大黄鱼不同地理野生群体的重测序数据和养殖群体的“宁芯 3 号”液相芯片测序数据，在此基础上进行群体遗传学分析，探讨了不同机器学习方法对大黄鱼不同群体的分类能力。研究表明，大黄鱼“宁芯 3 号”芯片是进行大黄鱼群体遗传学、种质资源鉴定的优秀工具，其将在未来大黄鱼研究中发挥日益重要的作用。本研究的创新点在于首次将机器学习方法应用于大黄鱼种质鉴定领域，机器学习对不同种质的鉴定准确性依赖于训练集的规模，因此，未来需搜集更多的大黄鱼不同种质资源样本补充目前已建立的种质资源数据库，建立大黄鱼种质鉴定标准，并开发配套的可视化计算机程序执行鉴定工作。

（作者声明本文无实际或潜在的利益冲突）

参考文献 (References):

[1] 凌建忠, 李圣法, 严利平. 东海区主要渔业资源利用状况的分析 [J]. 海洋渔业, 2006, 28(2): 111-116.
Ling J Z, Li S F, Yan L P. Analysis on the utilization of main fishery resources in the East China Sea[J]. Marine Fisheries, 2006, 28(2): 111-116 (in Chinese).

[2] 徐鹏, 柯巧珍, 苏永全, 等. 大黄鱼种质资源保护与利用现状及建议 [J]. 水产学报, 2022, 46(4): 674-682.
Xu P, Ke Q Z, Su Y Q, *et al*. Protection and utilization status and prospect of large yellow croaker (*Larimichthys crocea*) germplasm resources[J]. Journal of Fisheries of China, 2022, 46(4): 674-682 (in Chinese).

[3] 王俊杰. 探析种质及种质资源定义 [J]. 甘肃林业科技, 2007, 32(2): 1-4.
Wang J J. Analysis on definition of germ plasm and germ plasm resource[J]. Journal of Gansu Forestry Science and Technology, 2007, 32(2): 1-4 (in Chinese).

[4] 贾继增. 分子标记种质资源鉴定和分子标记育种 [J]. 中国农业科学, 1996, 29(4): 1-10.
Jia J Z. Molecular germplasm diagnostics and molecular marker assisted breeding[J]. Scientia Agricultura Sinica, 1996, 29(4): 1-10 (in Chinese).

[5] Liu Z J, Cordes J F. Erratum to “DNA marker technologies and their applications in aquaculture genetics” [Aquaculture 238 (2004) 1–37][J]. Aquaculture, 2004, 242(1-4): 735-736.

[6] David L, Rajasekaran P, Fang J, *et al*. Polymorphism in ornamental and common carp strains (*Cyprinus carpio* L.) as revealed by AFLP analysis and a new set of microsatellite markers[J]. Molecular Genetics and Genomics, 2001, 266(3): 353-362.

[7] Waldbieser G C, Wolters W R. Short Communication: definition of the USDA103 strain of channel catfish (*Ictalurus punctatus*)[J]. Animal Genetics, 2007, 38(2): 180-183.

[8] King T L, Kalinowski S T, Schill W B, *et al*. Population structure of Atlantic salmon (*Salmo salar* L.): a range-wide perspective from microsatellite DNA variation[J]. Molecular Ecology, 2001, 10(4): 807-821.

[9] Chen H D, Xie W B, He H, *et al*. A high-density SNP genotyping array for rice biology and molecular breeding[J]. Molecular Plant, 2014, 7(3): 541-553.

[10] Excoffier L, Dupanloup I, Huerta-Sanchez E, *et al*. Robust demographic inference from genomic and SNP data[J]. PLoS Genetics, 2013, 9(10): e1003905.

[11] Coll F, McNerney R, Guerra-Assunção J A, *et al*. A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains[J]. Nature Communications, 2014, 5: 4812.

[12] Montesinos-López O A, Martín-Vallejo J, Crossa J, *et al*. A benchmarking between deep learning, support vector machine and bayesian threshold best linear unbiased prediction for predicting ordinal traits in plant breeding[J]. G3 Genes| Genomes| Genetics, 2019, 9(2): 601-618.

[13] Alves A A C, Da Costa R M, Bresolin T, *et al*. Genome-wide prediction for complex traits under the presence of dominance effects in simulated populations using GBLUP and machine learning methods[J]. Journal of Animal Science, 2020, 98(6): skaa179.

[14] Schiavo G, Bertolini F, Galimberti G, *et al*. A machine learning approach for the identification of population-informative markers from high-throughput genotyping data: application to several pig breeds[J]. Animal, 2020, 14(2): 223-232.

[15] Miao J, Chen Z T, Zhang Z Y, *et al*. A web tool for the global identification of pig breeds[J]. Genetics Selection Evolution, 2023, 55(1): 18.

