畜牧兽医学报 2017,48(12):2258-2267
Acta Veterinaria et Zootechnica Sinica
doi: 10.11843/j.issn.0366-6964.2017.12.005

基于 GBLUP 与惩罚类回归方法的猪血液性状基因组选择研究

张巧霞[1],张玲妮[1],刘 飞[1],刘向东[1],刘小磊[1],赵书红[1,2],朱猛进[1,2*]
(1. 华中农业大学 农业动物遗传育种与繁殖教育部重点实验室,武汉 430070;
2. 生猪健康养殖协同创新中心,武汉 430070)

摘 要: 旨在探讨 GBLUP 与惩罚类回归方法用于猪血液性状基因组选择的相关问题。以本实验室收集的免疫资源猪群体 13 个血液性状为分析对象,结合 Illumina 公司猪 SNP60K 基因芯片分型数据,以加性模型和加性-显性模型为基础,利用 GBLUP 和 3 种惩罚类回归方法(ridge、lasso 与 elastic-net)开展基因组选择分析。研究发现,基因组选择的准确性与性状芯片遗传力估计值呈正相关。交叉验证分析结果表明,4 种方法对 13 个血液性状预测准确性最高的性状均是 MCV(平均红细胞体积),而加性模型和加性-显性模型的预测准确性在不同性状中的表现不同。在多数性状中,lasso 和 elastic-net 回归的预测准确性低于 ridge 回归和 GBLUP 法,但在 NE%(嗜中性细胞百分比)等少数性状中则刚好相反。综上说明,没有适用于所有性状的最佳基因组预测方法,基因组预测方法的选择应考虑目标性状的遗传特性。本研究为猪免疫性状基因组选择的实际应用提供了重要参考信息。

关键词: 猪;血液性状;基因组选择;GBLUP;惩罚类回归
中图分类号: S828.2 **文献标志码:** A **文章编号:** 0366-6964(2017)12-2258-10

A Study of Genomic Selection on Porcine Hematological Traits Using GBLUP and Penalized Regression Methods

ZHANG Qiao-xia[1], ZHANG Ling-ni[1], LIU Fei[1], LIU Xiang-dong[1], LIU Xiao-lei[1], ZHAO Shu-hong[1,2], ZHU Meng-jin[1,2*]
(1. *Key Laboratory of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education, Huazhong Agricultural University, Wuhan* 430070, *China*;
2. *The Cooperative Innovation Center for Sustainable Pig Production, Wuhan* 430070, *China*)

Abstract: This study aimed to explore the application of GBLUP and penalized regression methods in genomic selection of the hematological traits in pigs. We chose 13 hematological traits from the immune resource population collected by our laboratory as the analyzed objects. We used the genotyping data of Illumina PorcineSNP60 Genotyping Beadchip to conduct the genomic selection analysis, in which GBLUP and 3 penalized regression methods (ridge, lasso and elastic-net) were used based on additive model and additive-dominance model. The results showed that the accuracy of genomic selection was positively correlated with estimated values of chip heritabilities of traits. The results of cross-validation analysis showed that the MCV (mean corpuscular volume) had the highest prediction accuracy among 13 hematological traits. The prediction accuracy of additive model and additive-dominance model were different in different traits. In total trend, the predic-

收稿日期: 2017-03-14
基金项目: 国家自然科学基金面上项目(31372302;31672392);国家高技术研究发展计划(2013AA102502);湖北省公益性科技研究项目(2012DBA25001);国家生猪产业技术体系项目(CARS-35)
作者简介: 张巧霞(1991-),女,河南信阳人,硕士,主要从事动物遗传育种研究,E-mail:qiaoxiazhang@webmail.hzau.edu.cn
* **通信作者:** 朱猛进,博士,副教授,硕士生导师,主要从事统计基因组学研究,Tel:027-87281306,E-mail:zhumengjin@mail.hzau.edu.cn

tion accuracy of the lasso and elastic-net regressions were lower than that of the ridge regression and GBLUP. But in a few traits, such as NE%, it was opposite. In conclusion, there is no optimal genomic prediction method that could be suitable for all traits, and we should consider the genetic characteristics of the target traits when choosing a genome evaluation method. This research provides important reference information for the practical application of genomic selection for immune traits in pigs.

Key words: porcine; hematological trait; genomic selection; GBLUP; penalized regression

我国是当今世界猪肉产量与消费量最大的国家,养猪业在农业生产中占据着举足轻重的作用。随着我国养猪业的不断发展,猪病也呈现出病原谱复杂化、爆发周期密集化的趋势,猪病已成为制约我国养猪业健康、稳定发展的重要因素之一。培育高抗病力猪种,可以从源头提高猪群抗病力,既符合新时期绿色健康养殖的理念,又能在某种程度上达到降低猪群发病几率、控制猪场疫情的目的。所以,近年来猪的抗病育种工作开始受到国内外学者的高度重视。

