中国水产科学 2022 年 2 月, 29(2): 245–251
Journal of Fishery Sciences of China

研究论文

DOI: 10.12264/JFSC2021-0231

简单回归尺度转换实现半滑舌鳎性逆转基因的高效定位

黄岩[1, 3], 宋禹昕[2, 3], 蒋丽[3], 杨润清[3]

1. 上海海洋大学水产科学国家级实验教学示范中心, 上海 201306;
2. 南京农业大学无锡渔业学院, 江苏 无锡 214081;
3. 中国水产科学研究院生物技术研究中心, 北京 100141

摘要: 在间断性状全基因组关联分析中, 当基因组数据存在复杂群体分层时, 广义线性模型需要同时考虑上百个协变量, 其求解速度会大大下降而且还会产生异常解。本研究目的是把简单回归结果中显著位点的效应值和遗传力的尺度转化为可解释的广义线性回归结果。首先对亲缘关系矩阵进行谱分解, 特征向量作为主成分(PC), 矫正间断性状中的群体分层; 再求解每一个主成分的回归系数, 并将众多协变量与其各自的回归系数相乘, 得到的乘积合并为一个新的协变量; 然后将它作为简单回归的协变量, 逐个对标记进行关联检验; 最后对筛选获得的候选数量性状核苷酸(QTN)进行广义线性模型回归分析, 将效应和方差转化为广义线性回归模型尺度。采用本研究提出的方法与直接考虑主成分的广义线性回归模型, 分别对半滑舌鳎(*Cynoglossus semilaevis*)的性逆转性状进行全基因组关联分析, 结果表明, 本研究方法的 QTN 检测效率更高, 共检测出 6 个 QTN, 其中 5 个 QTN 位于 Z 染色体上, 1 个 QTN 位于 W 染色体上, 并且在基因组控制方面, 本研究方法的基因组控制值与直接考虑 PC 的广义线性回归模型的基因组控制值相同, 均为 1.01, 处于较优水平。结论认为, 基于主成分分析的简单回归尺度转换方法能够在保证准确率的情况下提升QTN的检测效率, 实现间断性状快速稳健的全基因组关联分析, 同时检测出的QTN能为半滑舌鳎性逆转性状的研究提供理论指导。

关键词: 全基因组关联分析; 半滑舌鳎; 性逆转; 主成分; 尺度转换; 简单回归模型; 广义线性回归模型
中图分类号: S917 **文献标志码:** A **文章编号:** 1005–8737–(2022)02–0245–07

全基因组关联分析(genome-wide association study, GWAS)是在全基因组范围内, 以数以万计的单核苷酸多态性(single nucleotide polymorphism, SNP)作为分子标记, 探究基因变异与性状之间关联的研究。GWAS 已成为当今基因定位的主流方法, 广泛应用于人类疾病、动植物疾病、动植物性状改良、动植物遗传育种等方面。由于 GWAS 的样本来源于不同的种群, 每个种群独特的遗传历史、交配习惯、繁殖扩张以及随机变异都会使个体之间等位基因频率产生差异[1-2], 存在群体分层, 会导致关联分析过程中检验统计量膨胀, 从而降低对数量性状核苷酸(quantitative trait nucleotide, QTN)的检测效力。主成分分析法(principal component analysis, PCA)是当下应对该问题的主要策略[3-4]。

间断性状, 即类似于性别、是否患病等表型, 可用二进制或者是不连续数字表示的性状。在进行基因定位时, 间断性状和连续性状一样, 也需考虑除检验标记外的群体分层、家系结构和隐藏亲缘等混杂因素[5-6]。然而与呈现正态分布的连续

收稿日期: 2021-05-14; **修订日期:** 2021-05-25.
基金项目: 国家重点研发计划“蓝色粮仓科技创新”重点专项(2018YFD0900201); 中央公益性科研院所基本科研业务费专项资金项目(2019ZY09).
作者简介: 黄岩(1994–), 女, 硕士研究生, 研究方向为数量遗传学. E-mail: ayan0827@163.com
通信作者: 蒋丽, 副研究员, 研究方向为数量遗传学, E-mail: jiangl@cafs.ac.cn; 杨润清, 研究员, 研究方向为数量遗传学, E-mail: runqingyang@cafs.ac.cn

