



整合不同生物学先验信息的基因组选择预测准确性的比较研究

谭涛涛, 高炳熙, 杜志强*

(长江大学动物科学技术学院, 荆州 434025)

摘要: 旨在开发整合基因组信息、表型信息和功能注释信息筛选遗传位点的不同方法, 比较评估8种在复杂性状上的基因组选择(genomic selection, GS)方法的预测准确性。本研究将542只北京鸭和2 548只杜洛克猪基因型-表型数据通过3个维度, 基于基因组信息(连锁不平衡结合主成分分析(linkage disequilibrium and principal component analysis, LD_PCA), 结合其他主成分位点(LD and other principal component analysis, LD_outPCA), 结合皮尔逊相关系数分析(LD and Pearson correlation coefficient, LD_PCC))、表型信息(连锁不平衡结合全基因组关联分析(LD and genome-wide association study, LD_GWAS), 结合互作分析(LD and EPISNP, LD_EPI))和功能注释信息(权重分配法(weight distribution method, WDM)), 共8种不同筛选方法处理构建基因型矩阵, 比较不同方法的基因组育种值估计准确性。结果表明, 比较发现LD_PCC结合6种GS方法(GBLUP、BayesA、BayesB、BayesC、BayesLASSO和RRGBLUP), 在鸭龙骨长、猪百千克日龄、猪背膘厚和乳头数表型获得最高平均预测准确性(对比初始基因型矩阵提高约4%~13.9%)分别达到了0.709 3、0.440 0、0.497 4和0.440 1; 权重分配法(WDM)在猪背膘厚和乳头数表型对比未加权获得最高预测准确性(高约10%)。综上, 本研究发现WDM和LD_PCC可以有效提升GS的预测准确性, 为深入研究位点筛选方法影响GS准确性和在育种实践中的应用提供了方向和借鉴。

关键词: 基因组选择; 基因组关系矩阵; 位点筛选; 功能注释; 权重分配

中图分类号: S813.1

文献标志码: A

文章编号: 0366-6964(2026)03-1373-12

Comparative Study on the Prediction Accuracy of Genomic Selection Integrating Different Biological Priors

TAN Taotao, GAO Bingxi, DU Zhiqiang*

(College of Animal Science and Technology, Yangtze University, Jingzhou 434025, China)

Abstract: The aim of this study was to develop different methods for screening genetic variants by integrating genomic data, phenotypic information, and functional annotations, and to compare and evaluate the prediction accuracy of 8 genomic selection (GS) methods for complex traits. Genotype-phenotype data from 542 Beijing ducks and 2 548 Duroc pigs were analyzed using three dimensions of variant selection: 1) genomic information, including linkage disequilibrium combined with principal component analysis (LD_PCA), other principal component variants (LD_outPCA), and Pearson correlation coefficient analysis (LD_PCC); 2) phenotypic information, including LD combined with genome-wide association study (LD_GWAS) and EPISNP analysis (LD_EPI); and 3) functional annotation information, using the weight distribution method (WDM). In total, 8 different screening

收稿日期: 2025-08-18

基金项目: 湖北省技术创新计划项目(2024BBB014); 安徽省畜禽联合育种改良项目(2021-2025)

作者简介: 谭涛涛, 硕士生, 主要从事分子数量遗传学和生物信息学的研究, E-mail: Tantt924@163.com

* 通信作者: 杜志强, 主要从事动物遗传育种与繁殖研究, E-mail: zhqdu@yangtzeu.edu.cn

methods were applied to construct genotype matrices, and the prediction accuracies of different GS models were compared. The LD_PCC method, when combined with 6 GS models (GBLUP, BayesA, BayesB, BayesC, Bayes LASSO, and RRGBLUP), achieved the highest mean prediction accuracies for duck fossil length, pig age to 100 kg, backfat thickness, and teat number, improving by approximately 4%-13.9% compared with the initial genotype matrix, with accuracies of 0.709 3, 0.440 0, 0.497 4, and 0.440 1, respectively. The WDM method achieved the highest accuracies for backfat thickness and teat number, with improvements of about 10% compared with unweighted models. In conclusion, this study demonstrates that WDM and LD_PCC can effectively enhance the prediction accuracy of GS, providing valuable insights for further exploration of variant selection methods and their application in breeding practice.

Keywords: genome selection; genome relationship matrix; site screening; functional annotation; weight assignment

* **Corresponding author:** DU Zhiqiang, E-mail: zhqdu@yangtzeu.edu.cn

全基因组选择 (genomic selection, GS) 是由 Meuwissen 等^[1]提出的一种基于全基因组标记信息估算个体基因组育种值 (genomic estimated breeding value, GEBV) 的遗传评估方法^[2,3]。其核心在于利用高密度单核苷酸多态性 (single nucleotide polymorphism, SNP) 基因型和表型性状数据构建线性预测模型, 以提高遗传进展和选育效率^[4-6]。GS 预测准确性受训练群体的规模和遗传结构、SNP 标记的密度和质量、复杂性状的遗传基础等多种因素影响^[7-9]。

祁军英等^[10]通过 GWAS 筛选藏羊毛性状相关位点提高预测准确性; 贾宏霞等^[11]在肉牛的研究进展提到多种方法, 不同的 SNP 筛选策略会影响基因组关系矩阵 (genomic relationship matrix, GRM) 的构建和个体间亲缘关系评估的准确性, 从而影响 GEBV 的估算^[12]; 李棉燕等^[13]指出特征筛选 (feature selection) 对机器学习模型的表现至关重要, 但是已有研究中 SNP 筛选方法单一, 包括连锁不平衡 (LD) 筛选、主成分分析 (PCA) 筛选等, 容易出现如 LD 滑移导致的标记丢失、PCA 可能丢失关键遗传信息等问题^[14,15], 在构建基因组预测模型时, 对所有 SNPs 采用统一的权重处理方式也会导致信息单一或丢失个别关键 SNPs 的高影响信号^[16]。如何结合多种方法筛选 SNP, 以提高 GS 准确性, 仍是亟待解决的问题^[17]。因此, 本研究提出结合 LD、PCC、PCA、GWAS、EPISNP 及权重分配优化 SNP 筛选的方法^[18-21], 结合基因组最佳线性无偏预测 (genomic best linear unbiased prediction, GBLUP) 等模型在北京鸭和杜洛克猪数据集中进行验证, 采用

重复 5 次的五折交叉验证获取准确性。比较多种方法的预测准确性, 为在北京鸭龙骨长和胸宽性状以及杜洛克猪的百千克日龄、背膘厚和乳头数性状选育中使用基因组选择提供理论依据。

