

DOI: 10.12264/JFSC2021-0251

瓦氏雅罗鱼盐碱适应相关 InDels 位点的挖掘与分析

王双毅^{1,2}, 梁利群², 常玉梅², 孙博²

1. 上海海洋大学水产与生命学院, 上海 201306;

2. 中国水产科学研究院黑龙江水产研究所, 淡水鱼类育种国家地方联合工程实验室, 农业农村部淡水水产生物技术与遗传育种重点实验室, 黑龙江省特殊生境鱼类种质特性与抗逆育种重点实验室, 黑龙江 哈尔滨 150070

摘要: 为了探究瓦氏雅罗鱼(*Leuciscus waleckii*)极端盐碱环境下的适应机制, 采用比较基因组学的研究方法, 分别对来自内蒙古高盐碱湖泊达里湖(53.57 mmol/L, pH 9.6)和来自松花江流域的瓦氏雅罗鱼各 5 尾样本进行了基因组重测序, 重点对两个种群全基因组水平的插入/缺失(insertion-deletion, InDels)位点进行了比较分析。测序共得到 486.57 G 高质量数据, 3243686532 个 clean reads, 平均测序深度为 51.34 X。序列比对后进行 InDels 调用共获得非冗余 InDels 位点 983528 个。种群比较分析获得差异 InDels 位点 8176 个, 整合差异 InDels 和两个种群基因组强选择信号区域, 共获得与盐碱适应相关的 InDels 位点 325 个, 关联候选基因 176 个。富集分析显示, 176 个候选基因主要在离子转运、酸碱平衡和炎症免疫反应等生物学过程发挥重要作用。利用 EasyCodeML、MEME 和 FEL 3 种模型, 对离子转运和酸碱平衡等盐碱适应关键调控过程的候选基因 *abcc1*、*atp2b1*、*slc4a4*、*slc7a2* 和 *aqp4* 进行了进化分析和选择压力检测, 结果显示净化选择均对候选基因起主导作用, 但 *abcc1*、*atp2b1*、*slc4a4* 和 *slc7a2* 的受选择位点表现出不同程度的正向选择, 并且部分位点处于关键功能结构域中。蛋白三维结构分析显示, SLC4A4 和 SLC7A2 两个蛋白的少数位点与配体结合位点重合或非常接近, 推测盐碱胁迫环境介导的正向选择压力驱动了对这些位点的选择。本研究可为瓦氏雅罗鱼盐碱适应候选基因的深入挖掘及育种新型分子标记的开发提供参考和依据。

关键词: 瓦氏雅罗鱼; 全基因组重测序; 插入/缺失; 盐碱适应性; 选择压力; 适应性进化

中图分类号: S961

文献标志码: A

文章编号: 1005-8737-(2022)02-0184-16

瓦氏雅罗鱼(*Leuciscus waleckii*)隶属于鲤形目(Cypriniformes)、鲤科(Cyprinidae)、雅罗鱼亚科(Leuciscinae)、雅罗鱼属, 是我国北方知名的土著鱼类, 主要分布于黑龙江流域、辽河、黄河、内陆淡水及盐碱湖泊。瓦氏雅罗鱼具有极强的盐碱耐受能力, 能够耐受内蒙古达里湖碱度 53.57 mmol/L, pH 9.6 的极端不良环境, 并形成优势种^[1]。据地质学和地理学考证, 达里湖在更新世早期(11500~7600 CAL.BP)是一个淡水湖泊, 和古辽河水系相连; 到更新世中晚期(3450 CAL. BP 至今), 由于气候干燥, 降水减少, 地质下陷等原

因, 演变为封闭的内陆盐碱湖泊^[2]。而栖息于此的瓦氏雅罗鱼(碱水种), 被逐渐升高的碱度胁迫和选择, 在不足 1 万年的时间里实现了对达里湖高碱度高 pH 环境的快速适应^[3]。室内碳酸盐碱度急性毒理研究发现, 瓦氏雅罗鱼碱水种的耐碱能力是淡水种的 1.5 倍^[4]。群体遗传学研究显示, 瓦氏雅罗鱼碱水种和淡水种出现了明显的群体遗传结构分化, 推测地理隔离和局域适应(local adaptation)可能是造成二者分化的主要原因^[1,5-6]。因此, 瓦氏雅罗鱼独特的地质和演化历史为揭示硬骨鱼类在古环境中适应极端环境的分子机制研究提供

收稿日期: 2021-05-22; 修订日期: 2021-07-07.

基金项目: 国家重点研发计划项目(2019YFD0900405); 国家自然科学基金项目(31602136); 中国水产科学研究院中央级公益性科研院所基本科研业务费专项(2019ZD0601, 2020TD22).

作者简介: 王双毅(1996-), 男, 硕士研究生, 研究方向为鱼类生物信息学. E-mail: 18266810085@163.com

通信作者: 常玉梅, 研究员, 研究方向为鱼类分子生物学. E-mail: changyumei@hrfri.ac.cn

了理想的研究模型。

全基因组高通量测序技术的发展,使从基因组尺度解析瓦氏雅罗鱼盐碱适应的遗传基础成为可能。Xu 等^[6]通过对瓦氏雅罗鱼达里湖和黑龙江种群基因组 SNP 多态性进行扫描和比较分析,发现了一系列与硬骨鱼适应极端碱性环境相关的候选基因,功能分析发现这些基因参与了酸碱调节、离子转运、含氮废物排泄等多种调控通路。笔者近期通过消除地理隔离和空间背景差异,聚焦碱水和淡水两种异质环境,基于全基因组差异的 SNP 位点筛选鉴定了 21 个受盐碱正向选择的候选基因^[5]。这些研究表明达里湖种群已进化出独特的基因组适应性机制以应对极端盐碱环境。

相比于 SNP,插入缺失(insertion-deletion, InDel)作为第二种最常见的基因组变异类型,在稳定性和多态性上的表现更加突出^[7]。目前在动植物中这种分子标记已经开始应用于种群遗传结构分析以及通过全基因组关联研究(genome-wide association studies, GWAS)开展的分子辅助育种^[8-10],但在鱼类基因组中相关研究报道相对较少。

本研究选取瓦氏雅罗鱼达里湖和松花江种群为研究对象,利用全基因组重测序技术,筛选碱水种和淡水种基因组差异 InDel 突变位点,结合基因组选择清除分析和基因选择压力分析筛选受盐碱正向选择的 InDel 位点及其关联基因,并对候选 InDel 关联基因进行选择压力及蛋白质结构预测分析。旨在通过全基因组重测序挖掘 InDel 变异位点,为瓦氏雅罗鱼适应极端盐碱环境提供全面的、潜在的遗传基础证据,也为鱼类育种新型分子标记的开发提供参考和依据。

1 材料与方法

1.1 样本来源及 DNA 提取

瓦氏雅罗鱼分别采自内蒙古达里湖和松花江流域绥滨段,两个采样点的水质参数参照 Chang 等^[1]和常玉梅等^[11]的报道。每个群体各采集 5 尾样本,达里湖种群(碱水种, DL)平均体重和体长分别为(139.18±10.96) g 和(20.90±0.68) cm;松花江种群(淡水种, SH)平均体重和体长分别为(125.85±3.67) g 和(20.10±0.24) cm。鱼体经 MS-222 (Sigma-

Aldrich, 美国)麻醉后用肝素钠浸润过的 1.5 mL 注射器从尾静脉采集血液,放入液氮保存备用。血液基因组 DNA 使用 DNeasy Blood and Tissue Kit (Qiagen, 德国)试剂盒进行提取,1%的琼脂糖凝胶电泳检测 DNA 完整性,并使用 Nanodrop 8000 (Thermo Scientific, 美国)检测 DNA 浓度和纯度。检测合格的 DNA 置于-20 °C 保存备用。

1.2 DNA 文库构建及高通量测序

检验合格的 DNA 样品(>1 µg)通过 CovarisS2 随机打断成长度为 350 bp 的片段。采用 TruSeq Library Construction Kit (Illumina, 美国)进行建库,严格使用说明书推荐的试剂和耗材。DNA 片段经末端修复、加 ployA 尾、加测序接头、纯化、PCR 扩增等步骤完成整个文库制备过程。构建好的文库经质检后在 Illumina Hiseq4000 平台上进行 PE150 双末端测序。文库构建及高通量测序由北京诺禾致源科技股份有限公司合作完成。

1.3 测序数据质量评估

原始测序数据经过 FastX-Toolkit 软件(v0.0.14) (http://hannonlab.cshl.edu/fastx_toolkit/)进行过滤,过滤流程如下:(1) 过滤掉总体质量偏低(碱基质量分数 $Q < 20$)的 reads 序列;(2) 过滤去除 $Q > 20$ 但有效碱基数含量小于 50%的 reads 序列;(3) 过滤掉 3'端 $Q < 10$ 的碱基序列,确保测序碱基错误率小于 0.1;(4) 过滤掉所有 reads 序列中的接头序列;(5) 过滤掉测序深度不足而无法识别的碱基占比大于 10%的 reads 序列;(6) 过滤掉长度 < 20 bp 的 reads 序列。

1.4 InDel calling 和统计分析

使用 Burrows-Wheeler Alignment 软件(BWA, v0.7.17)^[12]以默认参数将过滤后的数据与瓦氏雅罗鱼参考基因组(NCBI: GCA_900092035.1)^[6]进行比对,使用 NGS QC 工具包^[13]对过滤 reads 进行进一步的质量控制;采用 Samtools 软件(v1.9)^[14]中的 rmdump 去除重复序列,使用基因组 GATK 工具包(v.3.5)^[15]对 InDel 周围的 reads 进行比对,其中 RealignerTargetCreator 查找需要进行精细比对的位置信息;InDelRealigner 对初始比对文件进行精细比对;最后使用 BaseRecalibrator 进行碱基质量分数的重新校准。

1.5 差异 InDel 位点挖掘

使用 Samtools 软件^[14] mpileup 鉴定 InDel 变异位点, 过滤标准如下: (1) $Q \geq 20$; (2) 每个个体测序覆盖深度(QualByDepth) >3 ; (3) 序列比对质量分数均方根(RMSMappingQuality) >20 。

