大豆科学 Soybean Science
http://www.haasep.cn
2018,37(3):353-358
DOI:10.11861/j.issn.1000-9841.2018.03.0353

对比 Bayesian B 等多种方法的大豆全基因组选择应用研究

唐 友[1],郑 萍[2],王嘉博[3],张继成[2]

(1. 吉林农业科技学院 电气与信息工程学院,吉林 吉林 132101; 2. 东北农业大学 电气与信息学院,黑龙江 哈尔滨 150030; 3. 黑龙江省农业科学院 畜牧研究所,黑龙江 哈尔滨 150086)

摘 要:为提高育种家选种选育的效率,运用统计学手段对两组大豆的基因型数据及对应的真实性状设置不同遗传力及有效位点数,模拟不同性状进行运算来预测精准值。采用交叉验证重复百次使预测值趋于稳定,对比分析 Bayesian B 及常见的几种基因组选择方法。结果图形直观展示出 Bayesian B 方法效果较好且较稳定,说明 Bayesian B 方法在大豆全基因组选择上有较明显的优势,可以在大豆育种分析中辅助使用。同时,介绍了 Bayesian B 方法的应用,以便指导大豆品种选育。

关键词:Bayesian B;基因组选择;交叉验证;选种选育

Application Research for Soybean Genomics Selection by Comparing Bayesian B and Other Methods

TANG You[1], ZHENG Ping[2], WANG Jia-bo[3], ZHANG Ji-cheng[2]

(1. Electrical and Information Engineering College, Jilin Agricultural Science and Technology University, Jilin 132101, China; 2. College of Electrical and Information, Northeast Agricultural University, Harbin 150030, China; 3. Institute of Animal Husbandry, Heilongjiang Academy of Agricultural Sciences, Harbin 150086, China)

Abstract: In order to improve the breeding efficiency, different heritability and effective loci were set up by the genotype data and corresponding real characters of two groups of soybean to simulate the accuracy of different characters, statistical methods were used to simulate different characters to predict precise values. Cross validation repeated hundreds of times to stabilize the prediction value, compare and analyze Bayesian B and several common methods of genomic selection. The results showd that the Bayesian B method was better and more stable, indicating that the Bayesian B method had obvious advantages in the selection of soybean whole genome, which could be used in soybean breeding. At the same time, the application of Bayesian B method was introduced to guide the soybean breeding.

Keywords: Bayesian B; Genomic selection; Cross validation; Selection and breeding

传统大豆育种方式改进性状效果不明显[1],随着测序分型技术的改进及费用的大幅度降低,高通量测序产生了大量的大豆基因型数据,因此将分子标记辅助选择引入到大豆全基因组选择来提高育种分析效率,成为现在育种家们亟待考虑的问题[2]。而利用统计模型分析基因型数据和表型数据,进而进行关联运算是重要的研究手段。常见的方法有利用线性回归的最小二乘法(least squares)、gBLUP 法(genomic Best Linear Unbiased Prediction)、rrBLUP 法(ridge regression Best Linear Urbiased Prediction),但最小二乘法速度快效率低,gBLUP 法比较适合在动物预测模型中,rrBLUP 法则需要引入个体关系作为预测依据,这 3 种方法对于大豆每个标记效果的计算对于基因组预测影响都考虑不足[3]。而 Bayesian B 方法通过设置有效抽样比例,并对每个标记进行计算和裁剪,使其作为预测的主要因素,可以使构建的预测模型更稳定精准,近年来应用比较广泛。因此,本文对大豆基因型数据和表型性状数据进行训练,并通过交叉验证计算精准值,对比分析这 4 种方法,以期找到比较适合大豆育种分析的全基因组选择方法来辅助育种家提高育种效率。并在此基础上,介绍筛选出的更准确且稳定的方法,以帮助用户选种选育优良品种,达到提高产量或增大效益的目的。

收稿日期:2018-02-05
第一作者简介:唐友(1979-),男,博士,教授,高级工程师,主要从事生物信息学、软件工程研究。E-mail:tangyou@neau.edu.cn。
通讯作者:张继成(1980-),男,博士,高级工程师,主要从事农业机器化研究。E-mail:zhangjicheng@neau.edu.cn。

1 材料与方法

1.1 材料

真实数据主要来源于 NCBI(https://www.ncbi.nlm.nih.gov/gene)和在公开发表的论文上下载的公共数据。主要数据涉及 2 组大豆数据:标记 2 个表型性状,170 个大豆个体基因型数据的2 174 个遗传标记(marks);307 个个体的7 125个 marks。