继最佳线性无偏预测(Best Linear Unbiased Prediction,BLUP)、分子标记辅助选择(Marker-Assisted Selection,MAS)、分子标记聚合育种(Pyramiding Breeding)之后,基因组选择(Genomic Selection,GS)已开始逐渐成为当前家畜育种的重要手段[1]。基因组选择最早由 T. H. Meuwissen 等[2]于 2001 年提出。基因组选择可视为在全基因组水平开展标记辅助选择的一种新遗传评估方法,与其他遗传评估方法相比较,基因组选择具有预测准确率高、世代间隔短、选择效率高[3-4]等优点。目前,基因组选择已在奶牛育种中得到了广泛应用,如中国[5]、澳大利亚、新西兰、美国、法国、德国、荷兰[6-10]等国已相继展开了基于基因组选择的奶牛育种工作,并取得了显著成效[6, 11-12]。近年来,以丹育(DanBred)为代表的国际大型猪育种公司已在育种实践中全面应用基因组选择技术,但国内猪基因组选择技术的实际应用还处于起步阶段。

猪抗病力及其组分性状多为中低遗传力性状[13],常规育种取得的遗传进展较慢。随着基因组学技术的高速发展,以猪基因组信息为基础的分子育种技术已成为猪经济性状改良的重要手段[14]。基因组选择已被学界公认优于传统选择指数或 BLUP 法,基因组选择技术为猪抗病力的遗传改良提供了有效手段。血液性状是猪抗病力的重要组分性状,血液指标可间接反映猪个体的健康状况与抗病潜力。血液性状采集、测定相对容易,对猪造成的伤害和应激较小,无论是抗病力机理解析,还是育种实践应用,血液性状均是较为理想的指示性状。因此,开展猪血液性状基因组选择相关的研究具有重要的意义。

鉴于此,本研究拟利用我室已有的免疫试验猪群数据,开展与猪抗病力密切相关的血液免疫性状的基因组选择研究,利用 GBLUP、ridge 回归、lasso 回归、elastic-net 回归 4 种常用基因组选择方法,通过预测准确性的比较,筛选猪血液性状基因组选择合适的方法,从而为猪血液性状基因组选择的实际应用提供重要参考信息。

1　材料与方法

1.1　试验猪群

试验猪群来自我室构建的免疫资源群体,原始数据样本包含大白猪、二花脸猪的 F_2 遗传设计交配后代 1 月龄仔猪 394 头,所有入试个体来自同一猪场,饲养管理条件和方式、营养水平、免疫程序均一致,系谱、测定批次、性别信息完整。数据经预处理,剔除表型缺失值,保留样本含量达到 200 以上的性状,最后纳入分析的性状包括白细胞(WBC)、嗜中性粒细胞(NE)、淋巴细胞(LY)、单核细胞(MO)、嗜酸性粒细胞(EO)、嗜中性粒细胞百分比(NE%)、淋巴细胞百分比(LY%)、单核细胞百分比(MO%)、红细胞(RBC)、血红蛋白(HGB)、平均红细胞体积(MCV)、红细胞分布宽度(RDW)、血小板(PLT)等 13 个血液性状。

1.2　基因组 SNPs 数据处理

免疫试验猪群 SNPs 数据由 Illumina 公司猪 60K 芯片(Illumina PorcineSNP60 Genotyping Beadchip)测定。用经典的酚/氯仿法(CITE HERE)从试验猪群采集的耳朵或尾巴组织提取全基因组 DNA。所有 DNA 样本经检测合格,终浓度标准化为 50 ng·μL^{-1},然后送交商业化公司完成基因芯片杂交试验。获得原始数据后,对基因型数据开展质量控制分析,质控标准

设置：基因型缺失率 10%、检出率或杂交阳性率(Call rate)90%、最小等位基因频率(MAF)1%、以及是否偏离哈代温伯格平衡(HWE)($P>0.05$)。基因型缺失值采用 R 程序包 synbreed 的 codeGeno 命令进行填充(Imputation)，填充参数 nmiss=0.1，最后获得 59 559 个有效 SNPs 用于基因组选择的交叉验证分析。

1.3 基因组选择模型

1.3.1 GBLUP 模型 GBLUP 用全基因组标记构建的关系矩阵(即 G 矩阵)代替了传统 BLUP 由系谱构建的分子血缘矩阵(即 A 矩阵)[7,11]。GBLUP 的模型：

$$y = X\beta + Zu + e^{[7,11,15]}$$

其中，y 为观测值向量，X 为固定效应设计矩阵，Z 为随机效应设计矩阵，β 为固定效应向量，u 为随机效应向量，e 为残差向量。

1.3.2 Ridge 回归模型 在全基因组回归模型 $y=X\beta+e$ 中，由于 SNP 数目(p)远大于样本含量(n)，不能用常规最小二乘方法(Ordinary Least Squares，OLS)估计每个标记的参数，构成了所谓的 p>>n 回归问题。惩罚类回归利用稀疏假设(相当于假设基因组中绝大部分 SNP 效应为零)对最小二乘估计过程施加正则化项(Regularizer)或惩罚项(Penalty term)约束，将常规最小二乘法转化为惩罚最小二乘法(Penalized Least Squares)，从而实现非零标记效应的参数估计。Ridge 回归的惩罚项叫做 L2 惩罚，即 ridge 回归使用 L2 惩罚最小二乘法(L2 Penalized Least Squares)实现 p>>n 的高维回归参数估计问题：