使用 SnpEff 软件(v4.1)^[16]根据瓦氏雅罗鱼参考基因组和注释文件创建本地数据库, 对获得的高质量 InDel 突变位点进行位置和效应信息注释。完成注释后, 首先收集达里湖和松花江种群基因组上所有位置坐标相同的 InDel 区域位点, 其次与松花江种群个体间共享位点(同一种群个体间 InDel 位点变化一致)比较, 筛选出达里湖种群差异共享位点及其对应基因。此外, 笔者前期基于遗传分化指数(fixation index, F_{st})和种群核苷酸多样性比值 π ratio (SH/DL)进行了达里湖种群和松花江种群基因组上的选择信号检测, 将 F_{st} 和 π ratio (SH/DL)占前 5%的部分($F_{st} \geq 0.375$, π ratio ≥ 4.160)作为基因组具有强选择性扫描信号的区域^[5]。在本研究中, 笔者以强选择扫描信号区域的 F_{st} 和 π ratio (SH/DL)值为参考, 过滤得到位于该区域的达里湖种群 InDel 差异共享位点及其候选基因。

1.6 基因功能富集分析

将候选基因递交至 KOBAS 3.0^[17] (<http://kobas.cbi.pku.edu.cn/>)进行 GO 和 KEGG 富集分析^[18-19]。在 Fisher 精确检验的基础上, 采用 Benjamini-Hochberg 多重测试对 P 值进行调整, 设定 0.05 为显著性阈值。

1.7 进化选择压力分析

根据富集分析结果, 选择参与离子转运和酸碱平衡关键调控过程的候选基因 mRNA 序列进行选择压力分析。首先从雅罗鱼基因组数据库 (<http://www.fishbrowser.org/database/amur-ide/>)中调取达里湖种群候选基因 mRNA 序列, 然后检索 GenBank 数据库(www.ncbi.nlm.nih.gov/genbank)收集底鳉(*Fundulus heteroclitus*)、尼罗罗非鱼(*Oreochromis niloticus*)、大西洋鲑(*Salmo salar*)、大黄鱼(*Larimichthys crocea*)、斑马鱼(*Danio rerio*)、鲤(*Cyprinus carpio*)、鲫(*Carassius auratus*)等 12 种硬骨鱼对应基因 mRNA 序列(表 1)。将每

个物种所得序列与达里湖种群序列进行 Blast 比对以确保序列相似性。通过 MEGA X 软件 ClustalW 程序^[20]以默认比对参数进行多序列比对, 比对结果利用 TBtools 软件^[21]进行修剪后通过最大似然法(maximum likelihood, ML)构建系统发育树确定各物种间的系统发育关系, 自验值设定为 1000 以验证进化树的可信程度。使用 EasyCodeml 软件(v4.8)^[22]中的 CodeML 算法估算候选基因的非同义/同义替换率比率 dN/dS (ω), 系统发育树(.tree)和转换为 .phy 的多序列比对文件作为输入数据。基于该软件中的位点模型检验候选基因的正向选择假设, 该模型允许选择压力随位点的不同而变化, 即假设系统发育树的不同分支所受选择压力相同, 但不同氨基酸位点经历的选择压力不同。本研究考虑了 4 对替换模型, M0 (one ratio) vs M3 (discrete), M1a (Nearly Neutral) vs M2a (Positive Selection), M7 (β) vs M8 (β and $\omega > 1$)以及 M8 (β and $\omega > 1$) vs M8a (β and $\omega = 1$), 与模型 M1a、M0、M7 相对应的 M2a、M3、M8s 假定条件均允许 $\omega > 1$ 。使用似然比检验(LRT)比较了模型的拟合度, 以模型中的 Bayes Empirical Bayes (BEB)标准来确定正向选择压力位点, 若后验概率 ≥ 0.95 , 表明该位点受到显著的正向选择压力。此外, 为避免假阳性正向选择位点的出现, 本研究还考虑了同义率变化和重组的其他测试, 包括混合效应模型(MEME)和固定效应模型(FEL)。两种测试方法通过 Datamonkey 在线服务器(<http://www.datamonkey.org>)完成, 满足以下条件的位点鉴定为正向选择位点: (1) $\beta > \alpha$; (2) 似然比检验 $P < 0.1$ 。通过 EasyCodeML、MEME 和 FEL 其中两种或两种以上测试的位点被认为是强正向选择压力位点。

1.8 正向选择位点功能预测

为了确定受正向选择氨基酸位点的功能, 将瓦氏雅罗鱼候选基因氨基酸序列使用 HMMER^[23]与斑马鱼序列进行同源搜索; 使用 MEGAX 软件的 MUSCLE 程序进行比对以确定瓦氏雅罗鱼正向选择位点在斑马鱼上的等效位置; 使用 Pfam 在线分析系统^[24]确定斑马鱼同源序列上的等效位

表 1 不同物种 mRNA 序列登录号
Tab. 1 mRNA accession number of different species

物种 species	基因 gene				
	<i>abcc1</i>	<i>atp2b1</i>	<i>slc4a4</i>	<i>slc7a2</i>	<i>aqp4</i>
斑马鱼 <i>Danio rerio</i>	XM_001341859.7	XM_021474965.1	XM_009301742.3	NM_001008584.1	NM_001358313.1
鲤 <i>Cyprinus carpio</i>	XM_019063140.1	LHQP_01004580.1	LHQP_01006640.1	LHQP_01044284.1	XM_019126192.1
太平洋鲑 <i>Oncorhynchus kisutch</i>	XM_031814424.1	N.A.	XM_031816383.1	XM_020502083.2	XM_031833596.1
青鳉 <i>Oryzias latipes</i>	XM_020705549.1	XM_020714117.2	XM_023958122.1	XM_011480522.3	XM_011486891.3
尼罗罗非鱼 <i>Oreochromis niloticus</i>	XM_005461613.4	XM_019346817.2	XM_005451543.4	XM_005460484.4	XM_005476355.4
金线鲃 <i>Sinocyclocheilus rhinoceros</i>	XM_016441595.1	XM_016272439.1	XM_016481766.1	XM_016446453.1	XM_016532039.1
鲫 <i>Carassius auratus</i>	MH_420555.1	XM_026236876.1	XM_026211369.1	XM_026281018.1	XM_026195121.1
斑点叉尾鮰 <i>Ictalurus punctatus</i>	XM_017490307.1	XM_017494455.1	XM_017452158.1	XM_017474119.1	XM_017455409.1
金鲈 <i>Perca flavescens</i>	XM_028598933.1	N.A.	XM_028577137.1	XM_028589476.1	XM_028592985.1
底鳉 <i>Fundulus heteroclitus</i>	XM_012860936.3	XM_021309511.2	N.A.	XM_021318690.2	XM_012875407.3
大黄鱼 <i>Larimichthys crocea</i>	XM_010755057.3	XM_027291879.1	XM_010736564.3	XM_010732897.3	XM_010730011.3
遮目鱼 <i>Chanos chanos</i>	XM_030790406.1	XM_030778801.1	XM_030772709.1	XM_030765002.1	XM_030770711.1

点是否处于功能结构域或功能结构域位点周围。此外,使用蛋白空间结构分析服务器 I-TASSER^[25]通过折叠辨识模拟的方法预测候选基因的蛋白质三维结构,进一步了解基因编码蛋白的功能意义,模型评估通过 C-score 和 TM-score 完成。对于 C-score,可信度区间为[-5,2];对于 TM-score,大于 0.5 表明拓扑结构具有较高的准确性,而小于 0.17 表明拓扑结构准确性较差^[25-26]。最终的蛋白质三维结构模型通过 PyMOL 软件(<https://www.schrodinger.com/pymol/>)进行展示。

2 结果与分析

2.1 测序数据统计

样品经过测序和质控共产生了 3243686532 个高质量 clean reads, 包含 486.57 G 的测序数据。其中 DL 组有 1659962056 reads 和 249 G 碱基数, SH 组有 1583724476 reads 和 237.57 G 碱基数(表 2)。与初始 reads 相比,保留下来的高质量 clean

reads 占 96.7%, 其中, phred score>20 的 reads 占 97.3%, phred score>30 的 reads 数占 92.6%。GC 含量较稳定,在 39.30~39.88 之间浮动。reads 比对至参考基因组后,统计结果显示,DL 组总的 Mapping reads 数为 1364105229,平均比对率为 82.17%,平均测序深度约为 54.4 X;SH 组总的 Mapping reads 数为 1287038455,平均比对率为 81.27%,平均测序深度约为 51.34 X (表 2)。

2.2 InDel 检测与分布统计

DL 和 SH 两组所有个体共有 983528 个 InDels 被保留下来,在各个染色体上,分布的 InDels 数量为 21244~67765 (图 1)。总体上各个染色体上分布的 InDels 数量有所差异或偏倚,平均每条染色体上每 Mb 分布 2.16~6.89 个(图 2)。在保留的 983528 个 InDels 中,DL 组有 420095 个,SH 组有 563433 个。InDels 长度为 1~29 bp,插入多于缺失,主要以 1~8 bp 的小片段为主。单碱基 InDels 最多,其中 DL 组插入数和缺失数分别为 294851 个和

表 2 瓦氏雅罗鱼达里湖和松花江种群个体测序数据及比对结果统计

Tab. 2 Summary of sequencing and mapping statistics of *Leuciscus waleckii* individuals from Lake Dali Nor (DL) and Songhua River (SH) populations

样品 sample	过滤碱基数/G clean bases	过滤读长 clean reads	比对读长 mapped reads	唯一比对读长率/% uniquely mapped reads	测序深度/X sequencing depth
DL1	48.49	323299350	265765182	82.20	53.00
DL2	48.22	321441642	263095847	81.85	52.50
DL3	47.74	318252248	261486938	82.16	52.10
DL4	55.94	372930972	307344683	82.41	61.30
DL5	48.61	324037844	266412579	82.22	53.10
合计 total	249	1659962056	1364105229	N.A.	N.A.
平均 average	49.8	331992411	272821046	82.17	54.40
SH1	48.35	322305900	260945026	80.96	52.00
SH2	47.90	319304476	259156951	81.16	51.70
SH3	46.85	312324164	254561923	81.51	50.80
SH4	46.06	307058642	251545155	81.92	50.20
SH5	48.41	322731294	260829400	80.82	52.00
合计 total	237.57	1583724476	1287038455	N.A.	N.A.
平均 average	47.51	316744895	257407691	81.27	51.34