[16] Zhou T, Chen B H, Ke Q Z, *et al*. Development and evaluation of a high-throughput single-nucleotide polymorphism array for large yellow croaker (*Larimichthys crocea*)[J]. Frontiers in Genetics, 2020, 11: 571751.

[17] 周涛, 陈葆华, 柯巧珍, 等. “宁芯 2 号”大黄鱼基因组育种芯片的开发及验证 [J]. 中国水产科学, 2022, 29(1): 41-48.
Zhou T, Chen B H, Ke Q Z, *et al*. Development and evaluation of a breeding array for genomic selection of large yellow croaker (*Larmichthys crocea*)[J]. Journal of Fishery Sciences of China, 2022, 29(1): 41-48 (in Chinese).

[18] Wang J Y, Miao L W, Chen B H, *et al*. Corrigendum to “development and evaluation of liquid SNP array for large yellow croaker (*Larimichthys crocea*)” [Aquaculture, volume 563, 2023/739021][J]. Aquaculture, 2023, 565: 739152.

[19] Ding J, Gao Z Y, Wang J Y, *et al*. Genome-wide association and transcriptome analysis provide the SNPs and molecular insights into the hypoxia tolerance in large yellow croaker (*Larimichthys crocea*)[J]. Aquaculture, 2023, 573: 739547.

[20] Tong B C, Wang J Y, Miao L W, *et al*. Development of an informative SNP panel for molecular parentage analysis in large yellow croaker (*Larimichthys crocea*)[J]. Aquaculture, 2023, 575: 739728.

[21] Wang J Y, Bai Y L, Zou X Q, *et al*. First genomic prediction of single-step models in large yellow croaker[J]. Marine Biotechnology, 2023, 25(4): 603-611.

[22] Zeng J J, Long F, Wang J Y, *et al*. GWAS reveals heritable individual variations in the inherent swimming performance of juvenile large yellow croaker[J]. Aquaculture, 2022, 559: 738419.

[23] Purcell S, Neale B, Todd-Brown K, *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses[J]. The American Journal of Human Genetics, 2007, 81(3): 559-575.

[24] Weir B S, Goudet J. A unified characterization of popu-

lation structure and relatedness[J]. Genetics, 2017, 206(4): 2085-2103.

[25] Alexander D H, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals[J]. Genome Research, 2009, 19(9): 1655-1664.

[26] Magris G, Marroni F, D'Agaro E, *et al*. ddRAD-seq reveals the genetic structure and detects signals of selection in Italian brown trout[J]. Genetics Selection Evolution, 2022, 54(1): 8.

[27] Chen B H, Zhou Z X, Shi Y, *et al*. Genome-wide evolutionary signatures of climate adaptation in spotted sea bass inhabiting different latitudinal regions[J]. Evolutionary Applications, 2023, 16(5): 1029-1043.

[28] 徐鹏, 吕红皂, 陈葆华, 等. 一种快速鉴定南北方花鲈群体的 InDel 分子标记方法: 201910892376.0[P]. 2019-09-20.
Xu P, Lü H Z, Chen B H, *et al*. InDel molecular marker method for rapidly identifying southern and northern lateolabrax maculatus populations: 201910892376.0[P]. 2019-09-20 (in Chinese).

[29] 田明诚, 徐恭昭, 余日秀. 大黄鱼形态特征的地理变异与地理种群问题 [J]. 海洋科学集刊, 1962(2): 79-97.
Tian M C, Xu G Z, Yu R X. The geographical variation of the morphological characters and the problem of geographical populations in the large yellow croaker, *Larimichthys crocea* (richardson)[J]. Studia Marina Sinica, 1962(2): 79-97 (in Chinese).

[30] 张其永, 洪万树, 杨圣云, 等. 大黄鱼地理种群划分的探讨 [J]. 现代渔业信息, 2011, 26(2): 3-8.
Zhang Q Y, Hong W S, Yang S Y, *et al*. Discussion on the division of geographic populations for the large yellow croaker (*Larimichthys crocea*)[J]. Modern Fisheries Information, 2011, 26(2): 3-8 (in Chinese).

[31] 林能锋, 苏永全, 丁少雄, 等. 大黄鱼群体遗传多样性的微卫星 DNA 分析 [J]. 福建农业学报, 2012, 27(7): 661-666.
Lin N F, Su Y Q, Ding S X, *et al*. Genetic analysis of large yellow croaker (*Larimichthys crocea*) stocks using polymorphic microsatellite DNA[J]. Fujian Journal of Agricultural Sciences, 2012, 27(7): 661-666 (in Chinese).

[32] Wang L, Shi X F, Su Y Q, *et al*. Loss of genetic diversity in the cultured stocks of the large yellow croaker, *Larimichthys crocea*, revealed by microsatellites[J]. International Journal of Molecular Sciences, 2012, 13(5): 5584-5597.