性状用剩余误差作为校正后的表型不同, 间断性状的表型形式较为特殊, 无法像连续性状一样估计表型的剩余误差, 这就需要引入广义线性模型(generalized linear model, GLM)中的 logit 回归, 对间断性状进行 GWAS。即使校正了一些固定效应协变量, logit 回归仍会受到群体分层的影响导致检验统计量膨胀。与线性模型(linear model, LM)相比较, 广义线性模型求解要消耗更多时间和计算机内存。当协变量比较多的时候, GLM 还会产生由近似引起的严重偏差问题[7-8]。

半滑舌鳎(*Cynoglossus semilaevis*)在动物分类上属于鲽形目(pleuronectiformes)、舌鳎科(cynoglossidae)、舌鳎属, 属于近海大型底栖暖温性动物, 分布于中国、朝鲜、日本, 在中国主要分布在黄海、渤海近海区域[9], 是目前中国重要的海水养殖鱼类品种之一[10]。不同性别的半滑舌鳎成体体重差异较大, 雌性舌鳎生长速度较快, 体重较大, 相比之下, 雄性舌鳎的体重较轻, 生长缓慢。在人工养殖过程中, 生理性别是雌性的舌鳎仅占 20%~35%[11]; 此外, 半滑舌鳎是性逆转生物, 即在发育过程中, 生理性别可能发生改变, 即由遗传雌性逆转为生理雄性, 该现象在昆虫[12]、爬行动物[13]、两栖动物[13]和鱼类[14]中时有发生。已有研究认为半滑舌鳎在 Z 染色体上可能存在导致性逆转的基因[15]。出于经济利益的考虑, 在半滑舌鳎养殖过程中控制性逆转的产生, 提高雌鱼的比例尤为重要。

半滑舌鳎性逆转性状在数量遗传学中被认为是间断性状, 对该类性状 QTN 的筛选通常采用广义线性回归分析, 对于存在群体分层的群体, 则可采用考虑主成分(principal component, PC)的广义线性回归模型进行基因定位。为了更好地校正存在于半滑舌鳎性逆转性状中的复杂群体分层, 以及提高 QTN 检测效力, 本研究将提出一种间断性状 GWAS 方法, 该方法可以将简单回归模型结果中显著位点的效应值和遗传力的尺度, 转化为可解释的广义线性回归模型结果, 降低了广义线性回归模型的求解难度和异常解出现的概率。本研究通过使用上述策略对半滑舌鳎性逆转性状进行 GWAS, 并与直接考虑 PC 的广义线性回归模型 GWAS 结果作比较, 检验本研究提出的方法对群体分层的校正效果以及 QTN 检测效力, 同时对检测出的影响半滑舌鳎性逆转性状的多态性核苷酸位点进行分析阐述, 为以后半滑舌鳎性逆转性状的研究提供理论依据。

1 材料与方法

1.1 实验群体基因型与表型的获取

本研究所用半滑舌鳎实验群体育成于河北黄骅。2013 年 3 月, 随机选取 6 尾亲本雌鱼和 11 尾表型雄鱼作为亲本, 混合养殖, 亲本间随机交配繁殖, 产生后代。半滑舌鳎性逆转通常发生在孵化后的前 90 d, 为避免低温性逆转现象发生, 对子代幼鱼进行持续 90 d、恒温 22 ℃的养殖。90 d 之后, 随机选取 268 尾半滑舌鳎进行 DNA 提取、亲代分析、遗传和表型性别检测, 去除缺失数据, 115 尾雌性遗传基因被用于 GWAS。采集这 115 只雌性半滑舌鳎的鱼鳍和性腺, 并用鱼鳍提取 DNA, 提取鱼鳍组织 DNA 的具体操作步骤在 Jiang 等[16]的文章中有详细的描述。采用 Chen 等[17]提出的方法确定遗传性别, 利用遗传雌性的生殖腺组织切片进行表型性别鉴定, 得到表型数据。基因型数据由上海欧易生物技术有限公司测序分析得到, 使用 Illumina HiSeq2500 平台对 115 尾遗传型雌性的半滑舌鳎的鱼鳍 DNA 进行单端测序, 采用 2b-RAD 方法对测序结果进行基因分型, 质量控制后, 最终获得共 17618 个 SNP 标记。

1.2 方法

1.2.1 广义线性回归关联检验 考虑群体分层时, 通常以亲缘关系矩阵的主成分去校正由群体分层导致的分层效应, 根据广义线性模型理论, 连接函数建立起了间断性状表型向量 $\boldsymbol{y}$ (0 或 1)的期望和被检验标记效应的关系:

$$\ln\left(\frac{\mu}{1-\mu}\right)=\boldsymbol{Xb}+\boldsymbol{z}\beta_1 \tag{1}$$

