Bayesian methods for jointly estimating genomic breeding values of one continuous and one threshold trait

Chonglong Wang[1☯]**, Xiujin Li**[2,3,4☯]**, Rong Qian**[1]**, Guosheng Su**[3]**, Qin Zhang**[2]**, Xiangdong Ding**[2]*****

1 Department of Pig Genetics and Breeding, Institute of Animal Husbandry and Veterinary Medicine, Anhui Academy of Agricultural Sciences, Hefei, China, **2** Laboratory of Animal Genetics, Breeding and Reproduction, Ministry of Agriculture of China, National Engineering Laboratory for Animal Breeding, College of Animal Science and Technology, China Agricultural University, Beijing, China, **3** Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Tjele, Denmark, **4** State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, North Third Road, Guangzhou Higher Education Mega Center, Guangzhou, Guangdong, PR China

☯ These authors contributed equally to this work.
* xding@cau.edu.cn

Abstract

Genomic selection has become a useful tool for animal and plant breeding. Currently, genomic evaluation is usually carried out using a single-trait model. However, a multi-trait model has the advantage of using information on the correlated traits, leading to more accurate genomic prediction. To date, joint genomic prediction for a continuous and a threshold trait using a multi-trait model is scarce and needs more attention. Based on the previously proposed methods BayesCπ for single continuous trait and BayesTCπ for single threshold trait, we developed a novel method based on a linear-threshold model, i.e., LT-BayesCπ, for joint genomic prediction of a continuous trait and a threshold trait. Computing procedures of LT-BayesCπ using Markov Chain Monte Carlo algorithm were derived. A simulation study was performed to investigate the advantages of LT-BayesCπ over BayesCπ and BayesTCπ with regard to the accuracy of genomic prediction on both traits. Factors affecting the performance of LT-BayesCπ were addressed. The results showed that, in all scenarios, the accuracy of genomic prediction obtained from LT-BayesCπ was significantly increased for the threshold trait compared to that from single trait prediction using BayesTCπ, while the accuracy for the continuous trait was comparable with that from single trait prediction using BayesCπ. The proposed LT-BayesCπ could be a method of choice for joint genomic prediction of one continuous and one threshold trait.

Introduction

With the developments of single-nucleotide polymorphism (SNP) chips and genotyping by sequencing, a huge number of genome-wide polymorphisms have been widely used in practical animal and plant breeding programs. Genomic selection (GS) can use the information of genome-wide markers to accurately predict the genetic merit of an animal without the need of

its own phenotypic information[1]. In a typical process of genomic prediction, SNP effects are estimated using a training population consisting of individuals with both SNP genotypes and phenotypes, and then these estimated effects are used to build a prediction equation to calculate the genomic estimated breeding values (GEBVs) for breeding candidates, based on their SNP genotypes. Therefore, an appropriate model is a key to accurately predict GEBVs in GS.

Many Bayesian models have been proposed to estimate GEBVs. In the first paper of genomic selection[1], two Bayesian methods (BayesA and BayesB) were presented to estimate genomic breeding values and have been extensively used in the subsequent studies of genomic selection[2–5]. However, there are two drawbacks of BayesA and BayesB. One drawback is that the full-conditional posterior distribution of a locus-specific variance has only one additional degree of freedom compared to its prior distribution regardless of the number of genotypes or phenotypes, and the shrinkage of SNP effects depends strongly on the scale parameter S_a^2 as pointed out by Gianola et al.[6]. The other one is that for BayesB the possibility π of a SNP having zero effect should be given as known. To overcome these two drawbacks, Habier et al.[7] proposed two new methods, BayesCπ and BayesDπ. For BayesCπ, a single variance is common to SNPs having non-zero effects instead of locus-specific variances, while for BayesDπ, the scale parameter, S_a^2, for the scaled inverse chi-square prior of the locus-specific variance is treated as an unknown with a Gamma(1,1) prior. Both BayesCπ and BayesDπ treat π as an unknown, and need to be inferred from the data.