1 材料与方法

1.1 数据来源

本研究选择了两批数据集, 涵盖 5 个表型, 分别来自北京金鸭有限公司提供的北京鸭数据和广东温氏食品集团 (中国广东) 提供的一个核心农场的杜洛克猪数据^[22,23]。北京鸭数据包含 542 个样本, 涉及胸宽 (breast width) 和龙骨长 (fossil length) 两个表型; 杜洛克猪数据包含 2 548 个样本, 涉及背膘厚 (backfat)、乳头数 (teat number) 和百千克日龄 (age to 100 kg) 3 个表型。对表型数据的分析结果见表 1。

使用 Beagle (5.2) 软件对缺失的基因型进行填充, 使用 PLINK (1.90b6.21) 软件^[24]对数据进行初步质控。mind 和 geno 参数设定为 0.02 过滤掉基因型缺失率大于 2% 的个体和 SNP 位点, maf 参数为 0.05 过滤掉最小等位基因频率 (minor allele frequency, MAF) 小于 0.05 的位点, hwe 参数设定为 1×10^{-4} 过滤掉 Hardy-Weinberg 平衡试验 ($P < 1 \times 10^{-4}$) 的 SNPs。最终个体样本数完整, 北京鸭剩余约 3.4 万个位点, 杜洛克猪剩余约 991.8 万个位点。

1.2 基于基因组信息的位点筛选

1.2.1 LD 通过使用 PLINK 工具计算 SNP 位点之间的 R^2 值来衡量 LD^[25]。设定 LD 阈值为 0.8, 若 R^2 值大于该阈值, 认为它们具有较强的连锁, 删

表1 表型统计结果

Table 1 Descriptive statistical analyses on traits

表型 Phenotype	样本 Sample	平均值 Mean	标准差 Standard deviation	最大值 Maxi value	最小值 Min value	遗传力 Heritability
鸭胸宽/cm Duck breast width	542	11.75	0.48	13.10	10.30	0.34
龙骨长/cm Duck fossil lenth	542	14.60	0.64	19.60	12.40	0.48
百千克日龄/day Age to 100 kg	2 548	154.66	9.32	203.01	124.47	0.26
背膘厚/mm Backfat thickness	2 548	10.91	2.06	19.70	5.48	0.32
乳头数/个 Teat number	2 548	10.72	1.06	15.00	8.00	0.27

除其中一个位点以减少冗余,具体过程见图1A。

1.2.2 LD_PCA LD_PCA主要通过机器学习库scikit-learn来实现^[26],通过线性变换将LD得到的SNP矩阵映射到一个新的坐标系,运算过程是将SNP矩阵进行归一化得到矩阵归一化矩阵X。

$$C = \frac{X^T X}{n-1}$$

通过上面协方差计算公式计算得到协方差矩阵。其中n表示样本数,通过协方差矩阵可以得到特征值和特征向量,进而得到主成分,如图1D所示。

1.2.3 LD_outPCA方法 LD_outPCA是指将连锁不平衡剔除位点PCA处理后的数据与剩余位点结合,保留LD剔除位点的主要信息和变异性,以最大程度地减少重要遗传信息的损失,见图1E。

1.2.4 LD_PCC方法 LD_PCC皮尔逊相关系数(Pearson correlation coefficient, PCC)^[27]是一种用于衡量两个变量之间线性关系强度和方向的统计方法。PCC的取值范围为-1~1,其具体含义如下:当PCC等于1时,表示两个变量之间存在完全正相关关系;当PCC等于-1时,表示两个变量之间存在完全负相关关系;当PCC等于0时,表示两个变量之间不存在线性相关性,表明它们彼此独立或不相关:

$$PCC(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

其中, X_i 、 Y_i 是第*i*个元素的观察值, $\bar{X}$ 、 $\bar{Y}$ 分别是X、Y的均值。

1.3 基于表型的先验标记

1.3.1 LD_GWAS方法 选择性连锁不平衡(SLDP)是Zhu等^[23]提出的一种SNP位点先验标记方法,对预测准确性有良好的提升,是指将GWAS计算得到的不同显著程度的位点标记为重要位点,

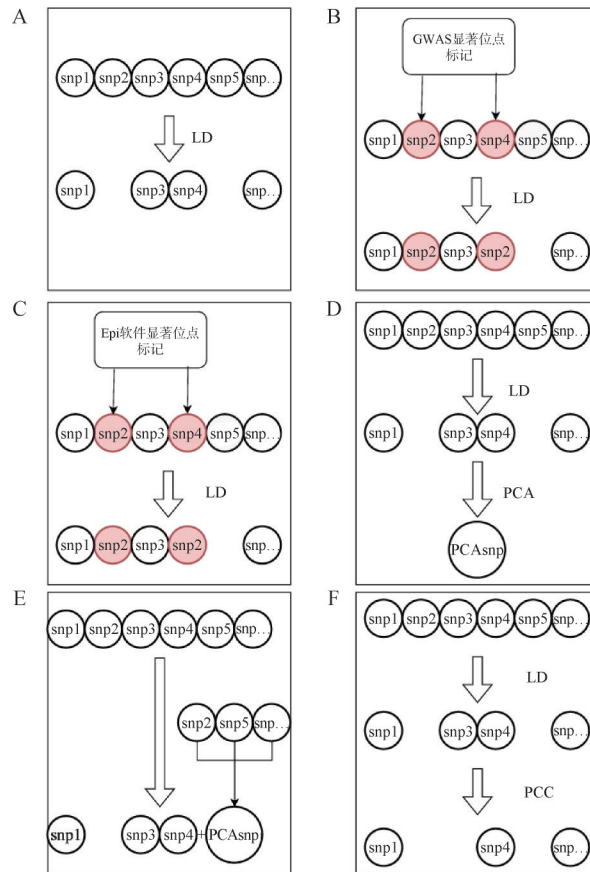
将其进行LD连锁不平衡剪枝时保留重要位点,同时减少干扰,从而提高准确性,过程见图1B。