针对基因型数据具体模拟表型过程:

首先,设置遗传力 h^2,模拟有效位点的个数(NQTN)符合正态分布,随机抽取 SNP marks 来获取有效的基因型数据,模拟不同情况下的表型性状。

其次,模拟运算过程,遗传方差的获取主要有两种方法,符合正态随机分布或几何分布。通过随机获取遗传值,与有效的 SNP 标记矩阵值进行运算,求取方差即是遗传方差。

再次,通过设定遗传力与所求方差运算,得到环境方差即为残差,然后取 0 到残差开方符合正态分布的随机抽取环境值。

最后,将模拟的每个个体遗传值和环境值求和,得到模拟性状数值。

1.2 基因组选择方法

当前基因组选择方法(genomic selection,GS)较多,本文比较最小二乘法、gBLUP、rrBLUP 和 Bayesian B 方法在大豆全基因组选择中的应用效果,着重介绍 Bayesian B 方法在育种选择中所适合的群体运算,几种方法的特点如下:

(1)最小二乘法

该方法利用最小化误差的平方和去寻找与数据最佳匹配的函数,再根据该方法简便地求出未知的数据,并使得这些求出的数据与实际数据之间的误差的平方和达到最小[4]。对每个标记作单点回归分析的公式如下:

$$y = \mu l_n + X_g + e$$

在模型里放入所选择最大效应值中的 m 点,然后进行评估运算。

$$y = \mu l_n + \sum X_i g_i + e$$

其特点:在随机情况下、有相当好的统计特性,并容易编程实现,在很多领域通用。但有两方面的缺陷:一是噪声模型时,得到的估计值不是无偏差、完全一致的;二是随着数据量增加,将出现所谓的"数据饱和"现象,容易高估标记效应。

(2)gBLUP

因为 gBLUP 扩展了一般线性模型,所以数据表现出很多的可变性。该方法根据分子标记亲缘信息构建关系矩阵来代替传统的系谱信息 A 矩阵,分析 BLUP 值[5]。假设所有染色体片段具有相同的效应方差,所得到的公式如下:

$$y = Wf + \sum_{j-1}^{n} X_j g_j + e$$

$$y = Wf + Z_a + \sum_{j-1}^{n} X_j g_j + e$$

其特点:gBLUP 对于多基因控制的性状预测效果更好,和传统的系谱信息相比更准确和容易操作,但是相对复杂,考虑因素多,模型稳定困难。

(3)rrBLUP

rrBLUP 又称为随机回归最佳线性无偏预测,它是一个随机效应的预测,具有线性、无偏和预测误差方差最小等统计学性质[6]。根据基因型效应估计个体的育种值得到的公式如下:

$$y = Wf + Z_a + e$$

$$V_{ar}(a) = G\sigma_a^2$$

其特点:能更有效地校正环境效应,充分利用所有亲属的信息,校正由于非随机交配造成的偏差,对不同群体进行联合遗传评定,提高育种值估计的精确性。

(4)Bayesian B

2001 年由 Meuwissen 等[7]提出,是 Alphabet 之一,主要构建马尔可夫链,并采用蒙特卡罗方法 MCMC(markov chain monte carlo)进行抽样运算。简单说是指使用先验概率按照设定的参数来确定部分假设有效,然后再通过后验概率假设进行实际抽样运算,结果收敛更稳定,且准确性更高[8]。贝叶斯 B 的基本运算如下:

假设部分 SNP 位点都有效应,标记位点其效应(∂_j^2)分布为方差(σ_j^2)服从尺度逆卡方分布的正态分布。并采用 t 分布标记为先验分布,标记效应符合正态分布,具体分布效应方差为 $\nu_\alpha S_\alpha^2 \chi_{\nu_\alpha}^{-2}$,符合尺度逆卡方分布,两个参数自由度 ν_α 和初度 S_α^2[9]。

$$a_j \mid \pi, \sigma_j^2 \sim (idd) \begin{cases} 0 & \sigma_j^2 = 0 \ \ with\ probability\ \pi \\ \sim N(0, \partial_j^2)\sigma_j^2 > 0 & with\ probability (1-\pi) \end{cases}$$

$where\ j = 1, \cdots, M$

$$\sigma_j^2 \mid v_\alpha, S_\alpha^2 \sim (idd) v_\alpha S_\alpha^2 \chi_{v_\alpha}^{-2}\ when\ \sigma_j^2 > 0$$