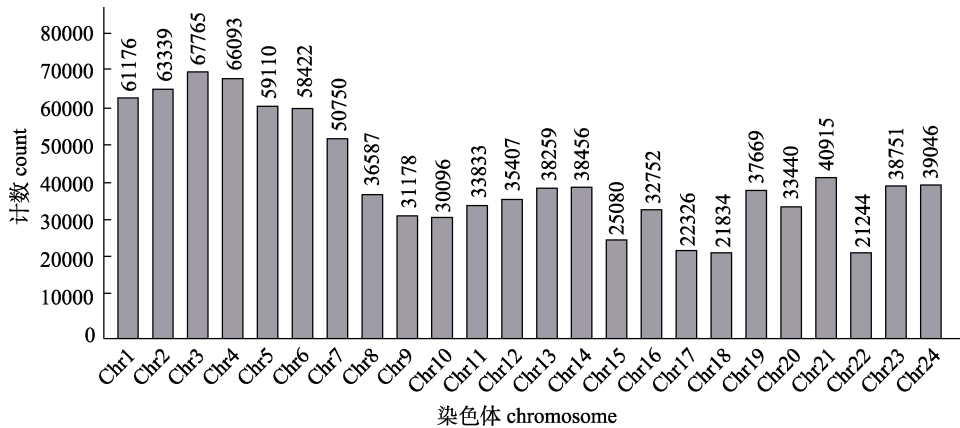


图 1 瓦氏雅罗鱼每条染色体上的插入/缺失位点(InDels)数量分布

Fig. 1 The number of insertion-deletion sites (InDels) on each chromosome of *Leuciscus waleckii*

30515 个, SH 组分别为 353277 个和 26539 个。其次为双碱基 InDels, 其中 DL 组插入数和缺失数分别为 18584 个和 13097 个, SH 组分别为 30870 个和 21302 个。随着 InDels 碱基数的增加, InDels 数量整体呈下降趋势(图 3)。

2.3 InDels 注释

DL 组总 InDels 数量为 420095 个。大多数 InDels 位于基因间区域(328682 个, 78.24%), 少部分位于基因区域(76289 个, 18.16%)、基因上游(8906 个, 2.12%)和下游(6218 个, 1.48%), 其中基因区域包括内含子(66173, 86.74%)、外显子(6340,

8.31%) 和非翻译区(3'-UTR, 2609 个, 3.42%; 5'-UTR, 1169 个, 1.53%)。编码区内移码插入总数(frameshift insertion)为 1818 个(28.68%), 编码区内移码缺失总数(frameshift deletion)为 2191 个(34.56%), 编码区内 InDels 突变获得终止子总数(stop-gain)为 186 个(2.94%), 编码区内 InDels 突变丢失终止子总数(stop-lost)为 280 个(4.41%); 非编码区内移码插入总数(non-frameshift insertion)和非编码区内移码缺失总数(non-frameshift deletion)分别为 793 个(12.50%)和 1072 个(16.91%)(图 4a)。

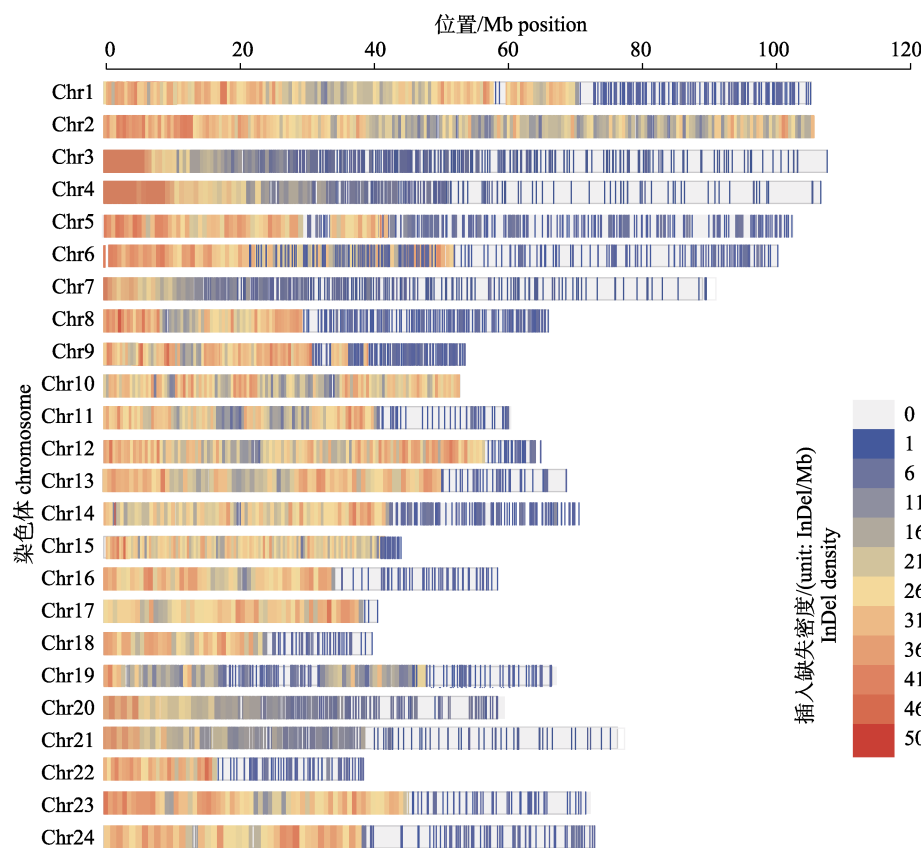


图 2 插入/缺失位点(InDels)在瓦氏雅罗鱼各染色体的分布情况

Fig. 2 Insertion-deletion sites (InDels) distribution on the chromosomes of *Leuciscus waleckii*

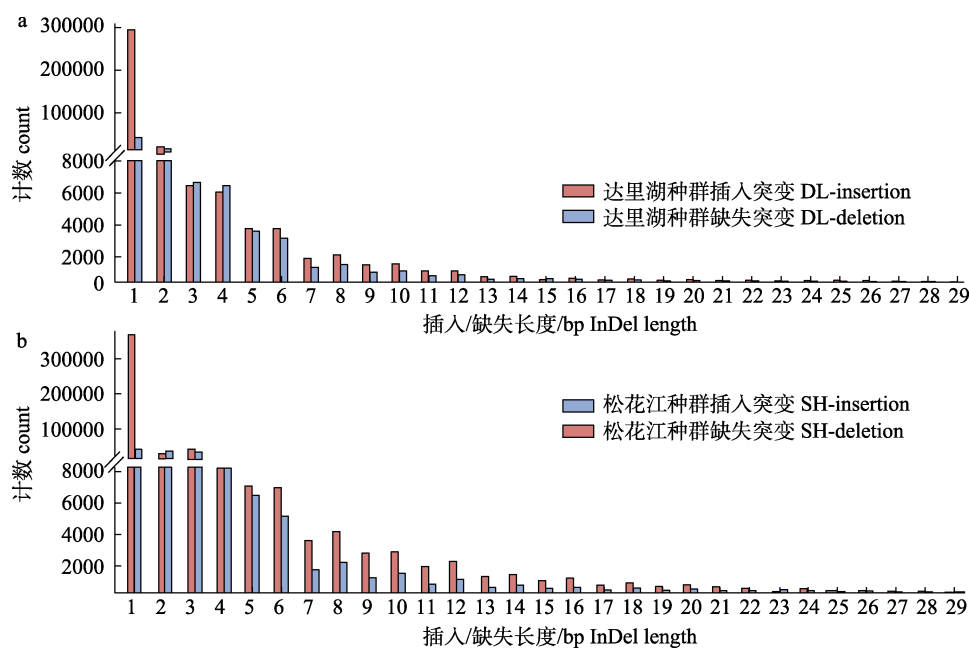


图 3 不同长度插入/缺失位点(InDels)在瓦氏雅罗鱼基因组的分布

a. 瓦氏雅罗鱼达里湖种群不同长度 InDels 基因组分布; b. 瓦氏雅罗鱼松花江种群不同长度 InDels 基因组分布.

Fig. 3 The distribution of different size insertion-deletion sites (InDels) in *Leuciscus waleckii* genome

a. The distribution of InDels length in *Leuciscus waleckii* from Lake Dali Nor (DL) population;

b. The distribution of InDels length in *Leuciscus waleckii* from Songhua River (SH) population.