1.3.2 LD_EPI方法 EPISNP是Ma等^[28]开发的一款软件,通过广义加性模型计算出不同位点间效应,包括加性、显性、互作等效应。标记EPISNP计算出的前1 000个效应显著的位点,再进行LD剪枝过程,见图1C。

1.4 基于功能注释信息的权重分配法

权重分配法是指将猪背膘和乳头数数据质控和LD筛选后,使用ANNOVAR软件对SNP位点进行注释。依据参考基因组数据库的功能注释对SNP进行分类并赋予不同的遗传权重。本研究使用了*Sus scrofa*11参考基因组数据^[29]以确定每个SNP位点在编码区、调控区、内含子区和其他功能区域的位置,并评估SNP对基因功能的潜在影响,SNP位点注释后,基于生物学先验信息,根据每个SNP所处的基因区域分配不同的遗传权重^[30-32]。

外显子区(exonic)的SNP因直接影响蛋白质的编码和功能,被赋予最高遗传权重1.0。剪接区域(splicing)的SNP变异可能会影响mRNA的剪接过程,从而改变蛋白质的结构和功能,遗传权重为0.9。位于同时作用于外显子和剪接区域(exonic; splicing)的SNP同样被赋予遗传权重1.0。非翻译区(UTR)的SNP在基因表达调控中具有重要作用,其中UTR3和UTR5的遗传权重分别为0.5和0.6。非编码RNA外显子区(ncRNA exonic)的SNP可能影响非编码RNA的功能,遗传权重为0.4。内含子区(intronic)的SNP尽管不直接编码蛋白质,但其变异可能会影响基因的表达调控,因此遗传权重为0.2。上游(upstream)和下游(downstream)区域的SNP变异可能会影响基因的启动和终止过程,遗传权重分别为0.4和0.3;同时位于上游和下游区域的SNP则赋予0.35的遗传权重。基因间区



LD. 连锁不平衡筛选;GWAS. 全基因组关联分析;Epi. 效应位点计算;PCA. 主成分分析;PCC. 皮尔逊相关系数计算。A. LD方法;B. LD_GWAS方法;C. LD_EPI方法;D. LD_PCA方法;E. LD_outPCA方法;F. LD_PCC方法

LD. Linkage disequilibrium screening; GWAS. Genome-wide association study; Epi. Effect locus calculation; PCA. Principal component analysis; PCC. Pearson correlation coefficient calculation. A. LD method; B. LD_GWAS method; C. LD_EPI method; D. LD_PCA method; E. LD_outPCA method; F. LD_PCC method

图1 多重特征筛选流程

Fig. 1 Multiple screening processes

(intergenic)的SNP变异对基因功能的影响较小,遗传权重为0.1,加权后生成新的SNP矩阵。通过这种权重分配法,在基因组选择过程中能够突出重要位点并降低噪音干扰,更精准地评估SNP效应,从而优化基因组选择模型的性能,提高表型预测的准确性。

1.5 基因组评估方法

本研究主要使用的基因组评估方法包括GBLUP^[33]、贝叶斯方法(如BayesA、BayesB、BayesC、Bayes LASSO)和RRGBLUP等^[34-36]。GBLUP是将基因组数据与性状数据进行线性组合,以估计个体的遗传值:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

其中, $\mathbf{y}$ 为猪的百千克日龄(age to 100 kg)、背膘厚(backfat thickness)和乳头数量(teat number)以

及北京鸭的胸宽(breast width)和龙骨长(fossil length)等性状向量;固定效应矩阵 $\mathbf{X}$ 仅包含截距项,由于缺乏其他固定效应信息,因此设为全为1的列向量; $\mathbf{b}$ 为固定效应系数; $\mathbf{Z}$ 为SNP位点矩阵; $\mathbf{u}$ 指基因效应的随机效应; $\mathbf{e}$ 为误差项。基因效应 $\mathbf{u}$ 被视为随机效应,其分布通常假定为多变量正态分布,并假设是所有SNP位点的效应相等,与GBLUP不同,BayesA使用贝叶斯框架,允许不同位点的效应不同,因此具有更大的灵活性。BayesB引入了稀疏性先验,鼓励某些位点的效应为零:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \sum_{i=1}^N \mathbf{Z}_i \alpha_i + \mathbf{e}$$

其中, $\mathbf{Z}_i, \alpha_i$ 分别表示每个SNP位点的设计矩阵和每个SNP位点的效应,允许BayesB进行位点选择,即确定哪些位点对性状具有显著影响。与

BayesB 相似,但 BayesC 的效应参数 α_i 包括了一个均匀分布,这进一步强调了稀疏性,更强烈地鼓励某些位点的效应为零。Bayesian Lasso 则引入了 L1 正则化项:

$$P(\beta|\lambda) = \prod_{j=1}^p \frac{1}{2\lambda} e^{-\frac{|\beta_j|}{\lambda}}$$

β 和 β_j 分别表示特征系数和回归系数, p 指的是回归系数的总数, λ 是控制稀疏程度的超参数,较小的使 β_j 更倾向于 0 (更稀疏),较大的 λ 允许较大的 β_j 。目的是鼓励参数向量 β 中的某些系数为零,从而实现稀疏性,通过后验分布来估计参数向量 β 的后验概率,进而确定特征对性状的影响程度,通常使用马尔科夫链蒙特卡罗 (MCMC) 采样等来实现。

RRGBLUP 是一种基于再生核希尔伯特空间 (RKHS) 的方法,结合岭回归 (ridge regression) 和 BLUP (best linear unbiased prediction) 进行基因组选择,它适用于捕捉复杂的非线性基因效应。为了确保结果的准确性和可比性,每个评估方法均进行了五折交叉运算,使得估计更加稳定。

2 结果

2.1 基于表型的先验标记和基因组信息的筛选结果

为了更直观地呈现各点位筛选方法处理后的剩余位点数,统计了所有方法处理 SNP 位点后剩余的位点数,统计结果见表 2、表 3、表 4。

表 2 位点筛选结果一

Table 2 Site screening results one

数据集 Dataset	初始标记 Origin markers	LD	GWAS<0.04	GWAS<0.03	GWAS<0.02	GWAS<0.01
鸭胸宽 Duck breast width	2 475	2 209	2 450	2 411	2 360	2 294
龙骨长 Duck fossil lenth	2 255	2 068	2 231	2 202	2 168	2 131

GWAS<. LD 数据结合 GWAS 阈值标记的位点。下表同

GWAS<. Indicates sites where LD data are falsely combined with GWAS threshold markers. The same as below

表 3 位点筛选结果二

Table 3 Site screening results two

数据集 Dataset	初始标记 Origin markers	LD	GWAS<1e-6	GWAS<1e-8	GWAS<1e-10	GWAS<1e-12
百千克日龄 Age to 100 kg	332 870	7 158	78 789	33 765	22 334	7 698
背膘厚 Backfat thickness	412 631	8 246	102 466	20 834	9 826	8 439
乳头数 Teat number	478 646	12 219	173 868	69 862	28 142	17 260