式中, μ 是 $\boldsymbol{y}$ 的期望, $\boldsymbol{Xb}$ 是固定效应项, β_1 为当前检验标记的加性遗传效应, $\boldsymbol{z}$ 为所对应的指示变量向量。

通常, 为了简化计算, 事先估计出 $\boldsymbol{X}\hat{\boldsymbol{b}}$ 并将它作为已知的回归项, 此时只需要用 $\boldsymbol{y}^*-\boldsymbol{X}\hat{\boldsymbol{b}}$ 代替 $\boldsymbol{y}$

求解标记的效应:

$$\hat{\boldsymbol{\beta}}_{\mathrm{l}} = (\boldsymbol{z}^T w \boldsymbol{z})^{-1} \boldsymbol{z}^T w (\boldsymbol{y}^* - \boldsymbol{X}\hat{\boldsymbol{b}}) \tag{2}$$

式中, $w=\mu(1-\mu)$是权重, $\boldsymbol{y}^* = \boldsymbol{z}\boldsymbol{\beta}_{\mathrm{l}} + \frac{\boldsymbol{y}-\mu}{w}$ 是由表型值和回归系数组成的新的因变量。

采用卡方统计量, 推断 SNP 与间断性状关联:

$$\chi^2 = \frac{\hat{\boldsymbol{\beta}}_{\mathrm{l}}^2}{(\boldsymbol{z}^T w \boldsymbol{z})^{-1}} \tag{3}$$

这个统计量服从自由度为 1 的卡方分布。

1.2.2　简单回归关联检验　考虑群体分层时, 通常以亲缘关系矩阵的特征向量作为基因组简单线性回归模型的协变量去校正由群体分层导致的分层效应, 简单回归模型变为:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{z}\boldsymbol{\beta}_{\mathrm{c}} + e \tag{4}$$

式中, $\boldsymbol{\beta}_{\mathrm{c}}$是简单回归模型求出的效应值, $\boldsymbol{\beta}_{\mathrm{c}}$可以直接用最小二乘法求解, $e \sim N(0, \sigma_e^2)$是回归的误差项。采用卡方统计量, 推断 SNP 与间断性状关联:

$$\chi^2 = \frac{\hat{\boldsymbol{\beta}}_{\mathrm{c}}^{\,2}}{(\boldsymbol{z}^T \boldsymbol{z})^{-1} \hat{\sigma}_e^2} \tag{5}$$

这个统计量服从自由度为 1 的卡方分布。

1.2.3　关联检验尺度转换　当广义线性模型的回归系数多, 群体小或标记等位基因频率较低时, 广义线性模型很难收敛, 理论和实践证明, GLM 与 LM 具有相同的统计效力[18], 所以本研究先采用(4)中的简单回归模型对每个标记进行检验, 并以 Bonferroni 矫正阈值为标准, 筛选候选 QTN。假设检测到了 q 个候选 QTN, 将这 q 个标记指示变量 $\boldsymbol{z}$ 逐个代入考虑群体分层的广义线性模型中重新求解 $\boldsymbol{\beta}_{\mathrm{l}}$, 将新求解出的 $\boldsymbol{\beta}_{\mathrm{l}}$代替简单回归模型求解出的 $\boldsymbol{\beta}_{\mathrm{c}}$, 这样标记的效应值就从简单回归尺度转化到了广义线性回归尺度。此时, 总检测到 QTN 的遗传力是 $\frac{\sum_{1}^{q}\mathrm{Var}(\boldsymbol{z}_i\boldsymbol{\beta}_{\mathrm{l}i})}{\sum_{1}^{q}\mathrm{Var}(\boldsymbol{z}_i\boldsymbol{\beta}_{\mathrm{l}i})+1}$, 其中第 i 个 QTN 的遗传力是 $\frac{\mathrm{Var}(\boldsymbol{x}_i\boldsymbol{\beta}_{\mathrm{l}i})}{\sum_{1}^{q}\mathrm{Var}(\boldsymbol{x}_i\boldsymbol{\beta}_{\mathrm{l}i})+1}$。

2　结果与分析

在 GWAS 中, Q-Q 图是算法统计性质的直接体现; 曼哈顿图常用于显著 QTN 的定位和筛选, 分布于水平线即 5% Bonferroni 校正阈值线(2.838×10^{-6})之上的点即为显著 QTN; 基因组控制(genomic control, GC)值可以直接显示是否存在群体分层, GC 值接近 1 说明基因组控制效果较好。