其特点:适应各种情况的预测,在个体性状数目少的情况下,有较高的准确性,但耗时长。

1.3 试验设计

试验硬件环境为 CPU 八核处理器、16G 内存的服务器。将获得的高通量大豆基因型数据转换成可以分析计算的数值格式,并通过 GAPIT2 工具[10]分析基因型数据特点。标记密度分布呈直方图和累计分布(图 1),对表型性状数据进行分析,散点图

表示表型值位置分布;直方图表示表型值呈正态分布;箱线图表示表型观察值整体分布统计,其中以平均值设为中位线,最大最小值为上下边界,异常值会超出两端边界;累积分布图表示表型值大小累积分布趋势(图 2)。通过分析基因型和表型数据构建训练群体,然后引入到 Bayesian B、最小二乘法、gBLUP、rrBLUP 基因组选择方法中进行运算,并设置交叉验证,采用皮尔森相关系数得到精准值,再迭代重复百次以上消除偶然性,最后对比结果以确定更精准且稳定的全基因组选择方法,获得更有参考价值的育种值,服务于育种家。

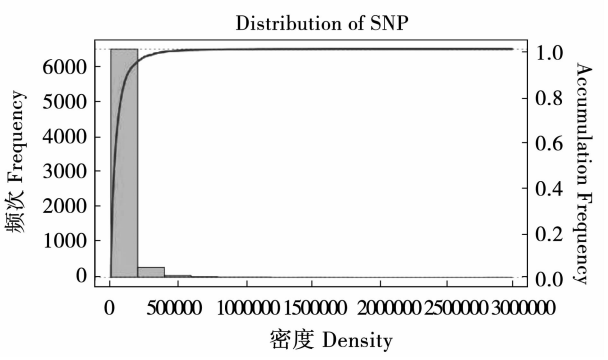
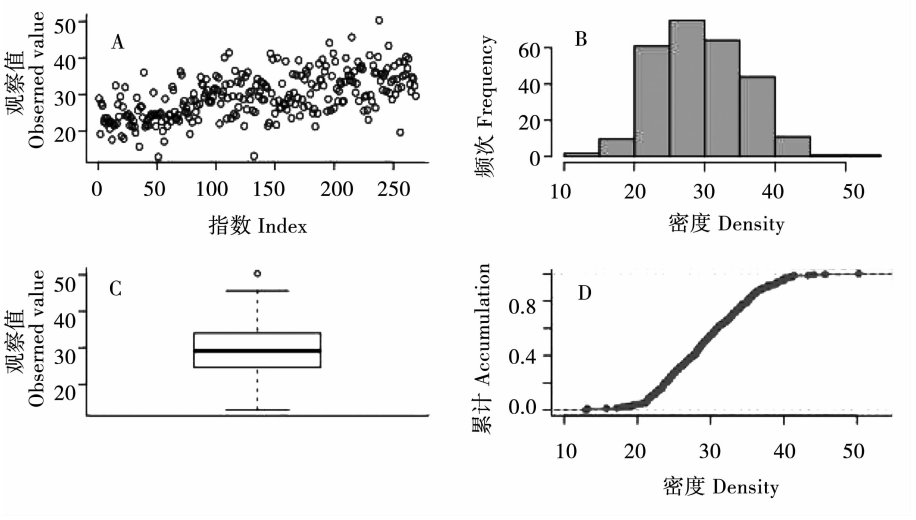


图 1　分析基因型数据

Fig. 1　Analysis of genotype data



A:散点图;B:直方图;C:箱线图;D:累积分布图。

A:Scatter diagram;B:Histogram;C:Box plots;D:Cumulative distribution diagram.

图 2　分析表型性状数据

Fig. 2　Analysis of phenotypic character data

2　结果与分析

2.1　4 种全基因组选择方法对比分析

2.1.1　*真实数据结果分析*　将两组真实大豆数据进行基因组选择测试,匹配基因型数据和表型数据形成训练群体,然后将不同性状引入到 4 个方法中运算,通过交叉验证计算皮尔森相关系数 Accuracy,也就是表型预测的精准值,百次平均值数据结果如表 1。表明 Bayesian B 方法在处理这些物种的一些性状时准确性高于最小二乘法、rrBLUP、gBLUP,说明对于表型数据预测可以采用该方法作为基因组选择的方法。