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$[16]。

式中，$\|y-X\beta\|_2^2$ 为最小二乘项，$\lambda\|\beta\|_2^2$ 为 L2 惩罚项，为 β 平方的 λ 倍。

1.3.3 Lasso 回归模型 Lasso 回归通过对基因组 SNPs 的回归系数实施 L1 惩罚(L1 Penalty)解决 p>>n 的回归问题：

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

式中，$\|y-X\beta\|_2^2$ 为最小二乘项，$\lambda\|\beta\|_1$ 为 L1 惩罚项，为 β 绝对值的 λ 倍。

1.3.4 Elastic-net 回归模型 Elastic-net 回归组合了 lasso 回归与 ridge 回归的特点，将 L1 惩罚和 L2 惩罚限制性条件同时纳入参数估计过程[17-19]：

$$\hat{\beta} = \left(\frac{1+\lambda_2}{n}\right)\{\arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda_2\|\beta\|_2^2 + \lambda_1\|\beta\|_1\}$$

式中，$\|y-X\beta\|_2^2$ 为最小二乘项，$\lambda_2\|\beta\|_2^2+\lambda_1\|\beta\|_1$ 分别对应 ridge 和 lasso 回归中的 L2、L1 惩罚项。

1.4 基因组选择参数设置和分析工具

基因组选择交叉验证(Cross-validation)的参数设置：重复数为 25 和 50，交叉验证倍数(k-fold)设为 5 和 10。对于 5-倍交叉验证，先将样本随机等分为 5 份，然后每 1 份样本依次作为测试集，剩余 4 份作为训练集，依次循环 5 次，使每份样本轮流充当 1 次测试集，估计预测值与实际值的相关系数，上述过程重复 25 或 50 次，最后用相关系数的平均值作为预测准确性的评估指标。在 10-倍交叉验证中，每次重复则将样本随机等分为 10 份，其中 9 份为训练集，1 份为测试集，依次循环 10 次。基于加性模型的 GBLUP 用于比较，而 ridge、lasso 和 elastic-net 回归则同时运行加性模型和加性-显性模型，其中加性模型以“0-1-2”编码基因型，而加性-显性模型则以“0-1-2”和“0-1-1”编码基因型，同时将加性基因型矩阵和显性基因型矩阵纳入分析模型。统计分析在 R 语言(3.2.2 版本)环境中完成，其中程序包 synbreed[20]用于基因型数据填充(Imputation)，惩罚类回归由程序包 glmnet[21]实现，GBLUP 由 rrBLUP[22]实现。

2 结 果

2.1 各性状遗传力的估计

遗传力由基因组 SNPs 信息构建的遗传关系矩阵替代传统加性血缘关系矩阵而估计，该遗传力又称为芯片遗传力(Chip heritability)。利用填充后的基因型矩阵，构建加性遗传关系矩阵，建立混合线性模型，用平均信息约束最大似然法算法(Average Information Restricted Maximum Likelihood，AIREML)估计遗传方差与剩余方差，通过加性遗传方差占总方差的比例估计各性状的遗传力。通过刀切法(Jackknife)依次抽去一条记录、保留其他记录估计遗传力，循环直至每条记录被依次抽去一次，再利用循环所得遗传力向量估计平均值和标准误，各性状的遗传力均值和标准误估计结果见表 1。MCV 性状的遗传力最高，HGB 次之，MO 与 MO% 的遗传力最低。除 NE、MO 和 MO% 性状属于低遗传力性状外，其他血液性状的遗传力估计值均达到了中、高遗传力的水平。

表 1　各性状遗传力估计结果

Table 1　Results of heritability estimation for each trait

性状 Trait	遗传力 Heritability	性状 Trait	遗传力 Heritability
WBC	0.377 0±0.014 5	MO%	0.034 1±0.015 5
NE	0.081 2±0.024 8	RBC	0.254 3±0.021 8
LY	0.325 8±0.018 1	HGB	0.503 6±0.012 2
MO	0.049 4±0.014 3	MCV	0.626 5±0.012 1
EO	0.277 8±0.019 4	RDW	0.471 3±0.019 2
NE%	0.344 0±0.021 3	PLT	0.183 0±0.019 5
LY%	0.426 7±0.015 5		

2.2　基于 GBLUP 的基因组选择交叉验证

使用不同重复次数和交叉验证倍数的参数组合,用 GBLUP 法对各性状开展基因组预测,用预测值与真实值之间的相关系数表示预测准确性,各性状的预测准确性见表 2(表 3～5 中的数字亦为相关系数估计的预测准确性)。预测准确性最高的是 MCV 性状,HGB 次之,准确性最低的是 MO 与 MO%性状。对照各性状遗传力估计值与 GBLUP 预测准确性,可发现 GBLUP 的预测准确性大致表现出与各性状遗传力呈正比的关系。

表 2　各性状基于 GBLUP 的交叉验证分析结果

Table 2　Results of cross-validation based on GBLUP for each trait

性状 Trait	repeat=10, k-fold=10	repeat=25, k-fold=10	repeat=50, k-fold=5
WBC	0.211 6	0.210 7	0.195 4
NE	0.103 7	0.104 5	0.070 2
LY	0.164 8	0.152 6	0.159 0
MO	0.017 9	0.017 9	0.009 1
NE%	0.162 7	0.172 7	0.163 2
LY%	0.221 1	0.240 3	0.233 7
MO%	0.009 3	0.011 8	0.010 1
RBC	0.121 9	0.174 9	0.125 1
HGB	0.267 9	0.289 1	0.258 7
MCV	0.315 0	0.320 5	0.316 8
RDW	0.184 4	0.230 0	0.197 6
PLT	0.059 2	0.081 1	0.058 9
EO	0.170 9	0.188 0	0.181 9