Wang et al.[8] extended the three Bayesian methods (BayesA, BayesB, and BayesCπ) for a Gaussian trait to be used for a threshold trait. The extended methods are correspondingly termed BayesTA, BayesTB and BayesTCπ, respectively. From the results of a simulation study on a threshold trait, the three BayesT methods performed better than the corresponding Bayesian methods when treating a threshold trait as a Gaussian trait, and BayesTCπ performed the best among the three new methods, so it was recommended as the method of choice for threshold traits in genomic selection[8]. In addition, a Bayesian multi-locus association mapping for threshold traits using a threshold model was also proposed by Iwata et al.[9], and their approach could reduce both false-positive and false-negative rates in detecting QTL to reasonable levels.

Currently, a single trait model is usually applied in practical genomic prediction. Theoretically, a joint genomic evaluation of genetically correlated traits should lead to more accurate predictions than single trait genomic prediction as shown in some studies[10–14]. However, these studies all focused on multiple continuous traits, and studies on a joint genomic prediction of continuous and discrete traits are scarce so far. In practical animal breeding, people often need to make selection for continuous and categorical traits simultaneously (e.g., birth weight and calving ease in cattle and growth rate and leg weakness in pig). In conventional breeding value prediction based on phenotypic and pedigree information, methods for joint analysis of continuous and threshold traits have been well established[15–18]. In this study, we proposed a BayesCπ bivariate model for a joint genomic prediction of one normal distributed trait and one threshold trait, which was termed as LT-BayesCπ. We validated LT-BayesCπ with simulated data and with a common data set from the 14th QTL-MAS workshop[19]. The accuracy of genomic prediction obtained from LT-BayesCπ was compared with that from BayesCπ or BayesTCπ based on single-trait model. Furthermore, factors affecting the performance of LT-BayesCπ were investigated as well.

Methods

Models

Let $\mathbf{y}_1' = \{y_{1,i}\}$ ($i = 1, 2, \ldots, n$) be the vector of observations for a continuous trait, $\mathbf{y}_2' = \{y_{2,i}\}$ ($i = 1, 2, \ldots, n$) be the vector of observations for a threshold trait and $\mathbf{l}' = \{l_i\}$ ($i = 1, 2, \ldots, n$) be

the vector of underlying latent variables or liabilities associated with the threshold trait. The linear-threshold model is

$$
\begin{bmatrix} \mathbf{y_1} \\ \mathbf{l} \end{bmatrix} = \begin{bmatrix} \mathbf{X_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{X_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{bmatrix} + \begin{bmatrix} \mathbf{Z} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z} \end{bmatrix} \begin{bmatrix} \mathbf{g_1} \\ \mathbf{g_2} \end{bmatrix} + \begin{bmatrix} \mathbf{e_1} \\ \mathbf{e_2} \end{bmatrix},
$$

where $\boldsymbol{\beta_{1(2)}}$ is a vector of fixed effects, $\mathbf{g_{1(2)}}$ is the vector of SNP effects, $\mathbf{e_{1(2)}}$ is the vector of the random residuals, $\mathbf{X_{1(2)}}$ is the incidence matrix for $\boldsymbol{\beta_{1(2)}}$, and $\mathbf{Z}$ is the matrix of genotype indicators with values 0,1 or 2 for genotypes 11, 12 and 22, respectively. Let $\mathbf{v'} = [\mathbf{y'_1}, \mathbf{l'}]$. It is assumed that, given $\boldsymbol{\beta}$ and $\mathbf{g}$, $\mathbf{v}$ is distributed as

$$
\mathbf{v} | \boldsymbol{\beta}, \mathbf{g}, \boldsymbol{R_e} \sim \mathrm{N}(\begin{bmatrix} \mathbf{X_1}\boldsymbol{\beta_1} + \mathbf{Z}\mathbf{g_1} \\ \mathbf{X_2}\boldsymbol{\beta_2} + \mathbf{Z}\mathbf{g_2} \end{bmatrix}, \mathbf{R}),
$$

where $\boldsymbol{\beta'} = [\boldsymbol{\beta'_1}, \boldsymbol{\beta'_2}]$, $\mathbf{g'} = [\mathbf{g'_1}, \mathbf{g'_2}]$, and $\mathbf{R} = \boldsymbol{R_e} \otimes \mathbf{I}$ with $\boldsymbol{R_e} = \begin{bmatrix} \sigma^2_{e_1} & \sigma_{e_{1,2}} \\ \sigma_{e_{1,2}} & \sigma^2_{e_2} \end{bmatrix}$.