表 1　不同的真实性状下采取不同基因组选择方法预测精准值的比较

Table 1　Comparison of different genome selection methods for predicting the accuracy of different real traits

大豆(个体,标记) Soybean(individuals,marks)	性状 Trait	gBLUP	Bayes B	rrBLUP	最小二乘法 Least squares
大豆(170,2174)	性状 1	0.88817	0.88692	0.88037	0.46432
Soybean(170,2174)	性状 2	0.53649	0.52730	0.47482	0.34123
大豆(307,7125)	性状 1	0.59447	0.61979	0.58615	0.39432
Soybean(307,7125)	性状 2	0.58778	0.60885	0.55003	0.42871

2.1.2　*模拟数据结果分析*　模拟大豆(307 个体,7 125个 marks)基因数据对应表型性状,通过设置大豆不同遗传力 h^2 和有效位点 NQTN 的个数来进行模拟,然后进行 4 个方法运算,计算结果对比,得到平均 Accuracy 数据(表 2)。显示这 4 个方法中,最小二乘法预测的准确性最低,Bayesian B 方法预测的准确性最高。

表 2 模拟在不同遗传力和 NQTN 个数的性状下采取不同基因组选择方法预测精准值的比较

Table 2 Comparison of the prediction accuracy of different genome selection methods under different genetic and NQTN numbers

大豆(个体,标记) Soybean(individuals,marks)	模拟性状 Stimulated trait(h^2,NQTN)	gBLUP	Bayesian B	rrBLUP	最小二乘 Least squares
大豆(307,7125)	0.25,5	0.12444	0.18909	0.14644	0.04678
Soybean(307,7125)	0.50,5	0.24424	0.50418	0.26028	0.21770
	0.75,5	0.38981	0.54868	0.39347	0.17231
	0.90,5	0.48283	0.55896	0.48562	0.34328
	0.25,50	0.18402	0.19495	0.13339	0.05621
	0.50,50	0.30793	0.33486	0.34092	0.14361
	0.75,50	0.42459	0.44023	0.44147	0.17778
	0.90,50	0.54581	0.55468	0.54315	0.24445

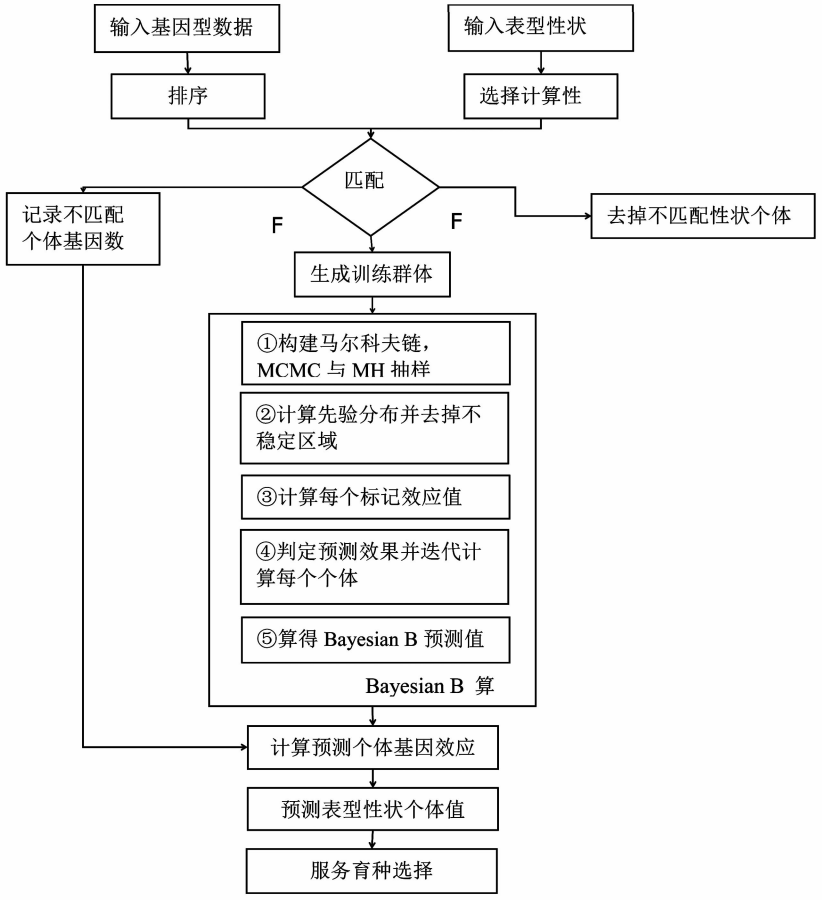


图 3 试验设计及实现流程图

Fig. 3 Experimental design and implementation flow chart

2.2 Bayesian B 方法应用

将基因型数据与实际测得的个体表型性状数据输入到 Bayesian B 方法中运算(图 3)。步骤如下:

(1)处理基因型数据将碱基标记转换成数值格式,并按照个体进行排序。同时将输入的表型数据进行选择。

(2)将处理的表型和基因数据按照个体进行匹配,去掉没有匹配上的个体表型性状,记录下没有匹配上的个体基因型数据,以备后面预测未知表型性状值时使用。

(3)将符合匹配的表型和基因型数据当作训练群体输入到 Bayesian B 方法中进行训练,为后续预测处理表型性状做好基础。

(4)构建 Bayesian B 方法预测模型,接收训练数据开始计算,计算方式调用底层实现统计计算具体实现细节分为 5 步:①构建马尔科夫链使用 MCMC 与 MH(Metropolos-Hasting)抽样对标记效应和

方差进行联合抽样[11-12];将初始设定先验值作为评定值去掉不稳定模型区域数据的阈值。②计算先验分布并去掉不稳定区域;训练每个匹配个体并计算效应值。③计算每个标记效应值;进行迭代完成训练群体每个个体的效应值计算。④判定预测效果并迭代计算每个个体;将所有效应值整合累积,并完成关联处理。⑤算得 Bayesian B 预测值。

(5)运用算法模型得到训练群体的预测值,计算没有匹配的个体表型性状的基因标记效应。

(6)将每个个体基因标记效应进行累加算出个体表型性状值。

(7)算得输入表型所有性状预测值,再计算综合育种值,服务于服务育种选择。

2.2.1　时间和效率对比　大豆(307 个体,7 125个 marks)某性状是在同一运算环境下进行时间和效率的对比,利用不同先验值进行测试,结果如表 3。直观展示先验值越大,处理时间就越长。无论先验值多大,所得到的 Accuracy 都是在一定范围内,而且有所提高,由此可见,先验值的大小,不仅决定了运算时间的长短,还对准确度的稳定性有很大的影响。

表 3　在贝叶斯 B 方法中取不同先验值进行精准值的运算所用时间和精准值的比较

Table 3　Comparison of the time and Accuracy in the computation of the accurate values of different prior values in the Bayesian B method

大豆(个体,标记) Soybean(individuals,marks)	时间和精准值 Time and accuracy	先验值 500 Prion value 500	先验值 5000 Prion value 5000	先验值 20000 Prion value 20000
大豆(307,7125)	Time(second)	65342.0	698756.8	27128768.8
Soybean(307,7125)	Accuracy	0.42816	0.46629	0.47638

2.2.2　育种值计算　通过 Bayesian B 方法对于大豆数据进行基因组选择,算得不同性状育种值如表 4。根据这些育种值,育种家可以采用加权方式或给定计算公式算取综合育种值,指导选种选育[13]。

表 4　Bayesian B 方法算取大豆性状育种值

Table 4　Bayesian B method for calculating the breeding value of soybean traits

序号 Code	个体 Individual	性状 1 育种值 Breeding value of trait 1	性状 2 育种值 Breeding value of trait 2	性状 3 育种值 Breeding value of trait 3	性状 4 育种值 Breeding value of trait 4
1	CC1	-1.144785753	-0.075067388	-0.075067388	-5.450764748
2	CC10	-2.169343332	-0.081929817	-0.081929817	-4.178595990
3	CC100	-0.177602911	0.049436254	0.049436254	-0.306000505
4	CC101	-2.972702508	-0.093927981	-0.093927981	-7.486184375
5	CC102	1.499222645	0.001017684	0.001017684	0.815709378
6	CC103	-0.651772411	-0.014369962	-0.014369962	-1.140591412
7	CC104	4.415172291	-0.084313487	-0.084313487	4.443670588
8	CC105	-0.340542284	-0.042535800	-0.042535800	1.316693580
9	CC106	-2.624320435	-0.081575082	-0.081575082	-5.289412543
10	CC107	1.768449444	-0.028466986	-0.028466986	4.271822384