2.1　基因组广义线性回归模型的 GWAS 结果

图 1 所示是半滑舌鳎性逆转性状考虑复杂群体分层的广义线性回归模型的 GWAS 结果。观察 Q-Q 图可以得知, 1 个 PC 和 3 个 PC 的 Q-Q 图前半段均未贴合理论线, 5 个 PC 的 Q-Q 图的前半段能够很好地贴合理论线, 后半部分上扬; 而且 1 个 PC 时的 GC 值为 2.37, 3 个 PC 时的 GC 值为 1.64, 5 个 PC 时的 GC 值为 1.01, GC 值和 Q-Q 图结果都显示, 1 个 PC 和 3 个 PC 时的群体结构未得到很好的校正, 存在假阳性。考虑 5 个 PC 的广义线性回归模型的 GC 值接近 1, 基因组控制效果较好, 观察曼哈顿图可见, 这时在 Z 染色体有 2 个 QTN 超过阈值线。

2.2　基因组简单回归模型尺度转换后的 GWAS 结果

半滑舌鳎性逆转性状使用基因组简单线性回归模型尺度转换后的 GWAS 结果如图 2 所示, 此方法把简单回归结果中显著位点的效应值和遗传力的尺度转化为可解释的广义线性回归结果。可以看出, 1、5、30 个 PC 的 Q-Q 图前面部分都能较好地贴合理论线, 尾部上扬, 说明 3 种情况都能对群体结构较好校正。计算 GC 值能直接了解基因组控制效果, 1 个 PC 的 GC 值为 1.1, 5 个 PC 的 GC 值为 1.1, 30 个 PC 的 GC 值为 1.01, 可以看出考虑 30 个 PC 的简单回归尺度转换方法对于群体分层的校正效果最好。由曼哈顿图可看出, 检测到 6 个 QTN, 5 个在 Z 染色体上, 1 个在 W 染色体上; 同时还检测到 13 号染色体的 1 个接近阈值的位点, 以及 Z 染色体上 3 个接近阈值的位点。

半滑舌鳎性逆转性状考虑 30 个 PC, 使用简单回归模型进行尺度转换, 然后进行 GWAS, 检出的 QTN 位点的详细信息如表 1 所示。表中 QTN

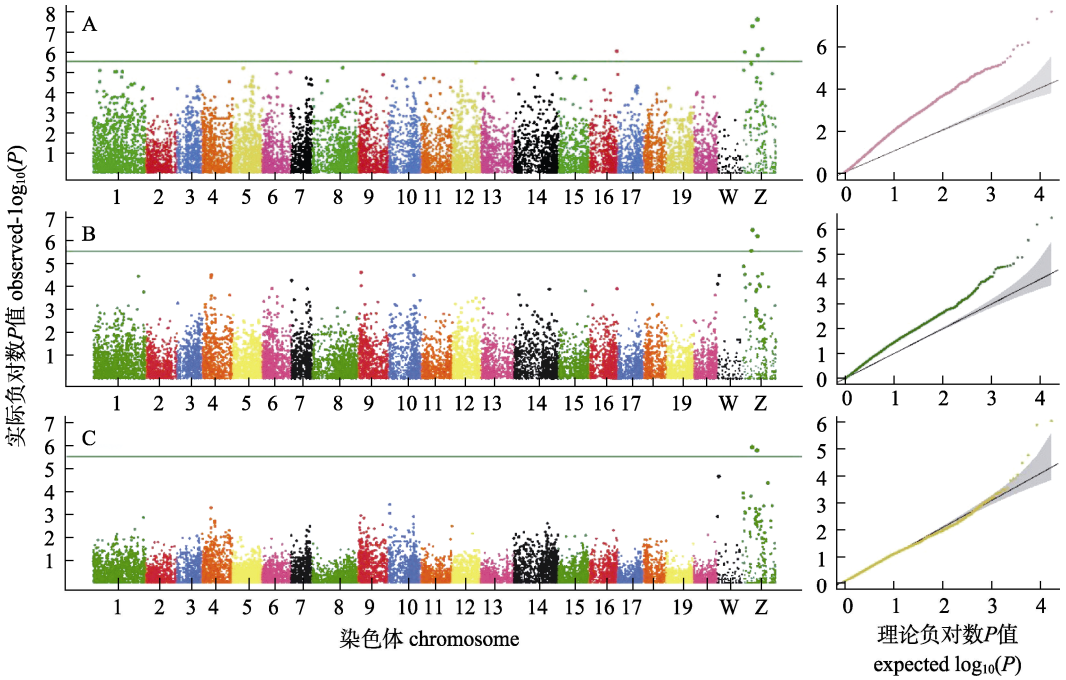

图 1 半滑舌鳎性逆转性状广义线性回归模型 GWAS 曼哈顿和 Q-Q 图

A. 1 个 PC, B. 3 个 PC, C. 5 个 PC

Fig. 1 Manhattan and QQ plots of generalized linear regression model GWAS on sex reversal traits in Half-smooth Tongue sole