Then, given $\boldsymbol{\beta}$, $\mathbf{g}$ and $\boldsymbol{R_e}$, the sampling model can be written as

$$
\mathrm{p}(\mathbf{v} | \boldsymbol{\beta}, \mathbf{g}, \boldsymbol{R_e}) \propto |\boldsymbol{R_e}|^{-\frac{n}{2}} \exp\left[-\frac{1}{2} \mathrm{tr}\left(\boldsymbol{R_e}^{-1} \boldsymbol{S_e} \right) \right]
$$

where $\boldsymbol{S_e} = \begin{bmatrix} \mathbf{e'_1}\mathbf{e_1} & \mathbf{e'_1}\mathbf{e_2} \\ \mathbf{e'_2}\mathbf{e_1} & \mathbf{e'_2}\mathbf{e_2} \end{bmatrix}$.

MCMC implementation of LT-BayesCπ

Prior distributions. In this study, the following prior distributions are assumed for building a hierarchical model.

For fixed effects $\boldsymbol{\beta}$:

$$
\boldsymbol{\beta} | \beta_{min}, \beta_{max} \sim \cup(\beta_{min}, \beta_{max})
$$

For SNP effects $\mathbf{g}$:

Each SNP has either a zero effect for both traits or non-zero effect for at least one trait with probabilities π and $1-\pi$, respectively. For the latter case, its prior distribution is bivariate normal, i.e.,

$$
\begin{bmatrix} \mathbf{g_1} \\ \mathbf{g_2} \end{bmatrix} | \boldsymbol{G_0} \sim N(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \boldsymbol{G_0} \otimes \mathbf{I}),
$$

where $\boldsymbol{G_0} = \begin{bmatrix} \sigma^2_{g1} & \sigma_{g1,2} \\ \sigma_{g1,2} & \sigma^2_{g2} \end{bmatrix}$.

For $\boldsymbol{G_0}$ and $\boldsymbol{R_e}$

$\boldsymbol{G_0}$ and $\boldsymbol{R_e}$ are assumed to follow inverse Wishart distributions:

$$
\mathrm{p}\left(\boldsymbol{G_0} | v_g, \boldsymbol{V_g} \right) \propto |\boldsymbol{G_0}|^{-\frac{1}{2}(v_g+3)} \exp[-\frac{1}{2} tr\left(\boldsymbol{G_0}^{-1} \boldsymbol{V_g}^{-1} \right)],
$$

$$
\mathrm{p}(\boldsymbol{R_e} |, \boldsymbol{V_e}) \propto |\boldsymbol{R_e}|^{-\frac{1}{2}(v_e+3)} \exp[-\frac{1}{2} tr(\boldsymbol{R_e}^{-1} \boldsymbol{V_e}^{-1})],
$$

where v_i and V_i ($i = g, e$) are the usual hyper-parameters of the scaled inverse Wishart distribution, which are assumed to be known. In particular, if $v_i = -3$ and $V_i = 0$, these two distributions reduce to improper uniform distributions [17].

For the thresholds **t**:

Suppose the threshold trait consists of k categories, then there are k-1 hypothetical thresholds (**t'** = $[t_1, t_2, \ldots, t_{k-1}]$) in the underling latent scale. These thresholds are assumed to follow ordered uniform distribution in the interval $[t_{min}, t_{max}]$. However, two of the thresholds must be fixed, so as to ensure the identifiability of the parameters. Typical choices are $t_1 = 0$ and $t_2 = 1$[17]. Therefore, there are k-3 unknown thresholds. The joint prior density of **t** is

$$p(\mathbf{t}) = (k-3)! \left(\frac{1}{t_{max} - t_{min}} \right)^{k-3} I(\mathbf{t} \in T),$$

where $T = \{(t_1 = 0, t_2 = 1, t_2, \ldots, t_{k-1}) | t_{min} \le t_1 \le \cdots \le t_{k-1} \le t_{max}\}$, I is an indicator vector. If $t \in T$, elements of I are 1, otherwise, elements of I are 0. In this study, we set t_{max} and t_{min} as $\mu \pm 10\sigma$ based on the normal distribution of liability.