SH 组总 InDel 数量为 563433 个。与 DL 组分布相似, 大多数 InDels 位于基因间区域(430068 个, 76.33%), 少部分位于基因区域(120743 个, 21.43%)、基因上游(7043 个, 1.25%)和下游(5579 个, 0.99%), 其中基因区域包括内含子(93709 个, 77.61%)、外显子(20164 个, 16.70%)和非翻译区(3'-UTR, 3513 个, 2.91%; 5'-UTR, 3357 个, 2.78%)。

编码区内移码插入总数为 5767 个(28.60%), 编码区内移码缺失总数为 6769 个(33.57%), 编码区内 InDels 突变获得终止子总数为 645 个(3.20%), 编码区内 InDel 突变丢失终止子总数为 250 个(1.24%); 非编码区内移码插入总数和非编码区内移码缺失总数分别为 3438 个(17.05%)和 3295 个(16.34%)(图 4b)。

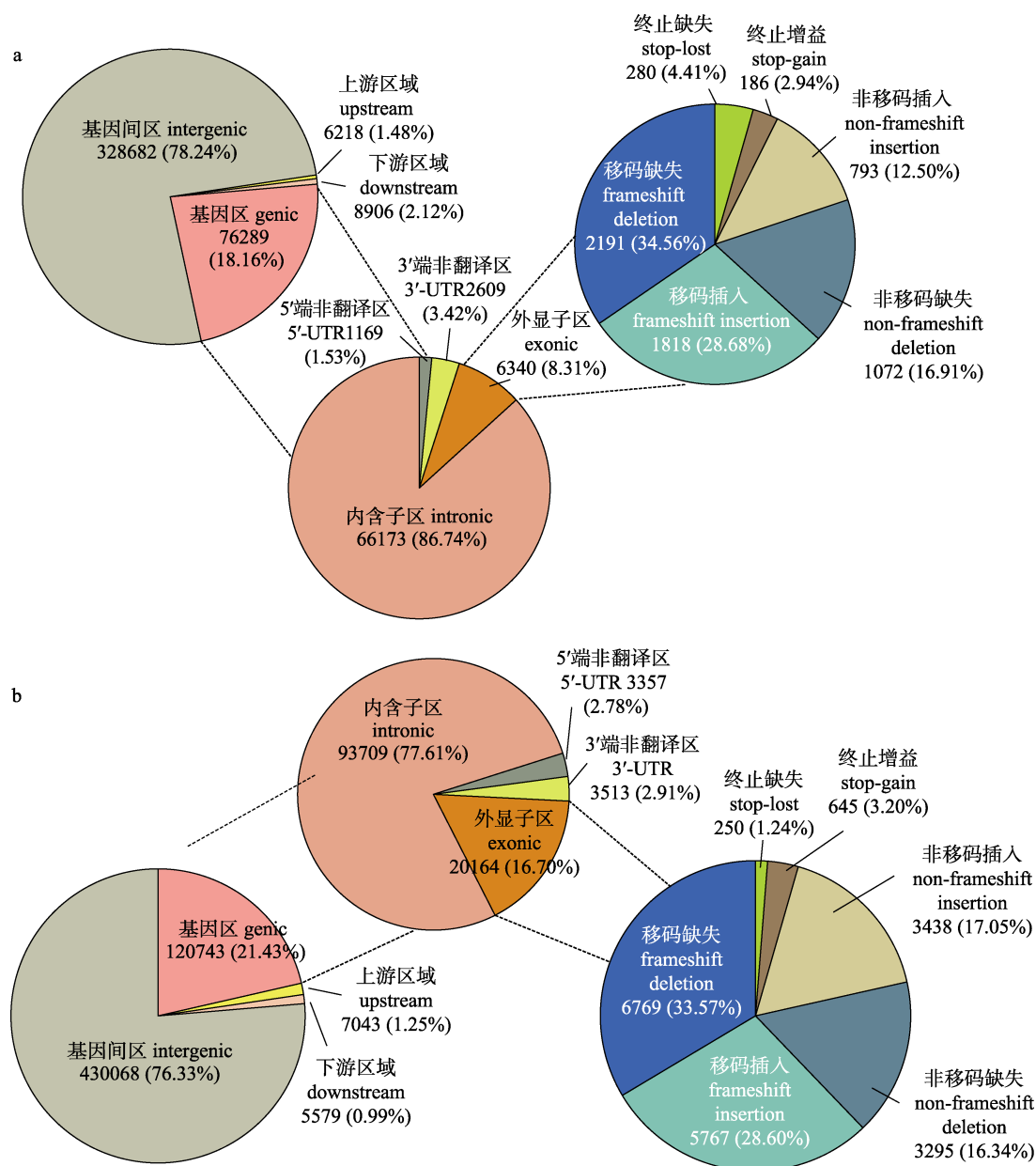


图 4 瓦氏雅罗鱼不同种群插入/缺失位点(InDels)区域功能分类

a. 瓦氏雅罗鱼达里湖种群 InDels 区域功能分类; b. 瓦氏雅罗鱼松花江种群 InDels 区域功能分类。

Fig. 4 Functional classification of the detected InDels for *Leuciscus waleckii* population groups

a. Functional classification of the detected InDels for *Leuciscus waleckii* from Lake Dali Nor (DL);

b. Functional classification of the detected InDels for *Leuciscus waleckii* from Songhua River (SH).

2.4 差异 InDels 位点筛选

DL 组与 SH 组共筛选得到 8176 个差异共享 InDels 位点, 其中基因间区域为 5755 个(70.39%), 基因上游和下游分别为 60 个(0.74%)和 55 个(0.68%), 基因区域中非翻译区 3'-UTR 为 23 个(0.29%), 5'-UTR 为 22 个(0.27%), 内含子为 2074 个(25.37%), 外显子为 187 个(2.26%)。通过计算 F_{st} 和 π ratio (SH/DL) 在两个种群中共筛选到 3259 个强选择信号区域^[5], 将 DL 差异共享位点与强选择信号区域取交集, 共得到 325 个差异共享位点关联到 176 个基因, 将这些基因确定为最终的功能候选基因。

2.5 基因功能富集分析

176 个候选基因共富集到 1613 个 GO 条目和 181 个 KEGG 通路中, 其中显著富集的 GO 条目有 193 个, KEGG 通路有 27 个($P < 0.05$)。GO 分析显示生物过程(biological process, BP)相关的候选基因富集程度较高的为免疫系统, 包括凋亡正调控(GO: 0043065)、免疫应答(GO: 0045087)、白细

胞介素-2 合成正调控(GO: 0032743)等过程; 其次为离子跨膜转运(GO: 0034220)和碱性氨基酸转运体活性(GO: 0015174)相关过程的候选基因也存在显著富集。细胞组分(cellular component, CC)中与能量代谢有关的候选基因存在显著富集, 包括线粒体(GO: 0005739)、线粒体内膜(GO: 000743)、呼吸链复合物 IV (GO: 0045277)等过程。分子功能(molecular function, MF)中转运调节基因富集程度较高, 包括 Na^+ 跨膜转运活性(GO: 0015081)和 ABC 型转运蛋白活性(GO: 0140359)等(图 5)。KEGG 分析显示参与细胞凋亡与免疫应答相关的功能区域显著富集, 包括 NOD 样受体通路(04621)、Toll 样受体通路(04620)、自噬(04215)、TNF 信号通路(04668)和 p53 信号通路(04115)等; 此外, 可能与逆境生境适应相关的功能区域也有显著富集, 包括 Ca^{2+} 信号通路(04020)、心肌收缩通路(04260)、肾素-血管紧张素系统(04614)以及应激激素区域, 如催乳素通路(04917)等(图 6)。

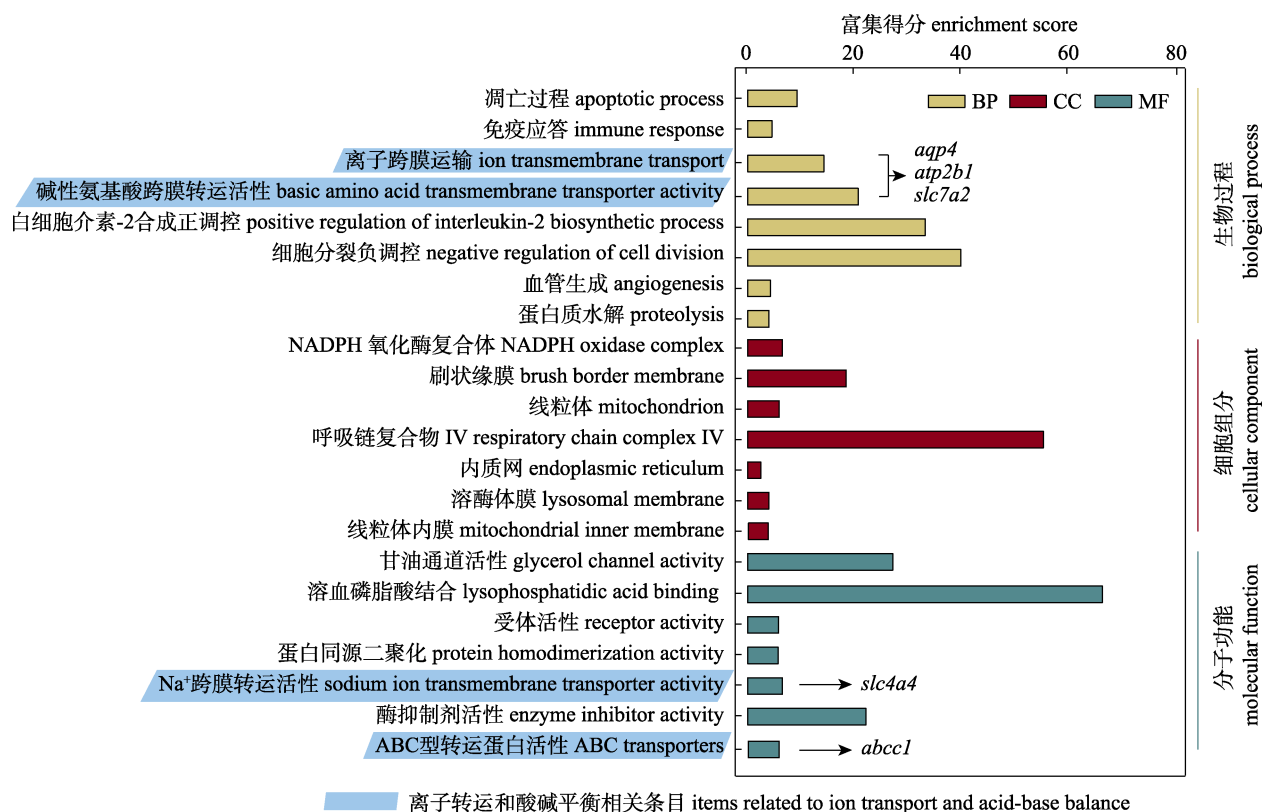


图 5 达里湖瓦氏雅罗鱼种群盐碱适应候选基因 GO 分析

Fig. 5 GO annotation of candidate genes for alkali-saline adaptation in *Leuciscus waleckii* from Lake Dali Nor population