[33] 徐兆礼, 陈佳杰. 东黄海大黄鱼洄游路线的研究 [J]. 水产学报, 2011, 35(3): 429-437.
Xu Z L, Chen J J. Analysis of migratory route of *Larimichthys crocea* in the East China Sea and the Yellow Sea[J]. Journal of Fisheries of China, 2011, 35(3): 429-437 (in Chinese).

[34] Chen B H, Bai Y L, Wang J Y, *et al*. Population structure and genome-wide evolutionary signatures reveal putative climate-driven habitat change and local adaptation in the large yellow croaker[J]. Marine Life Science & Technology, 2023, 5(2): 141-154.

[35] Araujo H A, Candy J R, Beacham T D, *et al*. Advantages and challenges of genetic stock identification in fish stocks with low genetic resolution[J]. Transactions of the American Fisheries Society, 2014, 143(2): 479-488.

[36] Liu S X, Palti Y, Martin K E, *et al*. Assessment of genetic differentiation and genetic assignment of commercial rainbow trout strains using a SNP panel[J]. Aquaculture, 2017, 468: 120-125.

[37] Kajungiro R A, Palaiokostas C, Pinto F A L, *et al*. Population structure and genetic diversity of Nile tilapia (*Oreochromis niloticus*) strains cultured in Tanzania[J]. Frontiers in Genetics, 2019, 10: 1269.

[38] Hamilton M G, Lind C E, Barman B K, *et al*. Distinguishing between nile tilapia strains using a low-density single-nucleotide polymorphism panel[J]. Frontiers in Genetics, 2020, 11: 594722.

Accurate identification of *Larimichthys crocea* genetic resources based on "NingXin III" chip and machine learning method

ZHAO Ji [1,2], FENG Miaosheng [1], KE Qiaozhen [1,2], WANG Jiaying [1,2], JIANG Tingsen [1], WU Xiongfei [3], PENG Shiming [4], BAI Yulin [1], SHEN Weiliang [3], ZHOU Tao [1], PU Fei [1], XU Peng [1*]

(1. *State Key Laboratory of Mariculture Breeding*, *College of Ocean and Earth Sciences*, *Xiamen University*, *Xiamen* 361101, *China*;
2. *State Key Laboratory of Marine Environmental Science*, *College of the Environment and Ecology*, *Xiamen University*, *Xiamen* 361101, *China*;
3. *Ningbo Academy Oceanology and Fishery*, *Ningbo* 315048, *China*;
4. *East China Sea Fisheries Research Institute*, *Chinese Academy Fishery Sciences*, *Shanghai* 200090, *China*)

Abstract: *Larimichthys crocea* is an important commercial fish in China, with an annual production of more than 250,000 tons in recent years. *L. crocea* is extremely rich in genetic resources, which consisted of wild populations distributed in natural sea areas and breeding lines obtained through decades of selection breeding. There is an urgent need to develop an accurate genetic identification method to distinguish different germplasm of *L. crocea* to efficiently protect, manage and utilize *L. crocea* genetic resources. However, the lack of high-throughput genotyping tools for *L. crocea* and the lack of representative samples of geographical populations have made accurate identifiing genetic resources difficult. Based on the previously developed 55K liquid SNP array ("Ningxin III") for *L. crocea*, the present study aims to carry out genetic identification for 21 *L. crocea* populations, including wild populations in coastal China, cultured populations in Fujian and Zhejiang, and multiple breeding lines. The results of population genetic analysis revealed that the *L. crocea* population could be divided into Nanhai, Mindong and Daiqu populations, among which the genetic differentiation of the Nanhai population was the most significant. The classification results of large yellow croaker populations based on machine learning methods showed that the identification accuracy rate of the geographical group to which unknown *L. crocea* individuals belong was more than 99%. The breeding lines to which unknown *L. crocea* individuals belong also had a very high identification accuracy rate. For example, after three generations of genetic selection, a new strain (GS3F3) that had strong resistance against *Cryptocaryon irritans* had an identification accuracy rate of 99% based on the neural network method. The present research showed that "Ningxin III" chip and machine learning methods could be used to implement quick and accurate genetic identification for *L. crocea*. The present study provided an effective tool for accurately identifying and managing the genetic resources of *L. crocea*, intellectual property protection for breeding materials and lines. Also, it provided a reference for the genetic identification of other aquatic organisms. In the future, it is necessary to establish a complete database covering all *L. crocea* germplasm resources and genetic identification standards, and develop a supporting visual computer program to perform identification work.

Key words: *Larimichthys crocea*; "Ningxin III" chip; germplasm identification; machine learning

Corresponding author: XU Peng. E-mail: xupeng77@xmu.edu.cn

Funding projects: National Key Research and Development Program of China (2022YFD2401002); National Science Fund for Distinguished Young Scholars (32225049); Seed Industry Innovation and Industrialization Project of Fujian Province (2021FJSCZY01)