For the probability of zero effect π:

$$p(\pi) \sim \cup(0, 1).$$

Joint posterior distribution. The parameter vector is augmented with the unobserved liabilities **l** for the threshold trait and with the indicator variables **δ** for SNP effects (indicating whether a SNP has an effect on the traits (with probability of 1 -π) or not (with probability of π)), and is represented as $(\Omega, \mathbf{l}, \boldsymbol{\delta})$, where $\Omega = (\boldsymbol{\beta}, \mathbf{g}, \mathbf{G_0}, \mathbf{R_e}, \mathbf{t}, \boldsymbol{\pi})$.

The joint posterior distribution of $(\Omega, \mathbf{l}, \boldsymbol{\delta})$ is

$$\mathbf{p}(\boldsymbol{\Omega}, \mathbf{l}, \boldsymbol{\delta} | \boldsymbol{y}_1, \boldsymbol{y}_2, \sigma_{e2}^2 = 1, t_1 = 0) \propto \mathbf{p}(\boldsymbol{y}_1, \boldsymbol{y}_2 | \boldsymbol{\Omega}, \mathbf{l}, \boldsymbol{\delta}) \mathbf{p}(\boldsymbol{\Omega}, \mathbf{l}, \boldsymbol{\delta}) = \mathbf{p}(\boldsymbol{y}_1, \mathbf{l} | \boldsymbol{\Omega}) \mathbf{p}(\boldsymbol{y}_2 | \boldsymbol{\Omega}, \mathbf{l}) \mathbf{p}(\boldsymbol{\Omega}, \boldsymbol{\delta})$$

Fully conditional posterior distributions. **Liabilities.** The fully conditional posterior distribution of liability l_i is a truncated normal distribution within the range from t_{j-1} to t_j, i.e, $l_i|ELSE \sim N(E(l_i|ELSE), Var(l_i|ELSE))$ and truncated within t_{j-1} to t_j with

$$E(l_i|ELSE) = \boldsymbol{x}_{2,i}' \boldsymbol{\beta}_2 + \boldsymbol{z}_{2,i}' \boldsymbol{g}_2 + \frac{\sigma_{e1,2}}{\sigma_{e1}^2} \left(y_{1,i} - \boldsymbol{x}_{1,i}' \boldsymbol{\beta}_1 - \boldsymbol{z}_{1,i}' \boldsymbol{g}_1 \right)$$

and

$$Var(l_i|ELSE) = \sigma_{e2}^2 \left(1 - \frac{(\sigma_{e1,2})^2}{\sigma_{e1}^2 \sigma_{e2}^2} \right)$$

Location parameters. Write the mixed model equations (MME) as $\mathbf{C}\hat{\boldsymbol{\theta}} = \boldsymbol{r}$, where **C** is left hand, $\boldsymbol{r}$ is right hand, and $\hat{\boldsymbol{\theta}}$ is a vector of unknown parameters, $\hat{\boldsymbol{\theta}} = \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{g} \end{bmatrix}$. The fully conditional posterior distribution of θ_i is

$$\theta_i | ELSE \sim N(\tilde{\theta}_i, C_{ii}^{-1}),$$

where $\tilde{\theta}_i = C_{i,i}^{-1}(r_i - \boldsymbol{C}_{i,-i}\boldsymbol{\theta}_{-i})$.

Dispersion parameters. The fully conditional posterior distributions of the co-variance matrix of SNP effects and residual effects are

$$\boldsymbol{G}_0 | ELSE \sim IW_2((\boldsymbol{V}_g^{-1} + \boldsymbol{S}_g)^{-1}, \nu_g + q)$$

$$\boldsymbol{R}_e | ELSE \sim IW_2((\boldsymbol{V}_e^{-1} + \boldsymbol{S}_e)^{-1}, \nu_e + n)$$

If the prior distributions of $\mathbf{G}_0$ and $\mathbf{R}_e$ are flat (in case of $\nu_i = $ -3 and $\boldsymbol{V_i} = \mathbf{0}$, $i = g, e$), these two distributions reduce to

$$\boldsymbol{G}_0 | ELSE \sim IW_2(\boldsymbol{S}_g^{-1}, q - 3)$$

$$\boldsymbol{R}_e | ELSE \sim IW_2(\boldsymbol{S}_e^{-1}, n - 3)$$

The fully posterior distributions of other parameters and the procedure of the Gibbs sampler are similar to single trait BayesTCπ which was described by Wang et al.[8].