表 4 位点筛选结果三

Table 4 Site screening results three

数据集 Dataset	LD_PCC	LD_outPCA	LD_PCA	LD_EPI
鸭胸宽 Duck breast width	1 933	2 310	291	2 335
龙骨长 Duck fossil lenth	1 812	2 170	287	2 150
百千克日龄 Age to 100 kg	1 111	7 256	120	1 595
背膘厚 Backfat thickness	1 056	8 342	110	2 972
乳头数 Teat number	1 393	12 338	128	4 233

2.2 预测精确度对比

本研究选取了皮尔逊相关系数 (PCC) 以及均方误差 (mean squared error, MSE) 作为衡量预测精度的手段。MSE 是一种常用的统计度量,用于衡量一个估计值或模型预测误差的平均平方值。MSE

通常用于评估模型的精度,特别是在回归分析和预测建模中:

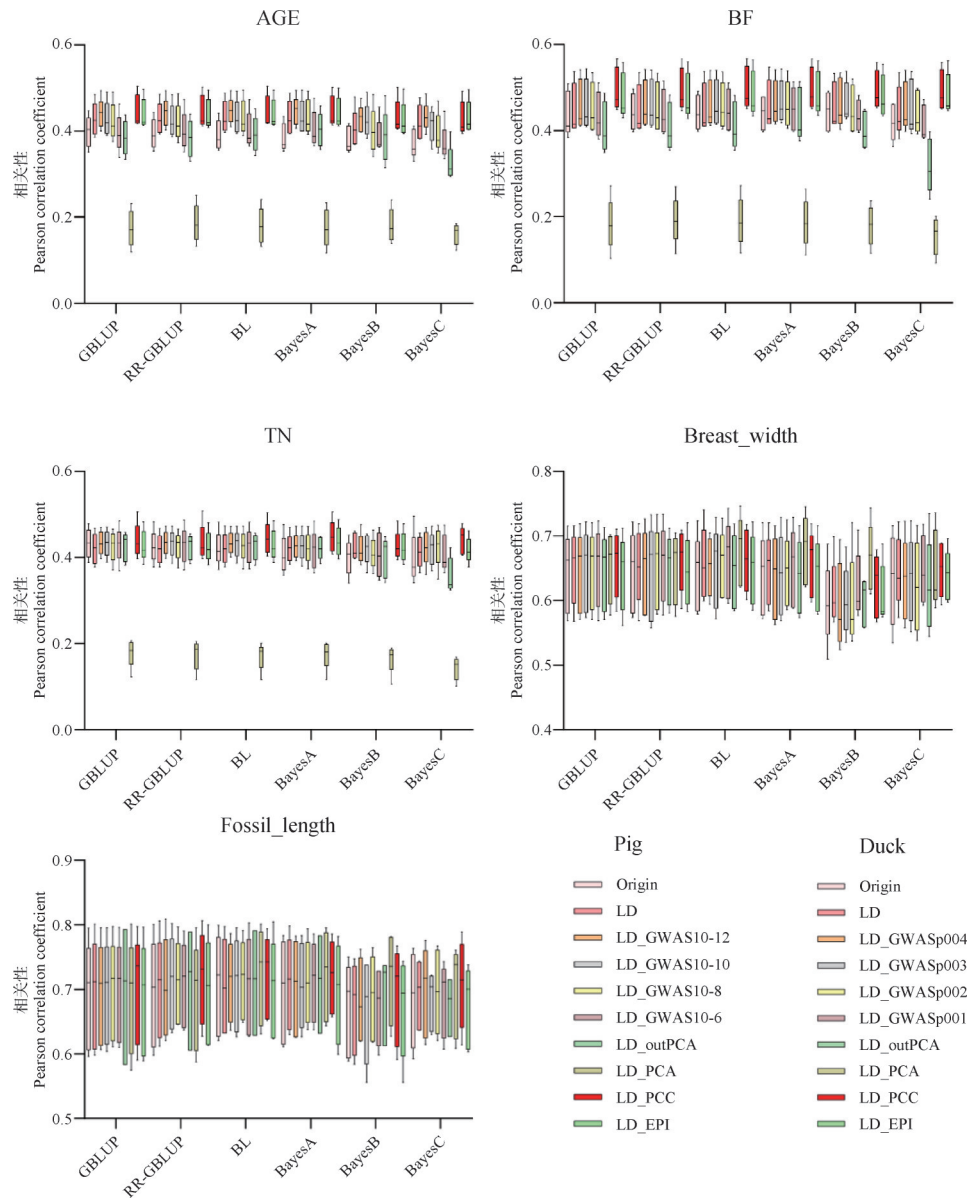
$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

对于每个观测值 i , 计算真实值 y_i 与预测值 $\hat{y}_i$ 之

间差的平方,最后计算平均值。MSE 的值越小越好,表示模型的预测与实际观测值之间的差异越小,模型越准确。取平均值并计算标准差得到结果见图 2。在北京鸭胸宽和龙骨长数据中,不同方法预测的相关性差别不大,其中 LD_PCC 融合处理的 GBLUP 方法相关性均最高,分别达到 0.710 0 和 0.791 6,MSE 分别为 0.122 5 和 0.220 6,处于较高

水平;猪的百千克日龄性状最好值同样是由 LD_PCC 的 GBLUP 方法达到了 0.503 7,其他方法的 MSE 异常所以只保留 GBLUP 方法数据;在猪背膘和猪乳头数的表型中 LD_PCC 的 GBLUP 模型均最好,分别达到了 0.566 5 和 0.506 2 的准确率。

为了更直观的比较各方法在 5 组数据的表现,针对每种 7 种方法分别对每组数据的 5 种基因组选



横坐标为不同统计模型。AGE. 猪百千克日龄;BF. 猪背膘厚;TN. 猪乳头数;Breast_width. 鸭胸宽;Fossil_lenth. 龙骨长。LD_GWAS+数据均表示LD结合GWAS方法,数据表示GWAS阈值

The abscissa represents different statistical models. AGE. Age to 100 kg; BF. Backfat thickness; TN. Teat number; Breast_width. Duck breast width; Fossil_lenth. Duck fossil lenth. LD_GWAS+data indicates LD-integrated GWAS with a specified GWAS significance threshold