A. 1 PC, B. 3 PCs, C. 5 PCs

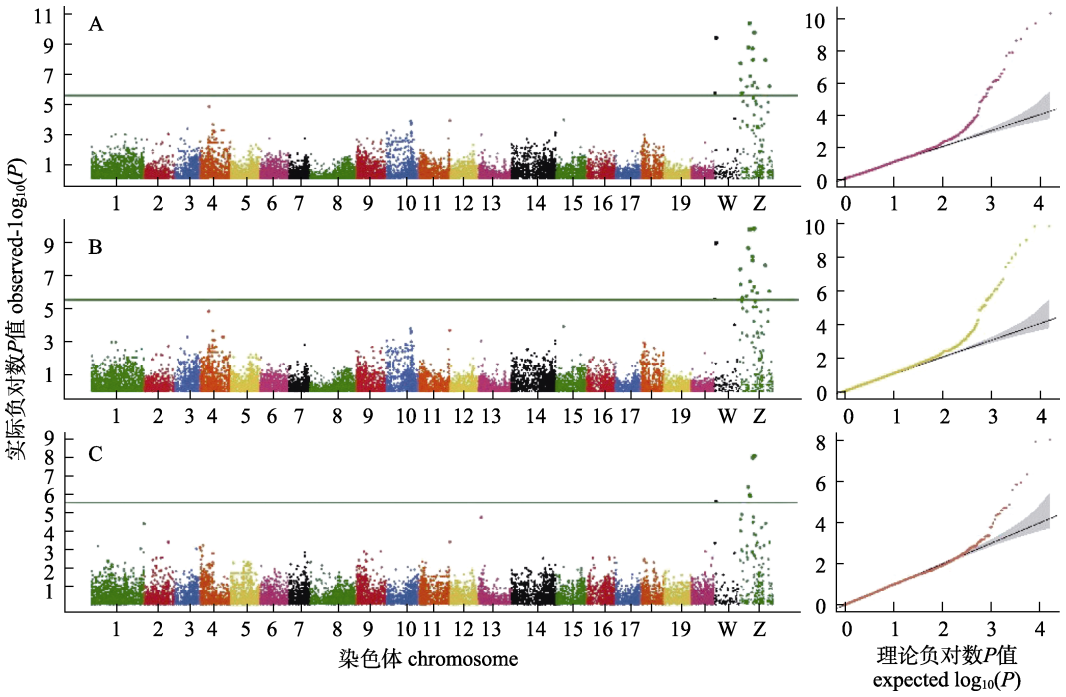

图 2 半滑舌鳎性逆转性状简单回归模型尺度转换 GWAS 曼哈顿和 Q-Q 图

A. 1 个 PC, B. 5 个 PC, C. 30 个 PC.

Fig. 2 Manhattan and QQ plots for scale transformation of simple regression model GWAS on sex reversal traits in Half-smooth Tongue sole

A. 1 PC, B. 5 PCs, C. 30 PCs.

遗传力说明各个位点能解释性逆转性状遗传变异的百分比，6 个 *P* 值超出阈值的 SNP 位点即为检测到的与半滑舌鳎性逆转性状相关的 QTN，其余 4 个接近阈值的位点，可作为候选 QTN。

表 1　检测到的与半滑舌鳎性逆转性状相关联的 QTN 信息

Tab. 1　QTN information associated with sex reversal trait detected in *Cynoglossus semilaevis*