It should be noted that when the categorical trait is binary, i.e., it has only two categories with one threshold, it is not possible to fix two thresholds. In this case, a usual parameterization is to fix the residual variance of the binary trait ($\sigma_{e_2}^2$) to be 1 and the threshold to be 0. With this parameterization, rather than adopting an inverse Wishart prior for $\boldsymbol{R}_e$, one can assign a conditional inverse Wishart prior (conditional on $\sigma_{e_2}^2 = 1$) and the fully conditional posterior distribution of $\boldsymbol{R}_e$ is conditional scaled inverted Wishart given $\sigma_{e_2}^2 = 1$ [17]. A general algorithm for drawing samples from such a distribution was proposed by Korsgaard et al. (1999) [20] (see also [17]). Following their algorithm, the realized values from the distribution $R_e | ELSE, \sigma_{e2}^2 = 1$ can be obtained in the following way:

Let $V = S_e^{-1} = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$

Sample x_1 from $W_1(V_{11}, n-3)$.

Sample x_2 from $N(V_{11}^{-1}V_{12}, x_1^{-1}V_{22.1})$, where $V_{22.1} = V_{22} - V_{12}^2 V_{11}^{-1}$.

Let $T_{11} = x_1^{-1} + x_2^2$, $T_{12} = -x_2$, and $T_{22} = 1$, then T_{11}, T_{12} and T_{22} are the realized values from the distribution $R_e | ELSE, \sigma_{e2}^2 = 1$.

Simulation study

Data simulation. To evaluate the proposed method LT-BayesCπ, we carried out a series of simulations using the multiple-trait genomic simulation software GPOPSIM [21]. For simplicity, we simulated a continuous and a binary threshold trait with different genetic correlation levels between them.

Briefly, the simulation started with a base population consisting of 100 individuals, followed by 1,000 non-overlapping historical generations with the same population size in each generation, denoted as generation -999 to generation 0. In each historical generation, 50 males were randomly mated with 50 females and each mating produced one male and one female offspring. All markers were monomorphic at the starting status in the base population, then polymorphic markers were generated in the following generations by mutation with a mutation rate of 1.25×10^{-3}, and reached the mutation-drift equilibrium status through genetic drift. After 1,000 historical generations, six generations, numbered from 1 to 6, were further generated. In

generation 1, the population size was expanded from 100 to 1,000 by increasing the number of offspring of each female in generation 0 from 2 to 20 (10 males and 10 females). From generation 1 to 5, 50 males were randomly selected from the 500 males to be sires of the next generation and all 500 females were used as dams. Each selected male mated randomly with 10 females and each female produced two offspring (one male and one female). Generations 1 and 2 were treated as training population across the scenarios studied, and generations 3–6 were validation (candidate) populations.

We simulated five chromosomes with a total length of 5 Morgan (1 Morgan per chromosome). On each chromosome, 2,000 markers were evenly distributed and every two adjacent loci were assumed to harbor a potential QTL. The final true QTL were randomly sampled from these potential QTL. Based on the distance between two adjacent loci, Haldane's mapping function was used to calculate the probability of recombination between adjacent loci.