2.3　基于 ridge 回归的基因组选择交叉验证

表 3 展示了不同重复次数和交叉验证倍数组合下,各性状基于 ridge 回归的交叉验证分析结果。由于基因组选择多以加性模型为主,同时考虑到加性-显性模型因纳入了加性基因型矩阵和显性基因型矩阵,基因型矩阵倍增而使得计算负担大、耗时长,所以在本研究中 3 种惩罚类回归的加性-显性模型均只分析了重复数为 25、交叉验证倍数为 10 的一种参数组合。如表 3 所示,ridge 回归在总体趋势上与 GBLUP 预测结果相似,其中预测准确性最高的是 MCV 性状,HGB 次之,准确性最低的是 MO 与 MO%性状。从性状内来看,不同交叉验证倍数的预测准确性略有差异,交叉验证倍数与预测准确性间的关系在不同性状中的表现并不一致。在多数性状中,随着交叉验证倍数的增加,基因组选择的验证结果准确性也随之增加。另外,比较相同交叉验证参数,除 LY、HGB、RDW 等少数性状外,基于 ridge 回归的加性-显性模型的预测准确性表现出比加性模型略高的趋势。

2.4　基于 lasso 回归的基因组选择交叉验证

使用加性模型和加性-显性模型的 lasso 回归分析结果见表 4。与 GBLUP、ridge 回归的结果相似,lasso 回归预测准确性最高的是 MCV 性状,准确性最低的也是 MO、MO%与 PLT 性状。但与 ridge 回归不同的是,lasso 的加性-显性模型预测准确性不及加性模型,而且在相同交叉验证参数组合下,MO、MO%和 PLT 的加性模型、以及 WBC、RDW 和 PLT 的加性-显性模型的预测准确性出现负相关系数估计值。而且,在 MO、MO%、NE 与 EO 性状

表 3 各性状基于 ridge 回归的交叉验证分析结果

Table 3 Results of cross-validation based on ridge regression for each trait

性状 Trait	加性模型 Additive model			加性-显性模型 Additive-dominance model
	repeat=10, k-fold=10	repeat=25, k-fold=10	repeat=50, k-fold=5	repeat=25, k-fold=10
WBC	0.213 7	0.207 8	0.198 9	0.217 5
NE	0.104 5	0.102 7	0.094 5	0.113 1
LY	0.182 5	0.200 4	0.202 4	0.185 1
MO	0.005 8	0.011 1	0.017 6	0.037 9
NE%	0.223 9	0.208 9	0.193 8	0.223 6
LY%	0.264 1	0.274 0	0.254 9	0.289 7
MO%	0.011 3	0.025 8	0.031 7	0.047 3
RBC	0.238 0	0.232 0	0.215 5	0.272 4
HGB	0.289 0	0.293 7	0.276 3	0.277 8
MCV	0.327 0	0.335 1	0.331 1	0.334 0
RDW	0.242 7	0.246 5	0.243 4	0.229 5
PLT	0.108 9	0.087 4	0.103 5	0.118 3
EO	0.149 8	0.152 4	0.130 2	0.218 7

表 4 各性状基于 lasso 回归的交叉验证分析结果

Table 4 Results of cross-validation based on lasso regression for each trait

性状 Trait	加性模型 Additive model			加性-显性模型 Additive-dominance model
	repeat=10, k-fold=10	repeat=25, k-fold=10	repeat=50, k-fold=5	repeat=25, k-fold=10
WBC	0.031 6	0.045 8	0.077 3	−0.038 3
NE	NA	NA	0.021 1	NA
LY	0.159 2	0.149 7	0.123 1	0.116 1
MO	NA	NA	−0.080 7	NA
NE%	0.246 3	0.251 1	0.235 1	0.201 9
LY%	0.204 5	0.238 4	0.195 0	0.180 0
MO%	−0.006 4	−0.029 4	−0.038 1	NA
RBC	0.211 3	0.239 1	0.157 8	0.231 6
HGB	0.233 2	0.261 4	0.203 2	0.246 5
MCV	0.312 5	0.309 5	0.290 4	0.280 6
RDW	0.074 4	0.094 3	0.066 6	−0.023 6
PLT	−0.027 1	−0.069 1	−0.016 2	−0.054 4
EO	NA	NA	NA	NA

中,lasso 回归出现了无效预测值或无区分度的预测值,导致预测值和实际值之间的相关系数估计值为无效值(NA)。这说明 lasso 回归在处理小样本数据时,其鲁棒性(Robustness)不及 GBLUP 和 ridge 回归。