染色体 chromosome	位置 location	SNP 编号 SNP ID	基因名称 gene name	Lm 效应 Lm effect	Lm 标准误 Lm standard error	Glm 效应 Glm effect	Glm 标准误 Glm standard error	QTN 遗传力/% QTN heritability	*P*
Z	9793913	M65579	Si:deky-193c22.1	−0.3032	0.0466	−1.2138	0.1864	0.189	8.24E-09*
Z	8643646	M65437	DAPK1	−0.4751	0.0735	−1.9015	0.2942	0.198	1.01E-08*
Z	5833328	M65164	ADGRD2	−0.2998	0.0538	−1.1999	0.2155	0.240	4.06E-07*
Z	6676874	M65244	FBXL17	−0.2540	0.0477	−1.0168	0.1910	0.236	1.08E-06*
Z	7256721	M65301	DMXL1	0.4827	0.0915	1.9322	0.3663	0.275	1.31E-06*
W	1218194	M63786	LOC103396896	0.5462	0.1067	2.1863	0.4270	0.275	2.40E-06*
Z	1633247	M64758	LOC103397473	−0.4116	0.0879	−1.6475	0.3516	0.270	1.26E-05
Z	9232660	M65511	无	0.3734	0.0811	1.4946	0.3245	0.275	1.70E-05
13	2140529	M40452	CD27	−0.9204	0.2006	−3.6840	0.8029	0.227	1.82E-05
Z	781540	M64676	MALT1	−0.2973	0.0657	−1.1902	0.2631	0.252	2.31E-05

注: *表示该 SNP 位点与性逆转性状显著相关($P<2.838\times10^{-6}$).

Note: * represent that the SNP location is significantly correlate with sex reversal trait ($P<2.838\times10^{-6}$).

3　讨论

本研究结果表明, 简单回归尺度转换方法和广义线性回归模型都可以很好地校正关联分析中的群体分层, 无论是评价指标中的 Q-Q 图还是基因组控制值, 都显示只要考虑作为协变量的主成分数目恰当, 就可以很好地控制全基因组关联分析中的假阳性。比较两种方法检出 QTN 的个数可知, 基于简单回归尺度转换的简单回归模型比直接考虑 PC 的广义线性回归模型的 QTN 的检出效力更高, 且使用简单回归尺度转换方法检测到的 QTN 高于已发表的方法, 不但检测到了 Z 染色体上的 QTN, 还检测到了 W 染色体的一个 QTN。

Z 染色体上位置为 9793913 的 SNP 位于 *si:deky-193c22.1* 基因上, 未有研究者对该基因的功能进行探究。Z 染色体上位置为 8643646 的 SNP 在 *DAPK1* (death-associated protein kinase 1)基因上, *DAPK1* 与细胞的死亡有关, Raval 等[19]研究发现, *DAPK1* 的表达缺失或降低是人类慢性 B 淋巴细胞白血病发病的原因; Siqueira 等[20]在不同性别牛胚胎对群落刺激因子 2 反应的研究中证实, *DAPK1* 的表达有受性别影响的趋势。Z 染色体上位置为 5833328 的 SNP 在 *ADGRD2* (adhesion G protein-coupled receptor D2)的基因上。Z 染色体上位置为 6676874 的 SNP 在 *FBXL17* (F-box and leucine-rich repeat protein 17)的基因上, Jiang 等[16]有关半滑舌鳎性逆转性状的研究发现, *FBXL17* 基因上位置为 6676874 的 SNP 与半滑舌鳎性逆转性状有关, 本研究所应用的方法也检测到了这一位点。Z 染色体上位置为 7256721 的 SNP, 在 *DMXL1* (Dmx like 1)基因上, 人类 *DMXL1* 基因编码一个 11 kb 的 RNA, 开放阅读框 3027 个氨基酸, 主要具有调节功能[21]。W 染色体上位置为 1218194 的 SNP, 是唯一检测到的 W 染色体上的 QTN, 在 *LOC103396896* 基因上, 还没有研究者对该基因功能进行深入的研究, 但是这个基因位于 W 染色体上, 有可能参与到性逆转性状的调控中。

除了检出的 5 个在 Z 染色体上的 QTN 和 1 个在 W 染色体上的 QTN 之外, 还检测到 Z 染色体上的另外 3 个 QTN, 但未超过阈值, 13 号染色体上也检测出了一个接近阈值的 QTN, 这几个位点如下: Z 染色体上位置为 1633247 的 SNP, 在 *LOC103397473* 基因上; Z 染色体上位置为 9232660 的 SNP, 未在基因编码区内; Z 染色体上位置为 781540 的 SNP, 在 *malt1* (MALT paracaspase 1)基因上, MALT1 是在抗原受体介导的淋巴细胞活化中起关键作用的信号蛋白[22]; 13 号染色体上检出的位点位置是 2140529, 在 *CD27* 基因上, CD27 是 T 细胞的共刺激分子, 可支持 naïve T 细胞的抗原特异性扩增, 若删除或改变编码 CD27 的基因, 免疫应答会发生延迟, 特异性 T 细胞数量减少[23]。这些位点虽未超过阈值, 但可作为候