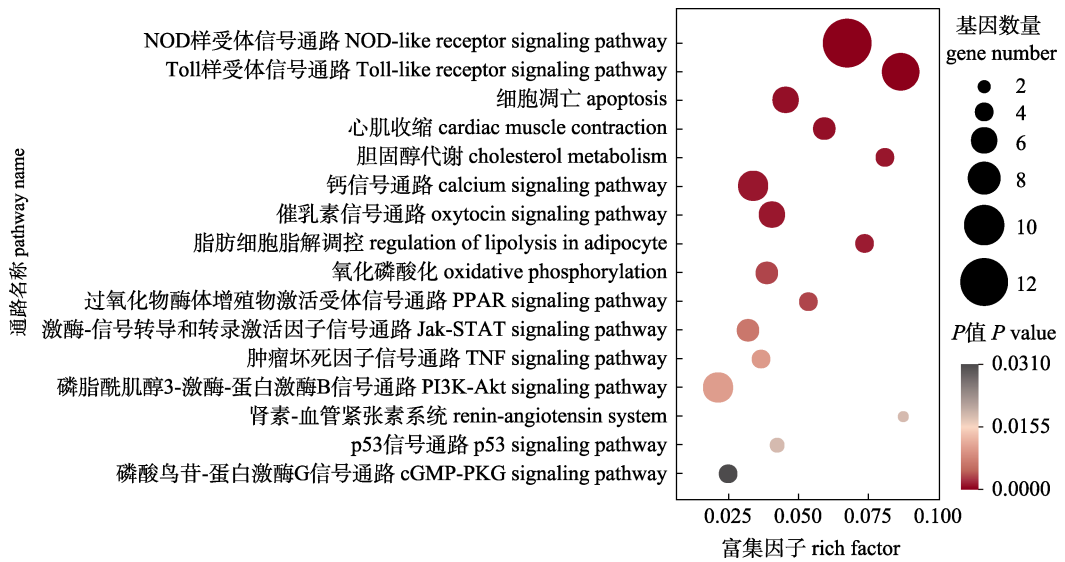


图 6 达里湖瓦氏雅罗鱼种群盐碱适应候选基因 KEGG 分析

Fig. 6 KEGG pathway analysis of candidate genes for alkali-saline adaptation in *Leuciscus waleckii* from Lake Dali Nor populations

2.6 关键调控基因选择压力分析

基于种群强选择信号区域、InDels 差异共享位点、富集的 GO 条目类型(图 5)以及基因生物学功能, *abcc1*、*atp2b1*、*slc4a4*、*slc7a2*、*aqp4* 被确定为与 DL 组离子转运和酸碱平衡等过程相关的关键调控基因, 这 5 个基因都处于强选择信号区域中, 并且都至少包含有一个差异共享 InDel 突变位点(表 3)。将这些关键调控基因首先利用 EasyCodeML 软件进行选择压力分析。对于 *aqp4*, 在 M0 模型下, $\omega=0.39436$, ω 值远小于 1, 没有直接的证据表明存在正向选择(表 4), 比较 M1a 和 M2a, 似然比检验 $2\Delta\ln L=0$, P 值为 1, 说明 M2a

并不优于 M1a, 分别比较 M7 vs M8 和 M8a vs M8, 似然比检验分别为 $2\Delta\ln L=9.44$ ($P<0.01$)和 $2\Delta\ln L=0.015$ ($P>0.5$), 说明 M8a 模型并不优于 M8, 综合表明 *aqp4* 基因主要受到较强的净化选择约束, 具有很高的保守性。对于基因 *atp2b1*, 在模型 M0 中, 进化速率 $\omega_0=1.10315$, 说明 *atp2b1* 在进化过程中受到了一定的正向选择, 对 4 个位点模型进行似然比检验发现, M3、M2a 和 M8 模型明显优于其相应的假设模型 M0、M1a、M7 与 M8a ($P<0.01$)(表 4), 表明各点之间的选择压力存在差异。其余基因 *slc4a4*、*slc7a2*、*abcc1* 的 M0 模型表明它们在进化过程中都受到了不同程度的净化选

表 3 达里湖瓦氏雅罗鱼种群盐碱适应关键调控基因插入/缺失位点信息

Tab. 3 The detailed information of the key regulatory genes and the related InDel loci for alkali-saline adaptation in *Leuciscus waleckii* from Lake Dali Nor (DL) populations

基因名 gene name	位置 position (scaffold_bp)	达里湖种群插入/缺失位点 InDel genotypes in DL	松花江种群插入/缺失位点 InDel genotypes in SH	F_{st}	π ratio (SH/DL)
<i>abcc1</i>	S2372_280895	ins (G->GA)	del (GTTTT->GTTTT)	0.496	6.187
<i>atp2b1</i>	S2457_79727	ins (TG->TGG)	del (TG->T)	0.550	8.135
	S2457_81204	ins (ATA->ATATACTTACTTA)	ins (ATA->ATAT)		
<i>slc4a4</i>	S156_79063	del (GAAT->G)	del (GAA->GA)	0.452	6.632
<i>slc7a2</i>	S2183_243547	ins (A->AT)	ins (AT->ATT)	0.617	11.743
	S2183_253398	ins (CG->CGG)	ins (CGG->CGGG)		
<i>aqp4</i>	S916_402279	del (AC->A)	del (ACTCT->ACT)	0.500	7.771

注: ins 表示插入, del 表示缺失。
Note: ins indicates insertion; del indicates deletion.

表 4 达里湖瓦氏雅罗鱼种群盐碱适应关键调控基因适应性进化分析

Tab. 4 Adaptive evolution analysis of key regulatory genes for alkali-saline adaptation in *Leuciscus waleckii* from Lake Dali Nor (DL) populations

基因 gene	M0	M1a	M2a	M3	M7	M8	M8a
<i>abcc1</i>	$\omega_0=0.93169$	$p_0=0.26242$	$p_0=0.23566$	$p_0=0.17768$	$p=0.11023$	$p_0=0.71712$	$p_0=0.27672$
		$p_1=0.73758$	$p_1=0.55983$	$p_1=0.49885$	$q=0.02848$	$p=0.37729$	$p=1.33245$
		$\omega_0=0.09249$	$p_2=0.20451$	$p_2=0.32346$		$q=0.22021$	$q=9.96033$
		$\omega_I=1.00000$	$\omega_0=0.09014$	$\omega_0=0.03549$		$(p_1=0.28288)$	$(p_1=0.72328)$
			$\omega_1=1.00000$	$\omega_1=0.76786$		$\omega=1.96411$	$\omega=1.00000$
			$\omega_2=2.19622$	$\omega_2=1.91167$			
<i>atp2b1</i>	$\omega_0=1.10315$	$p_0=0.22325$	$p_0=0.18519$	$p_0=0.22534$	$p=0.10641$	$p_0=0.80477$	$p_0=0.22325$
		$p_1=0.77675$	$p_1=0.66442$	$p_1=0.68357$	$q=0.02828$	$p=0.13154$	$p=0.00500$
		$\omega_0=0.00000$	$p_2=0.15039$	$p_2=0.09109$		$q=0.03480$	$q=6.06544$
		$\omega_1=1.00000$	$\omega_0=0.0000$	$\omega_0=0.00000$		$(p_1=0.19523)$	$(p_1=0.77675)$
			$\omega_1=1.00000$	$\omega_1=1.21228$		$\omega=3.17244$	$\omega=1.00000$
			$\omega_2=3.58350$	$\omega_2=4.39876$			
<i>slc4a4</i>	$\omega_0=0.49123$	$p_0=0.50319$	$p_0=0.48233$	$p_0=0.31338$	$p=0.42151$	$p_0=0.93159$	$p_0=0.66534$
		$p_1=0.49681$	$p_1=0.48557$	$p_1=0.52293$	$q=0.45488$	$p=0.45765$	$p=0.65833$
		$\omega_0=0.13607$	$p_2=0.03210$	$p_2=0.16369$		$q=0.54920$	$q=1.86869$
		$\omega_1=1.00000$	$\omega_0=0.13430$	$\omega_0=0.04725$		$(p_1=0.06841)$	$(p_1=0.33466)$
			$\omega_1=1.00000$	$\omega_1=0.56022$		$\omega=2.19614$	$\omega=1.00000$
			$\omega_2=3.09276$	$\omega_2=1.65694$			
<i>slc7a2</i>	$\omega_0=0.49050$	$p_0=0.49617$	$p_0=0.47347$	$p_0=0.35821$	$p=0.40579$	$p_0=0.93723$	$p_0=0.63033$
		$p_1=0.50383$	$p_1=0.48966$	$p_1=0.53030$	$q=0.42093$	$p=0.43899$	$p=0.67283$
		$\omega_0=0.12074$	$p_2=0.03686$	$p_2=0.11150$		$q=0.49897$	$q=2.27849$
		$\omega_1=1.00000$	$\omega_0=0.11839$	$\omega_0=0.05744$		$(p_1=0.06277)$	$(p_1=0.36967)$
			$\omega_1=1.00000$	$\omega_1=0.65093$		$\omega=2.53772$	$\omega=1.00000$
			$\omega_2=3.46105$	$\omega_2=2.06412$			
<i>aqp4</i>	$\omega_0=0.39436$	$p_0=0.56763$	$p_0=0.56763$	$p_0=0.53514$	$p=0.53076$	$p_0=0.58813$	$p_0=0.58857$
		$p_1=0.43237$	$p_1=0.03445$	$p_1=0.09459$	$q=0.63520$	$p=3.47112$	$p=4.59178$
		$\omega_0=0.12665$	$p_2=0.39791$	$p_2=0.37027$		$q=20.77246$	$q=28.01053$
		$\omega_1=1.00000$	$\omega_0=0.12665$	$\omega_0=0.10716$		$(p_1=0.41187)$	$(p_1=0.41143)$
			$\omega_1=1.00000$	$\omega_1=0.84815$		$\omega=1.00000$	$\omega=1.00000$
			$\omega_2=1.00000$	$\omega_2=0.84815$			

注: M0 表示单一比率; M1a 表示近中性; M2a 表示正选择; M3 表示离散; M7 表示 β 分布; M8 表示 β 与 $\omega>1$; M8a 表示 β 与 $\omega=1$ 。
Note: M0 (one ratio), M1a (neutral), M2a (positive selection), M3 (discrete), M7 (β), M8 (β & $\omega>1$), M8a (β & $\omega=1$).