2.5　基于 elastic-net 回归的基因组选择交叉验证

以 0.1 为步长,在 0.1 至 0.9 区间内筛选各性状 elastic-net 回归的最优 alpha 参数,预测准确性最高对应的 alpha 值为最优,其中图 1 显示了在重复数为 10、交叉验证倍数为 10 的加性模型下,各性状基因组预测的准确性与 alpha 值的对应曲线(该参数组合下 MO 和 EO 性状无数据)。图中每条曲线代表一个性状的基因组预测准确性随 alpha 值变化而变化的情况,每条曲线最高点对应的 X 轴值为该性状的优化 alpha 值。在经过 alpha 参数优化筛选后,得到加性模型和加性-显性模型 elastic-net 回归的交叉验证分析结果(表 5)。如表 5 所示,预测准确性最高的同样是 MCV 性状。与 lasso 回归分析相似,elastic-net 回归的加性-显性模型预测准确性亦不及加性模型。另外,elastic-net 回归分析结果也出现了预测值与实际值之间相关系数的负估计值和 NA 估计值。这不难理解,由于 elastic-net 回归兼具了 lasso 回归与 ridge 回归的特点,所以 lasso 回归所遭遇的统计问题,elastic-net 回归也不能避免。

2.6　4 种方法基因组预测效果的比较

图 2 为 3 种惩罚类回归方法与 GBLUP 在各个交叉验证参数组合以及不同模型下的各性状预测准确性的直接比较。为直观显示不同 alpha 值的预测准确性差异,图 2 同时给出了 elastic-net 回归全部 alpha 值的预测准确性,图中条柱从左至右依次为 elastic-net 回归(alpha 参数为 0.1、0.2、0.3、0.4、0.5、0.6、0.7、0.8、0.9)、ridge 回归、lasso 回归及 GBLUP 的预测准确性。从图 2 显示的趋势看,ridge 回归与 GBLUP 的预测准确性比 elastic-net 回归与 lasso 回归高,特别是在 WBC、NE、MO、MO%、RDW、PLT 等性状中,ridge 回归与 GBLUP 预测准确性明显高于 elastic-net 回归和 lasso 回归。单独比较 ridge 回归与 GBLUP 法的准确性,当重复次数较大(如 r=50)时,除 EO 性状外,在大多数性状中 ridge 回归的预测准确性都达到甚至超过了GBLUP方法,尤其是 RBC、LY、RDW 和 NE 性状,ridge 回归的预测准确性分别比 GBLUP 提高了 77.9%、28.0%、

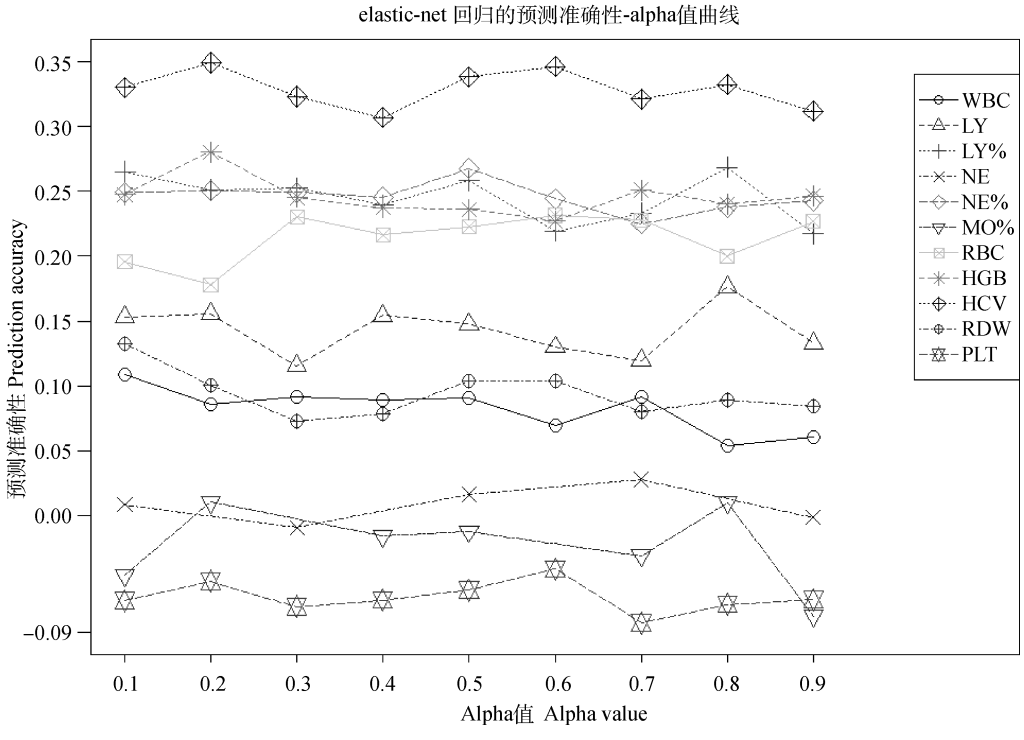


图 1　各性状 elastic-net 回归的优化 Alpha 参数筛选

Fig. 1　Optimal selection of alpha parameters of elastic-net regression for each trait