图2 预测准确性对比

Fig. 2 Comparison of prediction accuracy

择方法取平均值及最大值,得到表 5 和表 6。由表 5、表 6 可知,所有表型 LD_PCC 方法平均值在所有表型中预测相关性都处在前两位,LD_PCC 方法在猪的 3 个性状上比原始数据有 6.4%~13.9% 的提升,在鸭子的数据上有 4% 左右的提升。猪的 3 个

性状中表现最差的多重筛选方法为 LD_PCA,3 组数据均只有 0.167~0.180 的相关性,但此方法在鸭子的两个表型中表现良好,均高于原始数据相关性,其可能原因放在讨论部分分析。

表 5 杜洛克猪预测结果

Table 5 Duroc pig prediction results

表型 Phenotype	方法 Method	平均相关性 Mean PCC	标准差 STD
百千克日龄 Age to 100 kg	origin	0.386 1	0.031 9
	LD	0.424 0	0.037 3
	LD_GWAS10	0.429 7	0.036 2
	LD_PCA	0.176 0	0.037 4
	LD_outPCA	0.378 9	0.048 9
	LD_PCC	0.440 0	0.035 6
	LD_EPI	0.435 9	0.034 3
背膘厚 Backfat thickness	origin	0.437 9	0.040 9
	LD	0.455 1	0.050 4
	LD_GWAS10	0.461 1	0.050 5
	LD_PCA	0.180 8	0.050 2
	LD_outPCA	0.396 6	0.063 4
	LD_PCC	0.497 4	0.044 0
	LD_EPI	0.482 2	0.046 9
乳头数 Teat number	origin	0.413 4	0.041 7
	LD	0.419 8	0.031 2
	LD_GWAS12	0.428 2	0.029 6
	LD_PCA	0.167 3	0.032 0
	LD_outPCA	0.409 6	0.040 4
	LD_PCC	0.440 1	0.035 0
	LD_EPI	0.426 5	0.032 3

LD_GWAS 表示 LD 结合 GWAS 标记中方法效果最好的阈值。下同

LD_GWAS represents the threshold for the best performance of the LD combined with GWAS markers method. The same as below

2.3 计算时间对比

基因组学数据通常特别庞大,因此对设备计算性能要求较高。本次研究的运算 CPU 为 Intel(R) Xeon(R) CPU E7-4820 v4 @ 2.00GHz 四颗,共计 80 个线程,1TB 运行内存为了对比不同方法进行多重筛选后的计算性能,计算时间与位点数量呈高度的正相关。同时 GBLUP 虽位点增长速度比其他方法慢,位点较多的情况下,GBLUP 比其他方法运算速度更高,PCA 相关方法虽然计算时间最快,但是结果不稳定,在猪的 3 个性状上相关性极差。相比 PCA,LD_PCC 方法在获得最高预测精度同时所需

的计算时间较短,且准确率稳定提高。

2.4 基于功能注释信息的权重分配法结果

判断筛选方法后,我们将猪背膘和猪乳头数两组数据质控筛选后的 SNP 位点通过 ANNOVAR 软件进行了注释,其中 upstream、downstream 默认参数为上、下游各 1 kb。注释过程采用 *Sus scrofa*11 参考基因组数据库。通过注释,将这些 SNPs 位点分配到不同的基因区域。具体分布如表 7 所示。

2.5 权重分配法预测准确性比较

加权(weighted)和未加权过(default)的背膘厚和乳头数 SNP 矩阵在 6 种模型中的平均预测准确

表 6 北京鸭预测结果

Table 6 Peking duck prediction results

表型 Phenotype	方法 Method	平均相关性 Mean PCC	标准差 STD
鸭胸宽 Duck breast width	origin	0.632 6	0.060 2
	LD	0.643 7	0.052 6
	LD_GWAS1	0.648 7	0.054 5
	LD_PCA	0.662 1	0.051 9
	LD_outPCA	0.632 6	0.055 4
	LD_PCC	0.656 9	0.046 3
	LD_EPI	0.636 9	0.049 0
龙骨长 Duck fossil lenth	origin	0.680 3	0.072 7
	LD	0.695 2	0.067 6
	LDGWAS2	0.702 6	0.061 7
	LD_PCA	0.705 9	0.071 4
	LD_outPCA	0.691 4	0.068 1
	LD_PCC	0.709 3	0.065 4
	LD_EPI	0.686 9	0.071 6

表 7 杜洛克猪 SNPs 功能注释和分类

Table7 Functional annotation and classification of SNPs in Duroc pigs

区域 Region	计数 Count	百分比/% Percentage
基因间区 Intergenic	169 033	74.84
内含子 Intronic	50 617	22.40
外显子 Exonic	2 231	0.99
上游 Upstream	1 511	0.67
下游 Downstream	1 489	0.66
3'非翻译区 UTR3	237	0.10
5'非翻译区 UTR5	90	0.04
非编码 RNA 外显子 ncRNA exonic	62	0.03
剪接(作用)Splicing	46	0.02
上游;下游 Upstream;downstream	28	0.01
外显子;剪接 Exonic;splicing	2	0.001

性和标准差如表 8 所示。

由表 8 可知,加权过的背膘厚 SNP 矩阵在所有模型中的预测表现均优于未加权背膘厚 SNP 矩阵。在其中 4 种模型中,加权过的背膘厚 SNP 矩阵表现出显著的预测准确性提升。这表明,加权方法能够更有效地捕捉与背膘厚相关的加性效应等,提高预测性能。

加权过的乳头数 SNP 矩阵在所有模型中的预测表现也优于未加权过的乳头数 SNP 矩阵。在

GBLUP 模型中,加权后的乳头数 SNP 矩阵表现出显著的预测准确性提升,平均 PCC 从 0.349 0 提高到 0.389 7。然而乳头数预测准确性普遍低于背膘厚,可能是由于乳头数的遗传力较低,使预测难度增加^[37]。

如表 8 所示,展示了两种性状在不同模型中的预测表现。总体来看,加权过的 SNP 矩阵在背膘厚和乳头数的预测中均表现出较高的预测准确性,表明加权方法在基因组选择中具有广泛的适用性,有助于提高模型的鲁棒性和预测准确性。

表 8 不同模型在加权和默认 SNP 数据中的预测表现

Table 8 Predictive performance of different models on weighted and default SNP data