择($\omega<1$), 但 M3、M2a 和 M8 模型明显优于其相应的假设模型 M0、M1a、M7 与 M8a ($P<0.01$), 因此尽管在进化过程中受到不同程度的净化选择, 但在这些基因 M8 模型中也存在着发生正向选择的氨基酸位点。在模型 M8 中, 分别以 EasyCodeML、MEME 和 FEL 计算检测了 *abcc1*、*atp2b1*、*slc4a4*、*slc7a2* 和 *aqp4* 5 个候选基因的正向选择位点, 分别检测到 2、16、16、8 和 0 个强正向选择位点, 其中 *abcc1* 的两个位点达到极显著水平, 其余的位点均为显著水平(表 5)。

2.7 正向选择位点功能分析和蛋白质三维结构预测

将上述正向选择位点所在氨基酸序列与斑马鱼同源序列比对后确定了瓦氏雅罗鱼在斑马鱼正向选择位点的等效位置。*abcc1* 的两个氨基酸位点虽然达到了极显著水平, 但未发现与斑马鱼相关功能结构域及其附近位点的重合区域。*atp2b1* 受选择位点主要集中在阳离子转运 ATP 酶功能域的两端(N 端和 C 端)以及 E1-E2ATP 酶功能域。*slc4a4* 受选择位点与斑马鱼功能重叠区域主要集

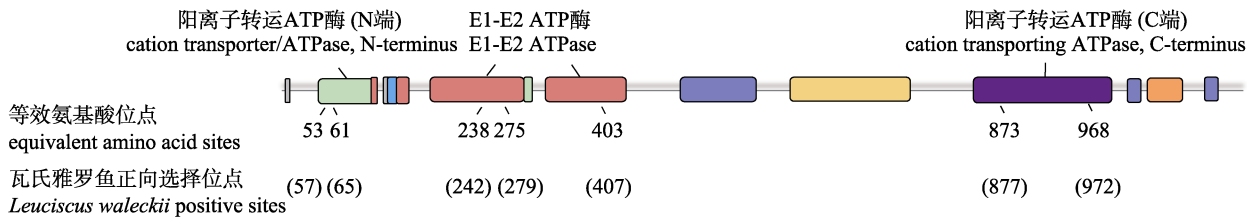
中在带 3 蛋白胞质段和 HCO_3^- 共转运结构域。*slc7a2* 受选择位点主要集中在斑马鱼氨基酸通透酶结构域(图 7)。由于 *slc4a4* 和 *slc7a2* 的部分受选择位点处于共转运和跨膜载体蛋白等关键调控结构域中, 因此本研究将这两个基因的氨基酸序列提交至 I-TASSER 进行三维蛋白结构预测。在

表 5 正向选择氨基酸位点
Tab. 5 Amino acid sites undergoing positive selection

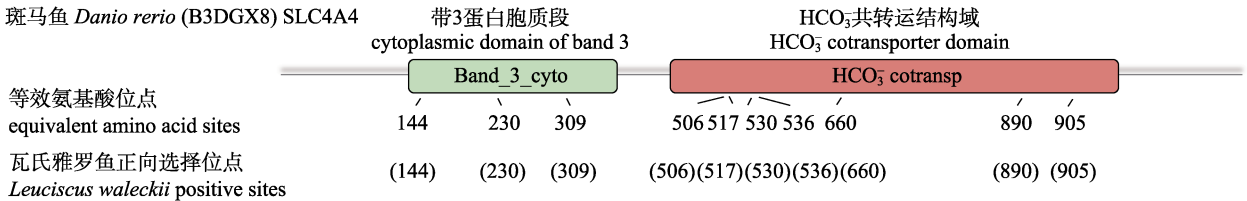
	M0 vs M3	M1a vs M2a	M7 vs M8	M8a vs M8	EasyCodeML (M8) vs MEME vs FEL
	$2\Delta\ln L^{P\text{-value}}$				
<i>abcc1</i>	660.16**	126.71**	138.17**	126.28**	32 th 128 th
<i>atp2b1</i>	144.55**	80.21**	90.10**	78.27**	33 th 39 th 40 th 43 th 57 th 65 th 141 th 151 th 242 th 279 th 328 th 407 th 619 th 840 th 877 th 972 th
<i>slc4a4</i>	1050.6**	60.43**	79.09**	57.41**	4 th 16 th 25 th 51 th 74 th 112 th 144 th 230 th 309 th 506 th 517 th 530 th 536 th 660 th 890 th 905 th
<i>slc7a2</i>	537.53**	36.22**	46.28**	37.47**	1 th 3 th 45 th 51 th 58 th 125 th 224 th 284 th
<i>aqp4</i>	126.95**	0	9.44**	0.015	not allowed

注: M0 表示单一比率; M1a 表示近中性; M2a 表示正选择; M3 表示离散; M7 表示 β 分布; M8 表示 β & $\omega>1$; M8a 表示 β & $\omega=1$. **表示似然比检验 $P<0.01$.
Note: M0 (one ratio), M1a (neutral), M2a (positive selection), M3 (discrete), M7 (β), M8 (β & $\omega>1$), M8a (β & $\omega=1$). ** indicates likelihood ratio $P<0.01$.

斑马鱼 *Danio rerio* (Q5RGZ4) ATP2B1



斑马鱼 *Danio rerio* (B3DGX8) SLC4A4



斑马鱼 *Danio rerio* (Q5PR34) SLC7A2

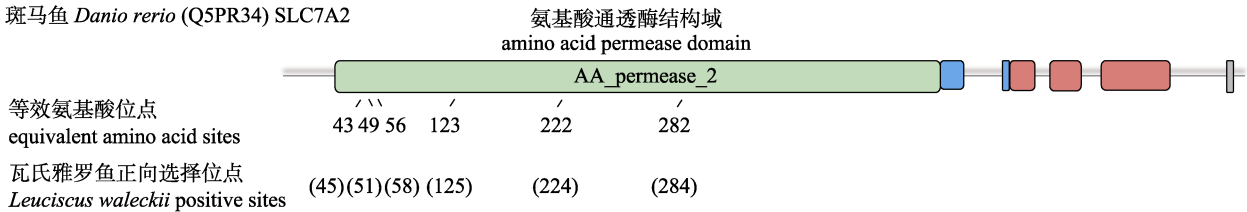


图 7 基于瓦氏雅罗鱼正向选择位点预测斑马鱼功能结构域等效位置

Fig. 7 Prediction of equivalent sites of *Danio rerio*'s functional domains based on positive selection sites of *Leuciscus waleckii*

SLC4A4 蛋白结构中, 检测到的正向选择氨基酸位点 660th (Phe)与配体结合位点重合, 144th (Phe)位点非常接近配体结合位点。在 SLC7A2 蛋白结

构中, 正向选择位点 45th (Leu)接近配体结合位点, 这些位点可能在配体结合中起关键作用, 此外, 在两个蛋白结构预测中均未发现酶激活位点(图 8)。

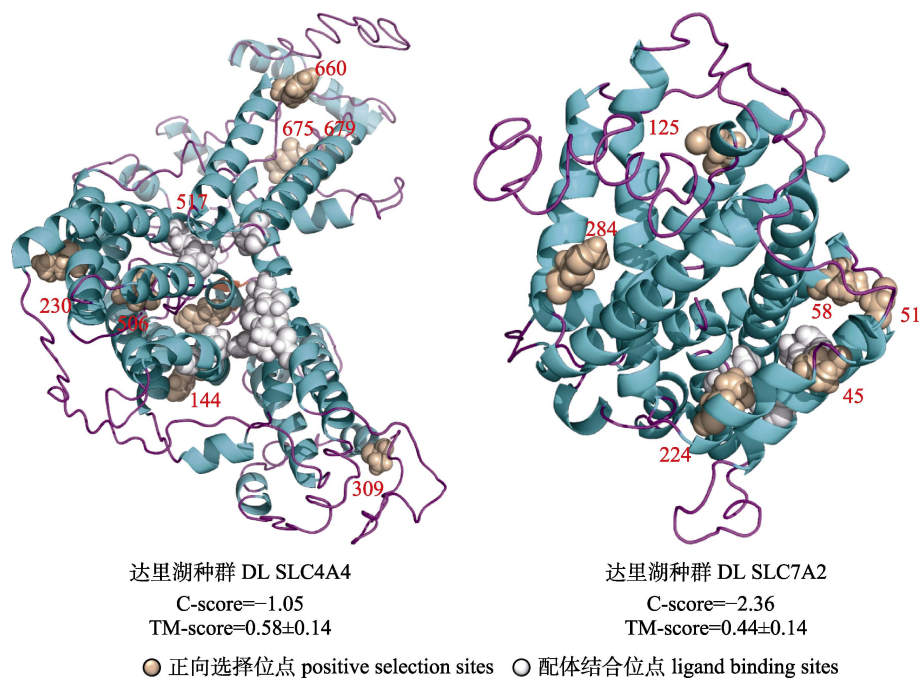


图 8 SLC4A4 和 SLC7A2 三维蛋白结构及正向选择位点所处位置

Fig. 8 Positive selection sites and three dimensional structure of SLC4A4 and SLC7A2

3 讨论

本研究借助于高通量测序技术, 筛选鉴定了瓦氏雅罗鱼两个生态种群的 InDels 位点, 得到的 InDels 数量与类型基本一致, 说明两个种群基因组上的 InDels 分布均匀, 没有因为起源背景或生活环境差异呈现偏态。总体来看, InDels 插入数大于缺失数, 数量集中在 1~8 bp, 其中 1~3 bp 最多。此外, 两个种群 InDels 注释变体大部分分布于基因间区和内含子等非编码区域, 编码区中分布的位点最少, 这与人类^[27]、荷斯坦牛^[28]、藏酋猴^[29]基因组数据上观察到的结果类似。