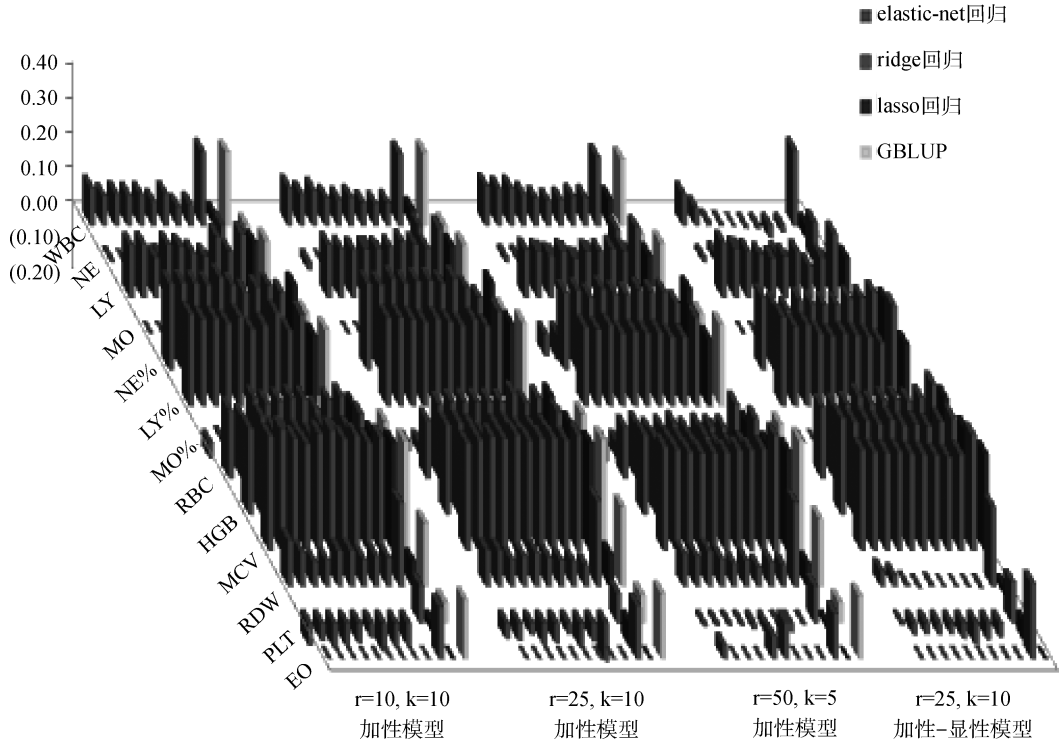
表 5 各性状基于 elastic-net 回归的交叉验证分析结果

Table 5 Results of cross-validation based on elastic-net regression for each trait

性状 Trait	加性模型 Additive model			加性-显性模型 Additive-dominance model
	repeat=10，k-fold=10	repeat=25，k-fold=10	repeat=50，k-fold=5	repeat=25，k-fold=10
WBC	0.108 7(alpha=0.1)	0.111 4(alpha=0.1)	0.114 0(alpha=0.1)	0.091 0 (alpha=0.1)
NE	0.027 7(alpha=0.7)	0.037 1(alpha=0.6)	0.024 5(alpha=0.4)	0.033 4(alpha=0.4)
LY	0.176 4(alpha=0.8)	0.160 9(alpha=0.7)	0.141 8(alpha=0.8)	0.159 1(alpha=0.1)
MO	NA	NA	−0.053 3(alpha=0.8)	NA
NE%	0.267 6(alpha=0.5)	0.262 8(alpha=0.3)	0.233 3(alpha=0.4)	0.210 6(alpha=0.1)
LY%	0.268 5(alpha=0.8)	0.256 7(alpha=0.2)	0.232 8(alpha=0.3)	0.258 7(alpha=0.1)
MO%	0.010 2(alpha=0.2)	−0.003 3(alpha=0.8)	−0.020 3(alpha=0.7)	NA
RBC	0.231 5(alpha=0.6)	0.243 6(alpha=0.8)	0.157 9(alpha=0.5)	0.235 1(alpha=0.4)
HGB	0.280 4(alpha=0.2)	0.263 4(alpha=0.2)	0.229 1 (alpha=0.1)	0.277 2(alpha=0.1)
MCV	0.349 4(alpha=0.2)	0.333 5(alpha=0.5)	0.298 8(alpha=0.1)	0.290 2 (alpha=0.1)
RDW	0.132 5(alpha=0.1)	0.104 9(alpha=0.1)	0.102 0(alpha=0.3)	0.044 9 (alpha=0.1)
PLT	−0.041 4(alpha=0.6)	−0.032 4(alpha=0.6)	−0.004 7 (alpha=0.9)	−0.005 4(alpha=0.1)
EO	NA	NA	0.104 2 (alpha=0.6)	NA

表中括号内数字为最优 alpha 值

The numbers in brackets are the optimal alpha values



Elastic-net 回归中 alpha 参数为 0.1、0.2、0.3、0.4、0.5、0.6、0.7、0.8、0.9

The alpha parameters of elastic-net regression are 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9

图 2 3 种回归方法与 GBLUP 在不同的重复次数与 k-fold 值以及不同模型下的预测准确性比较

Fig. 2 Comparison of 3 regression methods and GBLUP for predicting accuracy of each trait in different repetition times and k-fold values under different models

23.8%和 22.3%。不过,在加性模型中,NE%性状的预测准确性出现了相反的结果,elastic-net 回归与 lasso 回归的预测准确性要高于 ridge 回归与 GBLUP,而在加性-显性模型中,并没有这一现象。