模型 Model	背膘厚加权 Backfat thickness weighted	背膘厚未加权 Backfat thickness default	乳头数加权 Teat number weighted	乳头数未加权 Teat number default
GBLUP	0.410 8(0.028 2)	0.355 7(0.031 0)	0.389 7(0.022 8)	0.349 0(0.025 1)
BayesA	0.383 1(0.024 2)	0.358 6(0.026 4)	0.344 0(0.023 4)	0.322 8(0.030 7)
BayesB	0.349 4(0.029 9)	0.323 9(0.048 7)	0.306 5(0.032 1)	0.286 6(0.036 9)
BayesC	0.298 2(0.017 3)	0.257 2(0.037 1)	0.249 3(0.018 9)	0.240 5(0.032 7)
BL	0.400 3(0.022 7)	0.364 8(0.045 1)	0.366 3(0.018 1)	0.345 7(0.030 5)
BRR	0.395 7(0.023 0)	0.374 9(0.046 4)	0.353 7(0.027 8)	0.347 6(0.031 1)

BL. Bayes LASSO ; BRR. Bayesian Ridge Regression。括号中数字表示标准差

BL. Bayes LASSO; BRR. Bayesian Ridge Regression. The numbers in parentheses indicate the standard deviation

3 讨 论

本研究中预测准确性方面,LD_PCC方法有优越的表现,背膘厚和猪乳头数上均优于Zhu等^[23]使用的SLDP方法,可能源于该方法对不同染色体之间相关性的考虑,能够减轻共线性和降低特征的复杂度^[38],但试验中猪百千克日龄在除GBLUP的其他方法上MSE异常,可能是由于该性状受环境及管理因素影响较大,而模型中未能充分纳入这些固定效应信息,导致方差分解不稳定。尤其是贝叶斯类模型对方差先验较为敏感,在样本量有限或固定效应缺失的情况下容易出现收敛不稳,从而使预测误差(MSE)显著增大^[39]。

在Ma等^[28]的研究中,EPISNP方法能够计算位点之间的加性效应、互作效应和上位效应等,本研究中LD_EPI方法没有显著提升的原因可能是使用的育种模型主要计算加性效应,后续将使用深度学习等模型捕捉非线性效应。

LD_PCA和在大部分性状上的预测准确性都大幅度下降,这与吴华煊和杜志强^[27]的研究结果一致,可能是PCA会导致部分重要特征的丢失。LD_outPCA结果下降幅度不大,在鸭子的体宽性状上有上升,可能是通过PCA分析原来移除的位点,能够减少部分重要位点的丢失。

Lopez等^[40]使用的加权方法主要是根据GWAS结果进行权重分配,受限于测序数据量,并且不具有普适性,基于生物数据库中SNP位点的功能区域进行位点加权,不受限于表型数据,直接对外显子区和剪接区域SNP赋予较高权重而基因间区的SNP被赋予较低权重,对比未加权方法能显著提高模型的预测准确性^[41],可能主要是由于外显子区和

剪接区域的变异直接影响蛋白质的结构和功能,进而影响性状,而基因间区的变异对基因功能的影响较小^[42,43]。其次加权也能有效地降低噪音影响。低权重的SNP在模型中的影响力减小,使得主要依赖于重要功能区域的预测结果更加可靠^[44]。权重分配的合理性在很大程度上决定了模型的表现。

在运行时间方面,可以观察到在相同模型下,运算时间与位点数量正相关。不同模型对比中,随着位点数量增加,GBLUP的计算时间增长较慢。这是因为GBLUP的GRM是一个固定维度的对称矩阵^[45,46],位点增多时,每个位点对GRM的贡献逐渐减少,因此其计算复杂度增加较小。而当位点较少时,GBLUP的矩阵求解步骤相对开销较小,但整体数据量小可能导致计算速度较慢。同时,五折交叉验证在多次模型重估过程中会显著增加计算时间,其主要目的在于评估模型在新数据上的预测误差,并不专门针对模型预测的无偏性进行评估,因此在无偏性分析方面存在一定局限。相比之下,Leagarra和Reverter^[47]提出的LR(linear regression)方法可通过对部分与全部数据集预测值之间的回归,直接估计预测的无偏性和离散度。该方法计算复杂度较低,在需要评估模型预测无偏性时,LR方法在特定应用场景下具有优势。后续研究将尝试采用LR方法对模型的无偏性进行进一步验证与分析。

4 结 论

本试验以杜洛克猪以及北京鸭多个性状及对应基因型数据为研究对象,通过对比基因组选择预测结果与真实结果的相关性,可知LD_PCC方法在精确度上有很好的提升,而权重分配法在突出重要

位点和降低噪音干扰方面效果显著,使得预测结果提升显著。后续研究应进一步优化位点标记和权重分配方式,并探索更多新型算法和基因组先验知识,以提升基因组选择育种的精度。

参考文献(References):