在 InDel 注释结果的基础上, 通过每个种群个体间共享 InDels 位点信息筛选得到瓦氏雅罗鱼达里湖种群差异共享 InDels 位点, 通过与前期基于 SNP 位点筛选得到的差异基因组区域整合^[5], 确定了在基因组上一些可能与盐碱适应相关的候选基因。生物体对非生物胁迫的应答通常涉及多

种机制^[30], 筛选得到的候选基因在功能富集分析中显著富集于几个不同的功能区域, 这与瓦氏雅罗鱼盐碱适应机制复杂、多样、系统性强的特点相符^[11]。

通路富集分析发现与离子转运相关的基因显著富集, 其编码蛋白在细胞膜的流动性和渗透性方面发挥重要作用。如受正向选择压力较强的 *slc7a2* 和 *slc4a4*。研究显示, SLC7A2 作为渗透酶参与跨质膜转运和吸收阳离子氨基酸, 这种不依赖于 Na⁺ 的渗透调节系统被证实可通过氨基酸的有效转运使哺乳动物能够在胁迫环境下维持细胞的正常功能^[31]。NBC1 是基因 *slc4a4* 的表达产物, 属于 Na⁺/HCO₃⁻ 共转运蛋白, 是重要的 HCO₃⁻ 转运家族成员之一, 对维持鱼类血液酸碱平衡发挥重要作用^[32-33]。相比于淡水种, 达里湖种群受到了包括总碱度、pH 值以及氨氮代谢变化产生的协同毒性^[11], 使得血液中 HCO₃⁻ 浓度增加, 破坏体内酸碱平衡。课题组前期研究发现, 在相同碱度胁迫

迫下,随着胁迫时间的延长,*NBC1* mRNA 在达里湖种群中显著高表达,而在松花江种群表达量显著降低^[34],表明达里湖瓦氏雅罗鱼通过积极主动地排碱保持体内离子和酸碱平衡。关键调控基因 *atp2b1* 编码质膜钙泵蛋白,能够将 Ca^{2+} 从细胞质转运到细胞外来维持跨质膜的钙浓度梯度,更重要的是,质膜钙泵与渗透调节过程密切相关^[35],一项尼罗罗非鱼盐度耐受实验表明,在 0~32 的盐度递增条件下尼罗罗非鱼鳃和脑组织 *atp2b1* 和 Na^+/K^+ ATPase (NKA) mRNA 均显著高表达,从而实现在高渗环境下 Na^+/Cl^- 转运和 Ca^{2+} 内流^[36]。课题组前期关于达里湖瓦氏雅罗鱼的碱度耐受实验同样发现 *NKA* 随着碱度耐受时间的延长呈现显著高表达^[34],因此 ATP2B1 参与的 Ca^{2+} 泵激活可能与 *NKA* 的表达变化存在密切关系。此外,水体 pH 的升高将导致 H^+ 浓度的降低,使得鱼体排氨 ($\text{NH}_3 + \text{H}^+ = \text{NH}_4^+$) 受阻,最终导致“氨中毒”^[37],本研究鉴定得到的关键调控基因 *abcc1* 的表达产物是一种转运谷胱甘肽结合物的膜转运蛋白^[38],在欧洲舌齿鲈 (*Dicentrarchus labrax*) 和尼罗罗非鱼中,ABCC1 定位在细胞底端,将内容物运输至血液,对排出细胞内积累的环境毒物以及组织防御等过程至关重要^[39-41]。达里湖瓦氏雅罗鱼能够对抗 P_{NH_3} 梯度继续主动排氨^[11],推测 ABCC1 对于防止细胞内氨中毒发挥重要作用,但细胞解毒机制以及与氨氮代谢之间的关系仍需进一步探索。

通路分析还筛选到与免疫过程相关的基因显著富集,如 *xiap*、*ube2l*、*ube2d2*、*psme4b* 属于泛素-蛋白酶体系统,这一系统提供了控制蛋白质质量的基本方式,即增加蛋白的水解作用,加快蛋白质的更新速率,从而有选择性地清除在极端环境下由于生物合成错误、氧自由基或变性而引起的异常折叠或损坏^[42],表明修复细胞和清除坏死的细胞器也是瓦氏雅罗鱼盐碱适应的重要机制。

除了离子交换、转运及免疫的基因显著富集外,本研究还发现部分候选基因富集到了钙离子信号通路,包括编码钙黏蛋白 (*cdh4*, *cdh7*, *cdh11*, *cdhr1*), 紧密连接蛋白 (*cldn1*) 和细胞骨架蛋白 (*camsap1*) 相关的基因。其中 *cldn1* 在生物体处于压力环境时能够通过维持上皮细胞的完整性来充

当物理屏障,防止溶质和水自由通过上皮细胞和内皮细胞之间的细胞间隙。*camsap1* 编码细胞骨架蛋白,当细胞暴露于应激环境时,细胞可通过快速重组各种细胞骨架网络从而做出应答^[43]。

适应性进化(正向选择)是通过环境变化对蛋白质施加选择压力,以提高生物体在局域环境中的适应性^[44]。本研究推测达里湖瓦氏雅罗鱼在盐碱适应的进化过程中,可能通过调控基因的 SNP 或 InDel 等序列多态性,促使基因编码或调控序列的功能发生变化,以此达到对极端环境的适应。因此,除了对盐碱适应相关基因的功能研究外,本研究还重点关注了这些抗性相关基因在进化过程中受到的自然选择压力,对参与离子转运和酸碱平衡过程的关键调控基因进化速率进行了分析。笔者在选择压力检测分析中首先采用了 EasyCodeML 中的位点模型,因为在自然界的进化过程中,选择压力可能只是作用于某些位点,而不是作用于整个基因中^[45]。研究结果显示,*slc4a4*、*slc7a2*、*abcc1* 虽然主要经历了净化选择,但仍检测到小部分的正向选择氨基酸位点,并且 *slc4a4* 和 *slc7a2* 部分正向选择位点处于 HCO_3^- 共转运蛋白和氨基酸渗透关键功能结构域中。推测虽然整条基因在进化过程中受到较强的净化选择,但某些功能结构域的正向选择位点可能与盐碱胁迫环境的适应性进化、基因结构的改变有着重要的联系,盐碱环境胁迫可能是进化的主要驱动力。

4 结论

本研究对瓦氏雅罗鱼达里湖种群和松花江种群各 5 尾个体进行了基因组重测序,共鉴定得到 983528 个 InDels 位点。通过比较两个种群之间的差异 InDels 位点并整合已知的基因组强选择信号区域,筛选得到与盐碱适应相关的 325 个 InDels 位点及其关联候选基因 176 个。富集分析显示候选基因主要在离子转运、酸碱平衡及炎症免疫反应等生物学过程发挥重要作用。进一步的选择压力分析显示,关键调控基因 *abcc1*、*atp2b1*、*slc4a4*、*slc7a2* 和 *aqp4* 在进化过程中净化选择均起主导作用,但处于关键功能结构域的正向选择状态位点可能有助于瓦氏雅罗鱼达里湖种群快速

适应极端盐碱环境。本研究结果将为瓦氏雅罗鱼盐碱适应相关基因或标记的开发提供重要参考。

参考文献:

- [1] Chang Y M, Tang R, Sun X W, et al. Genetic analysis of population differentiation and adaptation in *Leuciscus waleckii*[J]. *Genetica*, 2013, 141(10-12): 417-429.
- [2] Geng K, Zhang Z C. The geomorphic characteristics and evolution of the lakes in Dalairuoe area of Neimenggu Plateau during the Holocene[J]. *Journal of Beijing Normal University (Natural Science)*, 1988, 24(4): 94-101. [耿侃, 张振春. 内蒙古达尔诺尔地区全新世湖群地貌特征及其演化[J]. *北京师范大学学报(自然科学版)*, 1988, 24(4): 94-101.]
- [3] Chi B J, Chang Y M, Yan X C, et al. Genetic variability and genetic structure of *Leuciscus waleckii* Dybowski in Wusuli River and Dali Lake[J]. *Journal of Fishery Sciences of China*, 2010, 17(2): 228-235. [池炳杰, 常玉梅, 闫学春, 等. 瓦氏雅罗鱼达里湖群体和乌苏里江群体的遗传多样性和遗传结构分析[J]. *中国水产科学*, 2010, 17(2): 228-235.]
- [4] Chi B J. Genetic variability of *Leuciscus waleckii* (Dybowski) and construction of the cDNA subtractive library of saline-alkaline tolerance[D]. Shanghai: Shanghai Ocean University, 2010: 12-17. [池炳杰. 瓦氏雅罗鱼遗传多样性分析和耐盐碱 cDNA 消减文库的构建[D]. 上海: 上海海洋大学, 2010: 12-17.]
- [5] Wang S Y, Kuang Y Y, Liang L Q, et al. Resequencing and SNP discovery of Amur ide (*Leuciscus waleckii*) provides insights into local adaptations to extreme environments[J]. *Scientific Reports*, 2021, 11: 5064.
- [6] Xu J, Li J T, Jiang Y L, et al. Genomic basis of adaptive evolution: The survival of Amur ide (*Leuciscus waleckii*) in an extremely alkaline environment[J]. *Molecular Biology and Evolution*, 2017, 34(1): 145-159.
- [7] Adedze Y M N, Lu X, Xia Y C, et al. Agarose-resolvable InDel markers based on whole genome re-sequencing in cucumber[J]. *Scientific Reports*, 2021, 11: 3872.
- [8] Misra G, Badoni S, Anacleto R, et al. Whole genome sequencing-based association study to unravel genetic architecture of cooked grain width and length traits in rice[J]. *Scientific Reports*, 2017, 7: 12478.
- [9] Grobet L, Royo Martin L J, Poncelet D, et al. A deletion in the bovine myostatin gene causes the double-muscling phenotype in cattle[J]. *Nature Genetics*, 1997, 17(1): 71-74.
- [10] Chen M Y, Yan H L, Wang K, et al. Goat SPEF2: Expression profile, indel variants identification and association analysis with litter size[J]. *Theriogenology*, 2019, 139: 147-155.
- [11] Chang Y M, Liang L Q. Advances of research of physiological and molecular mechanisms related to alkali-saline adaptation for fish species inhabiting alkali-saline water[J]. *Journal of Fisheries of China*, 2021, 45(5): 798-812. [常玉梅, 梁利群. 耐盐碱鱼类的生理和分子机制研究进展[J]. *水产学报*, 2021, 45(5): 798-812.]
- [12] Langmead B, Trapnell C, Pop M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome[J]. *Genome Biology*, 2009, 10(3): R25.
- [13] Patel R K, Jain M. NGS QC Toolkit: A toolkit for quality control of next generation sequencing data[J]. *PLoS One*, 2012, 7(2): e30619.
- [14] Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data[J]. *Bioinformatics*, 2011, 27(21): 2987-2993.
- [15] 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.
- [16] Cingolani P, Platts A, Wang L L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff[J]. *Fly*, 2012, 6(2): 80-92.
- [17] Xie C, Mao X Z, Huang J J, et al. KOBAS 2.0: A web server for annotation and identification of enriched pathways and diseases[J]. *Nucleic Acids Research*, 2011, 39(suppl_2): W316-W322.
- [18] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes[J]. *Nucleic Acids Research*, 2000, 28(1): 27-30.
- [19] Kanehisa M. Toward understanding the origin and evolution of cellular organisms[J]. *Protein Science*, 2019, 28(11): 1947-1951.
- [20] Kumar S, Stecher G, Li M, et al. MEGA X: Molecular evolutionary genetics analysis across computing platforms[J]. *Molecular Biology and Evolution*, 2018, 35(6): 1547-1549.
- [21] Chen C J, Chen H, Zhang Y, et al. TBtools: An integrative toolkit developed for interactive analyses of big biological data[J]. *Molecular Plant*, 2020, 13(8): 1194-1202.
- [22] Gao F L, Chen C J, Arab D A, et al. EasyCodeML: A visual tool for analysis of selection using CodeML[J]. *Ecology and Evolution*, 2019, 9(7): 3891-3898.
- [23] Finn R D, Clements J, Eddy S R. HMMER web server: Interactive sequence similarity searching[J]. *Nucleic Acids Research*, 2011, 39(suppl_2): W29-W37.
- [24] Mistry J, Chuguransky S, Williams L, et al. Pfam: The protein families database in 2021[J]. *Nucleic Acids Research*,