3　讨　论

本研究较为系统地探讨了 GBLUP 与 3 种惩罚类回归方法用于猪 13 个血液性状基因组选择的相关问题。首先,研究发现基因组选择的准确性与性状遗传力估计值呈正相关,4 种方法对 13 个血液性状预测准确性最高的性状均是 MCV(平均红细胞体积),准确性最低的均为 MO 与 MO%性状,而且性状遗传力对基因组选择准确性的影响与所使用的基因组选择方法无关。虽然基因组选择可提高低遗传力性状的选择效果,但这并不意味着基因组选择应该首选低遗传力性状,相反在育种实践中,为了取得最佳的遗传进展,应用基因组选择技术时仍需优先考虑高遗传力性状。

本研究重点探讨了 GBLUP、ridge、lasso 和 elastic-net 回归法对各性状基因组选择的适用性,发现 4 种基因组选择分析方法在不同性状中的预测准确性并不一致。对于 LY、MO%、HGB、RDW 和 PLT 性状,ridge 回归是预测准确性最高的方法,而对于 WBC、NE 和 MO 性状,在不同交叉验证参数组合下,ridge 回归与 GBLUP 各有优势,但总体上两者预测效果大致接近。虽然 GBLUP 是基因组选择最常用的方法之一,但与惩罚类回归方法相比,GBLUP 在预测准确性上并没有表现出优势。对于 3 种惩罚类回归方法,由于 elastic-net 回归方法兼具了 lasso 和 ridge 回归的优点[16],一般认为 elastic-net 回归的预测准确性应该高于 lasso 和 ridge 回归。但从本研究的实际结果来看,elastic-net 回归的预测准确性在多个性状上均不如 ridge 回归,只在 NE%等少数性状中,elastic-net 回归才优于 ridge 回归。从统计学角度讲,ridge 回归是实现"组选择"(Group selection),反映到基因组预测过程就是选中每个有效应的 LD 模块,而 lasso 回归则是从每个组中选择代表性变量,即只选中每个 LD 模块中的代表性 SNP,elastic-net 则是实现部分的组选择,选中 SNPs 界于整个 LD 模块与代表性 SNP 之间[22]。所以,可能是因为各血液性状调控基因在猪基因组中的分布特性(即 LD 结构),导致了 elastic-net 回归在多个性状上的表现均不如 ridge 回归。另外,本研究还发现,当交叉验证倍数较大,预测小样本数据时,ridge 回归和 GBLUP 法的鲁棒性要高于 lasso 回归和 elastic-net 回归。综合来看,对于基因组预测方法的选择,可能没有绝对的标准,最适基因组选择方法应是性状特异性的,没有适用于所有性状的最佳基因组选择方法,基因组选择方法的选择应考虑目标性状的遗传特性。

此外,本研究亦简单揭示了影响交叉验证结果的其他因素,包括模型(加性模型、加性-显性模型)、交叉验证倍数(k-fold)和重复数对交叉验证过程中基因组预测准确性的影响。虽然有效群体大小、染色体长度、标记数量与密度、参考群体的规模和结构、连锁不平衡状态、显性效应等诸多因素均有可能影响基因组选择的效果[23-24],但由于本研究分析的对象是已存在的真实数据集,其有效群体含量、猪基因组结构以及性状的遗传特性均已固定,这些因素均无法探讨。对于模型、交叉验证倍数和重复数,经研究发现,加性模型与加性-显性模型的预测准确性存在差异,从全部 13 个血液性状的预测结果来看,加性模型和加性-显性模型的预测准确性在不同性状中的表现不同,每种模型都有准确性最高或准确性相对较高的预测结果,显然并不存在适用于所有性状的最优模型。对于交叉验证倍数,大致表现出随着交叉验证倍数的增加,预测准确性随之增加的趋势,但超过一定的验证倍数后,其预测准确性可能反而下降,尤其是统计鲁棒性较低的 lasso 和 elastic-net 回归法,甚至可能出现无效预测或无区分度的预测结果。所以,在开展基因组选择的交叉验证时,需要综合性状遗传特性、以及样本含量等因素确定适宜的交叉验证倍数。由于交叉验证分组是采用随机分组法,理论上存在参考群和测试群均值出现显著差异的分组概率,所以在交叉验证过程中,重复次数越多,基因组预测的平均结果也越准确。不过,重复次数与计算量(计算时间)是对应的,重复次数应根据样本含量与计算机性能综合确定。

4　结　论

本研究应用 GBLUP 和 3 种惩罚类回归方法(ridge、lasso 与 elastic-net)对猪 13 个血液性状进行了基因组选择分析。研究发现,性状遗传力对基因组选择的准确性有显著的影响,无论何种方法,基因组选择准确性最高的均是 MCV(平均红细胞体积)性状,准确性最低的均是 MO(单核细胞)和 MO%

(单核细胞百分比)性状。在预测小样本数据时,ridge 回归和 GBLUP 法的鲁棒性要高于 lasso 回归和elastic-net 回归。除 NE%等少数性状外,在多数性状中 ridge 回归和 GBLUP 的预测准确性要优于 lasso 回归和 elastic-net 回归。本研究结果提示,没有适用于所有性状的最优方法,基因组选择方法是性状特异性的,基因组选择方法的选取,需要根据所研究性状的遗传特性进行优化和筛选。本研究为猪免疫性状基因组选择的实际应用提供了重要参考信息。

参考文献(References):

[1] 李娅兰,梅盈洁,刘敬顺,等. 基因组选择及其在猪育种中的应用[J]. 广东农业科学,2012,39(17):106-109.
LI Y L, MEI Y J, LIU J S, et al. Genomic selection and application on swine breeding[J]. *Guangdong Agricultural Sciences*, 2012, 39(17): 106-109. (in Chinese)

[2] MEUWISSEN T H, HAYES B J, GODDARD M E. Prediction of total genetic value using genome-wide dense marker maps[J]. *Genetics*, 2001, 157(4): 1819-1829.