- [1] MEUWISSEN T H E, HAYES B J, GODDARD M E. Prediction of total genetic value using genome-wide dense marker maps[J]. *Genetics*, 2001, 157: 1819-1829.
- [2] YUAN C, GILLON A, GALDRON D, et al. Evaluation of genomic selection models using whole genome sequence data and functional annotation in Belgian Blue cattle[J]. *Genet Sel Evol*, 2025, 57: 10.
- [3] MAKANJUOLA B O, MIGLIOR F, ABDALLA E A, et al. Effect of genomic selection on rate of inbreeding and coancestry and effective population size of Holstein and Jersey cattle populations[J]. *J Dairy Sci*, 2020, 103: 5183-5199.
- [4] MEUWISSEN T H E, HAYES B J, GODDARD M E. Genomic selection: A paradigm shift in animal breeding[J]. *Anim Front*, 2016, 6(1): 6-14.
- [5] 尹立林, 马云龙, 项 韬, 等. 全基因组选择模型研究进展及展望[J]. 畜牧兽医学报, 2019, 50(2): 233-242.
- YIN L L, MA Y L, XIANG T, et al. The progress and prospect of genomic selection models [J]. *Acta Veterinaria et Zootechnica Sinica*, 2019, 50(2): 233-242. (in Chinese)
- [6] MIN H, JING Z, ZHU Y, et al. Whole genome sequencing revealed genetic structure, domestication, and selection of Chinese indigenous ducks [J]. *Int J Biol Macromol*, 2025, 306: 141724.
- [7] 杨岸奇, 陈 斌, 冉茂良, 等. 基因组选择在猪杂交育种中的应用[J]. 遗传, 2020, 42(2): 145-152.
- YANG A Q, CHEN B, RAN M L, et al. The application of genomic selection in pig cross breeding [J]. *Hereditas*, 2020, 42(2): 145-152. (in Chinese)
- [8] 窦腾飞, 吴嘉浩, 吴姿仪, 等. 基因组选择和选配技术在猪育种中的应用[J]. 畜牧兽医学报, 2024, 55(7): 2795-2808.
- DOU T F, WU J H, WU Z Y, et al. Application progress of genomic selection and mating allocation techniques in pig breeding [J]. *Acta Veterinaria et Zootechnica Sinica*, 2024, 55(7): 2795-2808. (in Chinese)
- [9] LIU T F, QU H, LUO C L, et al. Genomic selection for the improvement of antibody response to Newcastle disease and avian influenza virus in chickens [J]. *PLoS One*, 2014, 9: 112685.
- [10] 祁军英, 裴全帮, 张文魁, 等. 全基因组选择信号鉴定高原型藏羊毛用性状候选基因及关联分析[J]. 畜牧兽医学报, 2024, 55(12): 5511-5526.
- QI J Y, PEI Q B, ZHANG W K, et al. Genome-wide selective signal identification and association analysis of candidate genes for tibetan sheep wool traits [J]. *Acta Veterinaria et Zootechnica Sinica*, 2024, 55(12): 5511-5526. (in Chinese)
- [11] 贾宏霞, 刘在霞, 周 乐, 等. 基因组选择在肉牛中的研究进展[J]. 畜牧兽医学报, 2024, 55(9): 3757-3768.
- JIA H X, LIU Z X, ZHOU L, et al. Research progress of genomic selection in beef cattle [J]. *Acta Veterinaria et Zootechnica Sinica*, 2024, 55(9): 3757-3768. (in Chinese)
- [12] 谈 成, 边 成, 杨 达, 等. 基因组选择技术在农业动物育种中的应用[J]. 遗传, 2017, 39(11): 1033-1045.
- TAN C, BIAN C, YANG D, et al. Application of genomic selection in farm animal breedin [J]. *Hereditas*, 2017, 39(11): 1033-1045. (in Chinese)
- [13] 李棉燕, 王立贤, 赵福平. 机器学习在动物基因组选择中的研究进展[J]. 中国农业科学, 2023, 56(18): 3682-3692.
- LI M Y, WANG L X, ZHAO F P, et al. Research progress on machine learning for genomic selection in animals [J]. *Scientia Agricultura Sinica*, 2023, 56(18): 3682-3692. (in Chinese)
- [14] MEUWISSEN T, EIKJE L S, GJUVSLAND A B. GWABLUP: genome-wide association assisted best linear unbiased prediction of genetic values [J]. *Genet Sel Evol*, 2024, 56: 17.
- [15] 林晓坤, 都萌萌, 周李生, 等. 敖汉细毛羊羊毛经济性性状的全基因组关联分析[J]. 畜牧兽医学报, 2024, 55(10): 4346-4359.
- LIN X K, DU M M, ZHOU L S, et al. Genome-wide association study of wool economic traits in Aohan fine wool sheep [J]. *Acta Veterinaria et Zootechnica Sinica*, 2024, 55(10): 4346-4359. (in Chinese)
- [16] ALEMU S W, LOPDELL T J, TREVARTON A J, et al. Comparison of genomic prediction accuracies in dairy cattle lactation traits using five classes of functional variants versus generic SNP [J]. *Genet Sel Evol*, 2025, 57: 20.

- [17] 黄红艳,张力允,黄智荣,等. 狮头鹅群体遗传多样性和体重体尺全基因组关联分析[J]. 畜牧兽医学报, 2024,55(9):3914-3924.
HUANG H Y, ZHANG L Y, HUANG Z R, et al. The study on population genetic diversity and genome-wide association study of body weight and size traits for Lion-head geese [J]. *Acta Veterinaria et Zootechnica Sinica*, 2024, 55(9): 3914-3924. (in Chinese)
- [18] 刘勃攻,杨 杰,肖 健,等. 利用BLUP和Bayes方法对湘佳黑凤鸡S1系产蛋性状基因组育种值预测效果的比较研究[J]. 中国家禽, 2025,47(1):1-7.
LIU B G, YANG J, XIAO J, et al. Comparative study on the effect of predicting genomic breeding values for laying traits of Xiangjia Heifeng S1 line chicken using BLUP and Bayes methods [J]. *China Poultry*, 2025, 47(1): 1-7 (in Chinese)
- [19] 杜永旺,黄 超,王一东,等. 结合GWAS先验标记信息的肉鸡RFI性状全基因组选择研究[J]. 畜牧兽医学报, 2022,53(10):3403-3411.
DU Y W, HUANG C, WANG Y D, et al. Genomic selection for RFI in broiler combining GWAS prior marker information [J]. *Acta Veterinaria et Zootechnica Sinica*, 2022, 53(10): 3403-3411. (in Chinese)
- [20] MICHAELSON J J, LOGUERCIO S, BEYER A. Detection and interpretation of expression quantitative trait loci (eQTL) [J]. *Methods*, 2009, 48: 265-276.
- [21] 袁泽湖,葛 玲,李发弟,等. 整合生物学先验信息的全基因组选择方法及其在家畜育种中的应用进展[J]. 畜牧兽医学报, 2021,52(12):3323-3334.
YUAN Z H, GE L, LI F D, et al. The Method of genomic selection by integrating biological prior information and its application in livestock breeding [J]. *Acta Veterinaria et Zootechnica Sinica*, 2021, 52(12): 3323-3334. (in Chinese)
- [22] DENG M T, ZHU F, YANG Y Z, et al. Genome-wide association study reveals novel loci associated with body size and carcass yields in Pekin ducks [J]. *BMC Genomics*, 2019, 20:1.
- [23] ZHU D, ZHAO Y, ZHANG R, et al. Genomic prediction based on selective linkage disequilibrium pruning of low-coverage whole-genome sequence variants in a pure Duroc population [J]. *Genet Sel Evol*, 2023, 55:72.
- [24] PURCELL S, NEALE B, TODD-BROWN K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses [J]. *Am J Hum Genet*, 2007, 81: 559-575.
- [25] BERG I V D, BOICHARD D, GULDBRANDTSEN B, et al. Using sequence variants in linkage disequilibrium with causative mutations to improve across-breed prediction in dairy cattle: a simulation study [J]. *G3 (Bethesda)*, 2016, 6: 2553-2561.
- [26] 白 露,王梦杰,马小春,等. 一种快速挖掘鸡品种特征性SNP标记集合的方法[J]. 畜牧兽医学报, 2023, 54(8):3252-3261.
BAI L, WANG M J, MA X C, et al. A method for quickly mining the characteristic SNP markers set of chicken [J]. *Acta Veterinaria et Zootechnica Sinica*, 2023, 54(8): 3252-3261. (in Chinese)
- [27] 吴华煊,杜志强. 基因型特征提取方法影响基因组选择预测准确性的研究[J]. 畜牧兽医学报, 2024, 55(6):2431-2440.
WU H X, DU Z Q. Methods of genotype feature extraction affecting the prediction accuracy of genomic selection [J]. *Acta Veterinaria et Zootechnica Sinica*, 2024, 55(6): 2431-2440. (in Chinese)
- [28] MA L, RUNESHA H B, DVORKIN D, et al. Parallel and serial computing tools for testing single-locus and epistatic SNP effects of quantitative traits in genome-wide association studies [J]. *BMC Bioinform*, 2008, 9:315.
- [29] ZHAO P J, ZHENG X R, YU Y, et al. Mining unknown porcine protein isoforms by tissue-based map of proteome enhances pig genome annotation [J]. *Genom Proteom Bioinform*, 2021, 19: 772-786
- [30] 潘 艺,韦佳霖,孙敬春,等. 全基因组关联分析筛选大白猪产木乃伊数性状候选基因[J]. 中国畜牧杂志, 2024,60(8):210-216.
PAN Y, WEI J L, SUN J C, et al. Genome-wide association analysis identifies candidate genes for mummified piglet traits in Large White pigs [J]. *Chinese Journal of Animal Science*, 2024, 60(8): 210-216. (in Chinese)
- [31] 王 琦,朱 迪,王宇哲,等. 全基因组SNP分型策略及基因组预测方法的研究进展[J]. 畜牧兽医学报, 2020,51(2):205-216.
WAN Q, ZHU D, WANG Y Z et al. Research progress of genomic-wide SNP genotyping and genomic prediction methods [J]. *Acta Veterinaria et Zootechnica Sinica*, 2020, 51(2): 205-216. (in Chinese)
- [32] SVENSSON E I, ARNOLD S J, BÜRGER R, et al. Correlational selection in the age of genomics [J]. *Nat Ecol Evol*, 2021, 5: 562-573.
- [33] MEUWISSEN T H. Accuracy of breeding values of