- 2021, 49(D1): D412-D419.
- [25] Yang J Y, Zhang Y. I-TASSER server: New development for protein structure and function predictions[J]. *Nucleic Acids Research*, 2015, 43(W1): W174-W181.
- [26] Zhang Y. I-TASSER server for protein 3D structure prediction[J]. *BMC Bioinformatics*, 2008, 9: 40.
- [27] Fujimoto A, Nakagawa H, Hosono N, et al. Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing[J]. *Nature Genetics*, 2010, 42(11): 931-936.
- [28] Jiang J P, Gao Y H, Hou Y L, et al. Whole-genome resequencing of Holstein bulls for indel discovery and identification of genes associated with milk composition traits in dairy cattle[J]. *PLoS One*, 2016, 11(12): e0168946.
- [29] Fan Z X, Zhao G, Li P, et al. Whole-genome sequencing of Tibetan macaque (*Macaca thibetana*) provides new insight into the macaque evolutionary history[J]. *Molecular Biology and Evolution*, 2014, 31(6): 1475-1489.
- [30] White C R, Alton L A, Frappell P B. Metabolic cold adaptation in fishes occurs at the level of whole animal, mitochondria and enzyme[J]. *Proceedings of the Royal Society B: Biological Sciences*, 2012, 279(1734): 1740-1747.
- [31] Ellingsen S, Narawane S, Fjose A, et al. Sequence analysis and spatiotemporal developmental distribution of the Cat-1-type transporter *slc7a1a* in zebrafish (*Danio rerio*)[J]. *Fish Physiology and Biochemistry*, 2020, 46(6): 2281-2298.
- [32] Esbaugh A J, Heuer R, Grosell M. Impacts of ocean acidification on respiratory gas exchange and acid-base balance in a marine teleost, *Opsanus beta*[J]. *Journal of Comparative Physiology B*, 2012, 182(7): 921-934.
- [33] Lee Y C, Yan J J, Cruz S A, et al. Anion exchanger 1b, but not sodium-bicarbonate cotransporter 1b, plays a role in transport functions of zebrafish H⁺-ATPase-rich cells[J]. *American Journal of Physiology Cell Physiology*, 2011, 300(2): C295-C307.
- [34] Chang Y M, Zhao X F, Liew H J, et al. Effects of Bicarbonate Stress on Serum Ions and Gill Transporters in Alkali and Freshwater Forms of Amur Ide (*Leuciscus waleckii*)[J]. *Frontiers in physiology*, 2021, 12: 676096.
- [35] Verboost P M, Schoenmakers T J, Flik G, et al. Kinetics of ATP- and Na⁺-gradient driven Ca²⁺ transport in basolateral membranes from gills of freshwater- and seawater-adapted tilapia[J]. *The Journal of Experimental Biology*, 1994, 186: 95-108.
- [36] Rengmark A H, Slettan A, Lee W J, et al. Identification and mapping of genes associated with salt tolerance in tilapia[J]. *Journal of Fish Biology*, 2007, 71(sc): 409-422.
- [37] Chen S G, Hou J L, Yao N, et al. Comparative transcriptome analysis of *Triplophysa yarkandensis* in response to salinity and alkalinity stress[J]. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics*, 2020, 33: 100629.
- [38] Keppler D, Cui Y, König J, et al. Export pumps for anionic conjugates encoded by MRP genes[J]. *Advances in Enzyme Regulation*, 1999, 39: 237-246.
- [39] Ferreira M, Costa J, Reis-Henriques M A. ABC transporters in fish species: A review[J]. *Frontiers in Physiology*, 2014, 5: 266.
- [40] Costa J, Reis-Henriques M A, Wilson J M, et al. P-glycoprotein and CYP1A protein expression patterns in Nile tilapia (*Oreochromis niloticus*) tissues after waterborne exposure to benzo(a)pyrene (BaP)[J]. *Environmental Toxicology and Pharmacology*, 2013, 36(2): 611-625.
- [41] Ferreira M, Santos P, Rey-Salgueiro L, et al. The first demonstration of CYP1A and the ABC protein(s) gene expression and activity in European seabass (*Dicentrarchus labrax*) primary hepatocytes[J]. *Chemosphere*, 2014, 100: 152-159.
- [42] Benedito-Palos L, Ballester-Lozano G F, Simó P, et al. Lasting effects of butyrate and low FM/FO diets on growth performance, blood haematology/biochemistry and molecular growth-related markers in gilthead sea bream (*Sparus aurata*)[J]. *Aquaculture*, 2016, 454: 8-18.
- [43] Kim S, Coulombe P A. Emerging role for the cytoskeleton as an organizer and regulator of translation[J]. *Nature Reviews Molecular Cell Biology*, 2010, 11(1): 75-81.
- [44] Morgan C C, Loughran N B, Walsh T A, et al. Positive selection neighboring functionally essential sites and disease-implicated regions of mammalian reproductive proteins[J]. *BMC Evolutionary Biology*, 2010, 10: 39.
- [45] Kamath P L, Getz W M. Adaptive molecular evolution of the Major Histocompatibility Complex genes, *DRA* and *DQA*, in the genus *Equus*[J]. *BMC Evolutionary Biology*, 2011, 11: 128.

Mining and analysis of InDels in response to alkali-saline stress in Amur ide (*Leuciscus waleckii*)

WANG Shuangyi^{1,2}, LIANG Liquan², CHANG Yumei², SUN Bo²

1. College of Fisheries and Life Sciences, Shanghai Ocean University, Shanghai 201306, China;
2. National & Local United Engineering Laboratory of Freshwater Fish Breeding; Key Laboratory of Freshwater Aquatic Biotechnology and Genetic Breeding, Ministry of Agriculture and Rural Affairs; Heilongjiang Province's Key Laboratory of Fish Stress Resistance Breeding and Germplasm Characteristics on Special Habitats, Heilongjiang River Fisheries Research Institute, Chinese Academy of Fishery Sciences, Harbin 150070, China

Abstract: To uncover the adaptative mechanism of the alkaline Amur ide (*Leuciscus waleckii*) population in Lake Dali Nor, Inner Mongolia, China, five samples of two populations, including one alkali form from Lake Dali Nor (DL), and one freshwater form from the Songhua River (SH), Heilongjiang, China, i.e., its historical origin, were analyzed using whole-genome resequencing technology. A total of 983528 nonredundant short insertions and deletions (InDels, 1–29 base pairs) were obtained. Among them, 8176 InDels that were polymorphic between the DL and SH populations were identified. Then, using integrated analysis of the known strong positive selection signal regions between the DL and SH populations, we identified a total of 176 potential candidate genes potentially affecting alkali-saline adaptation. Enrichment analysis showed that the genes were enriched mainly in inflammatory immune responses, ion transport, and osmotic regulation. Furthermore, the candidate genes *abcc1*, *atp2b1*, *slc4a4*, *slc7a2*, and *aqp4* for key regulatory processes of alkali-saline adaptation, such as osmotic regulation and acid-base balance, were identified via selective pressure analysis using EasyCodeML, MEME, and FEL. Analysis showed that purifying selection played a dominative role for all of these genes, but a few amino sites could be identified to be under positive selection in the selection model. Prediction of three-dimensional structure revealed that some sites of the two proteins SLC4A4 and SLC7A2 coincided or were very close to the ligand-binding sites, indicating that the positive selection pressure mediated by the alkali-saline stress may contribute to the development of rapidly adaptive evolution in Amur ide. Our findings provide a basis for further study and reveal key genes for alkali-saline adaptation in Amur ide.

Key words: *Leuciscus waleckii*; whole-genome resequencing; insertions and deletions (InDels); alkali-saline adaptation; selective pressure; adaptive evolution

Corresponding author: CHANG Yumei. E-mail: changyumei@hrfri.ac.cn