[3] DAETWYLER H D, VILLANUEVA B, BIJMA P, et al. Inbreeding in genome-wide selection[J]. *J Anim Breed Genet*, 2007, 124(6): 369-376.

[4] 王 晨,秦 珂,薛 明,等. 全基因组选择在猪育种中的应用[J]. 畜牧兽医学报,2016,47(1):1-9.
WANG C, QIN K, XUE M, et al. Application of genomic selection in swine breeding[J]. *Acta Veterinaria et Zootechnica Sinica*, 2016, 47(1): 1-9. (in Chinese)

[5] DING X, ZHANG Z, LI X, et al. Accuracy of genomic prediction for milk production traits in the Chinese Holstein population using a reference population consisting of cows[J]. *J Dairy Sci*, 2013, 96(8): 5315-5323.

[6] HAYES B J, BOWMAN P J, CHAMBERLAIN A J, et al. Invited review: genomic selection in dairy cattle: progress and challenges[J]. *J Dairy Sci*, 2009, 92(2): 433-443.

[7] VANRADEN P, TOOKER M. Methods to explain genomic estimates of breeding value[J]. *J Dairy Sci*, 2007, 90(S1): 374.

[8] MOSER G, TIER B, CRUMP R E, et al. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers[J]. *Genet Sel Evol*, 2009, 41: 56.

[9] WINKELMAN A M, JOHNSON D L, HARRIS B L. Application of genomic evaluation to dairy cattle in New Zealand[J]. *J Dairy Sci*, 2015, 98(1): 659-675.

[10] COLOMBANI C, LEGARRA A, FRITZ S, et al. Application of Bayesian least absolute shrinkage and selection operator (LASSO) and BayesCπ methods for genomic selection in French Holstein and Montbéliarde breeds[J]. *J Dairy Sci*, 2013, 96(1): 575-591.

[11] HABIER D, TETENS J, SEEFRIED F R, et al. The impact of genetic relationship information on genomic breeding values in German Holstein cattle[J]. *Genet Sel Evol*, 2010, 42: 5.

[12] SCHAEFFER L. Strategy for applying genome-wide selection in dairy cattle[J]. *J Anim Breed Genet*, 2006, 123(4): 218-223.

[13] 施启顺. 畜禽抗病育种研究进展[J]. 中国畜牧杂志,1995,31(6):48-51.
SHI Q S. Research progress of breeding for disease resistance in livestock and poultry[J]. *Chinese Journal of Animal Science*, 1995, 31(6): 48-51. (in Chinese)

[14] 彭中镇,赵书红,刘 榜,等. 自主育种是中国种猪企业提升种猪竞争力和走向世界的必由之路[J]. 中国猪业,2016,11(4):8-18.
PENG Z Z, ZHAO S H, LIU B, et al. Independent Breeding is the only way for Chinese Boar Breeding Enterprises to enhance their competitiveness and go to the world[J]. *China Swine Industry*, 2016, 11(4): 9-18. (in Chinese)

[15] EL-KASSABY Y A, KLÁPŠTĚ J, GUY R D. Breeding without breeding: selection using the genomic best linear unbiased predictor method (GBLUP)[J]. *New Forest*, 2012, 43(5-6): 631-637.

[16] OGUTU J O, SCHULZ-STREECK T, PIEPHO H P. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions[J]. *BMC Proc*, 2012, 6 S2: S10.

[17] ZOU H, HASTIE T. Regularization and variable selection via the elastic net[J]. *J R Stat Soc Series B Stat Methodol*, 2005, 67(5): 301-320.

[18] CHO S, KIM H, OH S, et al. Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis[J]. *BMC Proc*, 2009, 3(S7):

S25.

[19] WU T T, CHEN Y F, HASTIE T, et al. Genome-wide association analysis by lasso penalized logistic regression [J]. *Bioinformatics*, 2009, 25(6): 714-721.

[20] WIMMER V, ALBRECHT T, AUINGER H J, et al. Synbreed: a framework for the analysis of genomic prediction data using R[J]. *Bioinformatics*, 2012, 28(15): 2086-2087.

[21] FRIEDMAN J, HASTIE T, TIBSHIRANI R. Regularization paths for generalized linear models via coordinate descent[J]. *J Stat Softw*, 2010, 33(1): 1-22.

[22] ENDELMAN J B. Ridge regression and other kernels for genomic selection with R package rrBLUP[J]. *Plant Genome*, 2011, 4(3): 250-255.

[23] 张　哲, 张　勤, 丁向东. 畜禽基因组选择研究进展[J]. 科学通报, 2011, 56(26): 2212-2222.
ZHANG Z, ZHANG Q, DING X D. Advances in genomic selection in domestic animal[J]. *Chinese Science Bulletin*, 2011, 56(26): 2212-2222. (in Chinese)

[24] 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)

（编辑　郭云雁）