选 QTN, 为后续有关半滑舌鳎性逆转性状以及其他性状的全基因组关联分析研究奠定基础。

运用基于尺度转换的简单回归模型所检测出的与半滑舌鳎性逆转性状相关的位点, 除了 Z 染色体上位置为 6676874 的 SNP 与半滑舌鳎性逆转性状有关之外[16], 其余位点到目前为止没有报告或有明显的证据表明与性别决定的各种途径直接相关, 但它们仍值得被关注, 因为这些位点基本都在 Z 染色体或者 W 染色体上。众所周知, Z 染色体和 W 染色体是半滑舌鳎的性染色体, 雌性为 ZW, 雄性为 ZZ, 这两条染色体上的基因以及他们的连锁基因对于性别决定都可能具有潜在的影响, 任一位置的突变都有可能对决定性别的基因产生影响。另外 4 个接近阈值的位点可作为候选 QTN, 为研究者们研究半滑舌鳎性逆转性状提供了候选研究位点。

参考文献:

[1] Zhao J L, Li S L, Gao J, et al. Bare-bones regression scan for genome-wide mixed model association study[J]. Journal of Northeast Agricultural University, 2018, 49(7): 58-66. [赵敬丽, 李淑玲, 高进, 等. 全基因组混合模型关联分析的极速回归扫描法研究[J]. 东北农业大学学报, 2018, 49(7): 58-66.]

[2] Slatkin M. Inbreeding coefficients and coalescence times[J]. Genetical Research, 1991, 58(2): 167-175.

[3] Zhu X F, Zhang S L, Zhao H Y, et al. Association mapping, using a mixture model for complex traits[J]. Genetic Epidemiology, 2002, 23(2): 181-196.

[4] Price A L, Patterson N J, Plenge R M, et al. Principal components analysis corrects for stratification in genome-wide association studies[J]. Nature Genetics, 2006, 38(8): 904-909.

[5] Bulmer M G. The effect of selection on genetic variability[J]. The American Naturalist, 1971, 105(943): 201-211.

[6] Falconer D S, Mackay T, Longman P. Introduction to quantitative genetics[J]. American Journal of Human Genetics, 1990, 46(6): 1231.

[7] Schall R. Estimation in generalized linear models with random effects[J]. Biometrika, 1991, 78(4): 719-727.

[8] Gilmour A R, Anderson R D, Rae A L. The analysis of binomial data by a generalized linear mixed model[J]. Biometrika, 1985, 72(3): 593-599.

[9] Song C, Jiang L, Wang J W, et al. Studies on genetic features of sex reversal in *Cynoglossus semilaevis*[J]. Biotechnology Bulletin, 2015, 31(3): 207-212. [宋超, 蒋丽, 王景伟, 等. 半滑舌鳎性逆转的遗传特性研究[J]. 生物技术通报, 2015, 31(3): 207-212.]

[10] Wan R J, Jiang Y W, Zhuang Z M. Morphological and developmental characters at the early stages of the tonguefish *Cynoglossus semilaevis*[J]. Acta Zoologica Sinica, 2004, 50(1): 91-102. [万瑞景, 姜言伟, 庄志猛. 半滑舌鳎早期形态及发育特征[J]. 动物学报, 2004, 50(1): 91-102.]

[11] Liang Z, Chen S L, Zhang J, et al. Gonadal development process observation of half-smooth tongue sole in rearing population[J]. Journal of Southern Agriculture, 2012, 43(12): 2074-2078. [梁卓, 陈松林, 张静, 等. 半滑舌鳎养殖群体性腺发育观察[J]. 南方农业学报, 2012, 43(12): 2074-2078.]

[12] Narita S, Kageyama D, Nomura M, et al. Unexpected mechanism of symbiont-induced reversal of insect sex: Feminizing *Wolbachia* continuously acts on the butterfly *Eurema hecabe* during larval development[J]. Applied and Environmental Microbiology, 2007, 73(13): 4332-4341.

[13] Quinn A E, Georges A, Sarre S D, et al. Temperature sex reversal implies sex gene dosage in a reptile[J]. Science, 2007, 316(5823): 411.

[14] Nagahama Y. Molecular mechanisms of sex determination and gonadal sex differentiation in fish[J]. Fish Physiology and Biochemistry, 2005, 31(2-3): 105-109.