3　讨　论

全基因组选择是基因组育种的关键环节,在整个过程中,测得表型性状个体数常少于标记数,因此要进行的预测就异常艰难,通过以上介绍的几种方法,对真实数据和模拟数据预测结果显示:Bayesian B 方法对估计大豆育种值准确性很高,但实际应用 Bayesian B 方法并不十分普遍,主要原因是运行时间相对于其它方法较长,从统计模型计算上来看,Bayesian B 方法计算时间长于其它方法,所以在构建马尔可夫链时要根据标记数量来设定马尔科夫链和抽样数量,还需要改进方法加入判断参数设置的条件,进而避免人为设置参数影响精度和时间。对于运行时间长的问题还可选择高性能服务器,并将 GPU 计算引入,同时改进程序并行计算,通过提高资源利用率方式来解决。

4　结　论

通过对比几种基因组选择方法的优缺点,采用两组大豆基因型数据进行测试,在真实性状和模拟性状下采用交叉验证计算精准值进行对比,Bayesian B 方法效果较好。由于该方法参数设置在计算时会直接影响计算时间和精准值,因此根据基因型数据大小在采用该方法预测时可根据实际需

求进行设定,使其运行时间较短,准确性较高。另外,随着人们对于基因研究的进一步深入,能够为 Bayesian B 方法提供更多准确辅助信息,从而使得 Bayesian 方法的优势发挥到极致,由此可见该方法在大豆全基因组选择将会得到广泛应用。

参考文献

[1] 孙恺,孙健,舒小丽,等. 高异黄酮大豆突变体的筛选及其特性初步研究[J]. 核农学报,2016,30(11):2088-2095. (Sun K, Sun J, Shu X L, et al. Screening and characterization of a high isoflavone soybean mutant[J]. Journal of Nuclear Agriculture, 2016, 30 (11):2088-2095.)

[2] 薛永国,魏崠,唐晓飞,等. 黑龙江省育成大豆品种性状演变的分析[J]. 大豆科学,2015,34(3):361-366. (Xue Y G, W L, Tang X F, et al. Analysis of the character evolution of soybean cultivars bred in Heilongjiang province [J]. Soybean Science, 2015, 34(3): 361-366.)

[3] 穰中文,周清明. 水稻 Dirigent 基因家族生物信息学分析[J]. 湖南农业大学学报(自然科学版),2013,39(2):111-120. (Guo Q, Zhou Q M. Bioinformatics analysis of rice DIRIGENT gene family [J]. Journal of Hunan Agricultural University (Natural Science Edition), 2013,39(2): 111-120.)

[4] 周永正. 一般混合线性模型固定效应、随机效应与另一随机向量的联合估计[J]. 数学的实践与认识,2011(19):58-64. (Zhou Y Z. Joint estimation of fixed effects, random effects and another random vector in general mixed linear models[J]. Practice and Cognition of Mathematics,2011(19): 58-64.)

[5] Habier D, Fernando R L, Garrick D J. Genomic BLUP decoded: A look into the black box of genomic prediction [J]. Genetics, 2013,194:597-607.

[6] Wang C L, Qin Z, Li J, et al. Comparative study of estimation methods for genomic breeding values[J]. Science Bulletin,2016, 61(5):353-356.

[7] 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.

[8] Li B, Yu L X, Xi J J, et al. A widely applicable method for plant genomic DNA extraction[J]. Agricultural Biotechnology,2012,12 (1):143-186.

[9] 王欣,杨泽峰,徐辰武. 基于育种值预测的基因组选择方法的比较[J]. Science Bulletin,2015,60(10):925-935,983. (Wang X, Yang Z F, Xu C W. Comparison of genomic selection methods based on prediction of breeding value [J]. Science Bulletin, 2015,60 (10): 925-935,983.)

[10] Tang Y, Liu X L, Wang J B, et al. GAPIT Version 2: An enhanced integrated tool for genomic association and prediction[J]. Plant Genome,2016,9(2):1-9.

[11] Anna W, Jesus A, Petek S, et al. Garrick. Mixture models detect large effect QTL better than GBLUP and result in more accurate and persistent predictions[J]. Journal of Animal Science and Biotechnology,2016,7(4):468-473.

[12] Pan X P, Zhang G Z, Zhang J J, et al. Zoeppritz-based AVO inversion using an improved Markov chain Monte Carlo method[J]. Petroleum Science,2017,14(1):75-83.

[13] 王艳,韩英鹏,李文滨. 大豆分子标记研究新进展[J]. 大豆科学,2015,34(1):148-154,162. (Wang Y, Han Y P, Li W B. New progress in the research of soybean molecular markers [J]. Soybean Science, 2015,34(1):148-154,162.)