- 'unrelated' individuals predicted by dense SNP genotyping [J]. *Genet Sel Evol*, 2009, 41: 35.
- [34] ZHAO Y S, GOWDA M, LIU W X, et al. Accuracy of genomic selection in European maize elite breeding populations[J]. *Theor Appl Genet*, 2012, 124: 769-776.
- [35] BADRÉ A, ZHANG L, MUCHERO W, et al. Deep neural network improves the estimation of polygenic risk scores for breast cancer[J]. *J Hum Genet*, 2021, 66: 359-369.
- [36] 王延晖, 朱 波, 李俊雅. 基于显性模型的基因组选择中贝叶斯方法研究[J]. 畜牧兽医学报, 2017, 48(1): 60-67.
- WANG Y H, ZHU B, LI J Y. Bayesian models including dominant effects for genomic selection [J]. *Acta Veterinaria et Zootechnica Sinica*, 2017, 48(1): 60-67. (in Chinese)
- [37] ZHU H, ZHOU X. Statistical methods for SNP heritability estimation and partition: A review [J]. *Comput Struct Biotechnol*, 2020, 18: 1557-1568.
- [38] 习海娇, 李金泉, 张燕军, 等. 显性效应对内蒙古绒山羊产绒量和绒纤维直径育种值估计准确性的影响[J]. 畜牧兽医学报, 2025, 56(2): 571-581.
- XI H J, LI J Q, ZHANG Y J, et al. Influence of dominance effects on the accuracy of breeding value estimation of cashmere production and cashmere diameter in Inner Mongolia Cashmere goats [J]. *Acta Veterinaria et Zootechnica Sinica*, 2025, 56(2): 571-581. (in Chinese)
- [39] ZHANG C, KEMP R, STOTHARD P, et al. Genomic evaluation of feed efficiency component traits in Duroc pigs using 80K, 650K and whole-genome sequence variants[J]. *Genet Sel Evol*, 2018, 50: 14.
- [40] LOPEZ B I M, AN N, SRIKANTH K, et al. Genomic prediction based on SNP functional annotation using imputed whole-genome sequence data in Korean Hanwoo cattle [J]. *Front Genet*, 2021, 11: 603822.
- [41] ZHANG Z, OBER U, ERBE M, et al. Erbe improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies[J]. *PLoS One*, 2014, 9: e93017
- [42] 胡紫平, 苏艳芳, 宗文成, 等. 基于转录组和基因组数据筛选剑白香猪两头乌毛色的候选基因[J]. 中国畜牧兽医, 2025, 52(1): 13-24.
- HU Z P, SU Y F, ZONG W C, et al. Screening of candidate genes for two-end black coat color in screening of candidate genes for two-end black coat color in [J]. *China Animal Husbandry & Veterinary Medicine*, 2025, 52(1): 13-24. (in Chinese)
- [43] 吴相哲, 胡潘阳, 郑丽香, 等. 基于选择信号分析挖掘豫南黑猪生长性状关键基因[J]. 中国畜牧杂志, 2024, 60(8): 284-289.
- WU X Z, HU P Y, ZHENG L X, et al. Identification of key genes affecting growth traits in Southern Henan Black pigs based on selection signal analysis [J]. *Chinese Journal of Animal Science*, 2024, 60(8): 284-289. (in Chinese)
- [44] 盛中华, 陈国宏, 潘玉春, 等. 上海白猪(上系)基因组遗传变异检测与功能注释分析[J]. 畜牧兽医学报, 2018, 49(5): 887-896.
- SHENG Z H, CHEN G H, PAN Y C, et al. Identification and functional annotation of genome-wide genetic variation in Shanghai White Pigs (Shang) [J]. *Acta Veterinaria et Zootechnica Sinica*, 2018, 49(5): 887-896. (in Chinese)
- [45] CANO-GAMEZ E, TRYNKA G. From GWAS to function: using functional genomics to identify the mechanisms underlying complex diseases [J]. *Front Genet*, 2020, 11: 424.
- [46] CLARK S A, WERF J V D. Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values[J]. *Methods Mol Biol*, 2013, 1019: 321-330.
- [47] LEAGARRA A, REVERTER A. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method[J]. *Genet Sel Evol*, 2018, 50(1): 53.
- (编辑 郭云雁)