[15] Shao C W, Li Q Y, Chen S L, et al. Epigenetic modification and inheritance in sexual reversal of fish[J]. Genome Research, 2014, 24(4): 604-615.

[16] Jiang L, Li H D. Single locus maintains large variation of sex reversal in half-smooth tongue sole (*Cynoglossus semilaevis*)[J]. G3 Genes|Genomes|Genetics, 2017, 7(2): 583- 589.

[17] Chen S L, Tian Y S, Yang J F, et al. Artificial gynogenesis and sex determination in half-smooth tongue sole (*Cynoglossus semilaevis*)[J]. Marine Biotechnology, 2009, 11(2): 243-251.

[18] Yang J, Zaitlen N A, Goddard M E, et al. Advantages and pitfalls in the application of mixed-model association methods[J]. Nature Genetics, 2014, 46(2): 100-106.

[19] Raval A, Tanner S M, Byrd J C, et al. Downregulation of *death-associated protein kinase 1* (*DAPK1*) in chronic lymphocytic leukemia[J]. Cell, 2007, 129(5): 879-890.

[20] Siqueira L G B, Hansen P J. Sex differences in response of the bovine embryo to colony-stimulating factor 2[J]. Reproduction, 2016, 152(6): 645-654.

[21] Kraemer C, Enklaar T, Zabel B, et al. Mapping and structure of DMXL1, a human homologue of the DmX gene from *Drosophila melanogaster* coding for a WD repeat protein[J].

Genomics, 2000, 64(1): 97-101.

[22] Thome M. CARMA1, BCL-10 and MALT1 in lymphocyte development and activation[J]. Nature Reviews Immunology, 2004, 4(5): 348-359.

[23] Hendriks J, Gravestein L A, Tesselaar K, et al. CD27 is required for generation and long-term maintenance of T cell immunity[J]. Nature Immunology, 2000, 1(5): 433-440.

Efficiently mapping the sex reversal genes of half-smooth tongue sole, *Cynoglossus semilaevis* using simple regression scale transformation

HUANG Yan[1, 3], SONG Yuxin[2, 3], JIANG Li[3], YANG Runqing[3]

1. National Demonstration Center for Experimental Fisheries Science Education, Shanghai Ocean University, Shanghai 201306, China;
2. Wuxi Fisheries College, Nanjing Agricultural University, Wuxi 214081, China;
3. Research Centre for Aquatic Biotechnology, Chinese Academy of Fishery Sciences, Beijing 100141, China

Abstract: In genome-wide association analysis of discontinuous traits, when complex population stratification exists in genomic data, the generalized linear model needs to consider hundreds of covariables at the same time, which slows the calculation speed and presents abnormal solutions. This study aimed to transform the effect value and heritability scale of significant loci in simple linear regression results into interpretable generalized linear regression results. First, the eigenvectors solved by spectral decomposition of the kinship matrix were considered as the principal components (PCs) to correct the population stratification in the discontinuous traits dataset. Then, a new covariate was formed through the sum of the multiplications of each covariate, and its regression coefficient of the principal component was computed using a linear regression model. The new covariate was used as the covariable of simple regression to carry out correlation tests for markers one by one. Finally, the generalized linear model was used for regression analysis of candidate quantitative trait nucleotides (QTNs), and the effects and variance were transformed into the generalized linear regression model scale. The genome-wide association analysis of sex reversal traits in half-smooth tongue sole (*Cynoglossus semilaevis*) was conducted using the new method and the generalized linear regression model with direct consideration of principal components: The results show that the QTN detection efficiency of this method is higher, a total of 6 QTNs were detected, including 5 QTNs on Z chromosome and 1 QTN on W chromosome. In addition, in terms of genome control, the genome control value of the method in this study is the same as that of the generalized linear regression model which directly considers PC, which is at an optimal level of 1.01. Therefore, the simple regression scaling transformation method based on principal component analysis improved the detection power for QTN detection, while retaining the accuracy of results, with fast and robust genome-wide association analysis of discontinuous traits. In addition, the QTNs detected by the new method proposed in this study can provide theoretical guidance for the study of sex reversal traits in half-smooth tongue soles.

Key words: genome-wide association analysis; *Cynoglossus semilaevis*; sex reversal; principal component; scaling transformation; simple regression model; generalized linear regression model

Corresponding author: JIANG Li, E-mail: jiangl@cafs.ac.cn; YANG Runqing, E-mail: runqingyang@cafs.ac.cn