Two genetically correlated traits (denoted as Trait A and Trait B) were simulated. Trait A was a continuous trait with a normal distribution, and Trait B was a binary threshold trait with normally distributed underlying liabilities. The sampled true QTL were divided into three groups, Group1, Group2 and Group3. QTL in Group1 had pleiotropic effects on both traits, and QTL in Group2 and Group3 had effects on trait A only or trait B only, respectively. The allele substitution effects of each QTL in Group1 were sampled from a bivariate normal distribution with varied genetic correlation (r_{AB}) between the two traits and those in Group2 or Group3 were drawn from univariate normal distributions, see details in[14]. For any r_{AB} between traits A and B, we set the ratio of QTL in Group1, Group2 and Group3 as 0.8:0.1:0.1, except for the case of $r_{AB} = 0$, where the ratio of QTL in Group1, Group2 and Group3 was set as 0.0:0.5:0.5. The allele substitution effects were re-scaled to ensure that the total additive genetic variances of trait A and B were equal to 2.0 and 1.0, respectively. The environmental correlation was assumed to be 0.0, and the environmental effects on the two traits were sampled independently from univariate normal distributions.

True breeding values (TBV), which were generated by summing effects of all QTL, were added to environmental effects to produce phenotypic values of trait A and liability values of trait B. Genotypes and TBV were simulated for all individuals from generations 1 to 6, but phenotypic or liability values were only assigned to the 2,000 individuals in generations 1 and 2 (training population). For trait B, a threshold value was assigned according to the assumed incidence, and the observed phenotype of an individual with liability value lower than the threshold value was 0, otherwise it was 1.

We firstly simulated a standard scenario in which the following parameters were assigned: heritabilities of the two traits: $h_A^2 = 0.3$ and $h_B^2 = 0.1$, number of QTL = 60, genetic correlation between traits A and B: $r_{AB} = 0.50$, and the incidence (individuals with observation 1) for trait B = 0.30. To investigate the impacts of various factors on genomic prediction, alternative scenarios were generated by using one of the following parameters to replace the corresponding parameter in the standard scenario. These alternative parameters were: genetic correlation r_{AB} (0.00, 0.20, and 0.80), number of QTL (20, 200, and 500), heritability of the continuous trait A ($h_A^2 = 0.5$ and 0.8), heritability of the binary threshold trait B ($h_A^2 = 0.3$ and 0.5), and incidences of trait B (0.05, 0.1, and 0.5). For each scenario, 20 replicated datasets were simulated.

Data from the 14th QTL-MAS workshop. The common data set of the 14th QTL-MAS workshop[19] was also used to evaluate the proposed method LT-BayesCπ. This data set contains 3,226 individuals in five generations (F0-F4). All individuals have genotypes, and only 2,326 individuals in generations F0-F3 have phenotypic records on two traits: a quantitative trait Q and a binary threshold trait B. Individuals with phenotypic records (F0-F3) were treated as training population and those without phenotypic records (F4) as validation

(candidate) population. Five autosomal chromosomes were simulated, each 100Mbp long, and contained 10,031 biallelic SNP without any missing data and genotype errors. The quantitative trait Q was controlled by 37 QTL (30 additive QTL, 4 epistatic QTL and 3 imprinted QTL). Out of the 30 additive QTL, 22 also influenced the binary threshold trait B which was not controlled by any other QTL. The narrow-sense heritability (h^2) for trait Q was 0.52 for males and 0.39 for females, whereas h^2 for trait B was 0.48 for both males and females. The correlation between breeding values for the two traits was 0.59 for males and 0.68 for females.

Estimation of SNP effects

Three Bayesian methods were implemented to estimate SNP effects based on the training population. The proposed new method LT-BayesCπ was used for the joint analysis of both traits, while BayesCπ was used for the continuous trait and BayesTCπ for the threshold trait. For each model, the Markov chains were run for 50, 000 cycles of Gibbs sampling, and the first 30, 000 cycles were discarded as burn-in. All remaining samples of SNP effects after burn-in were averaged to obtain the estimates of SNP effects.

In the analysis of both simulated data sets by LT-BayesCπ, we assumed the values of the hyper-parameters v_g, $\mathbf{V}_g$, v_e and $\mathbf{V}_e$ of the prior distributions of $\mathbf{G}_0$ and $\mathbf{R}_e$ to be -3, $\mathbf{0}$, -3, and $\mathbf{0}$, respectively, such that they reduced to flat priors. In addition, in both simulated data sets, the threshold trait was binary. So, as mentioned above, we fixed the threshold to be 0 and the residual variance of the binary trait to be 1 in the analysis. We drew samples of the fully conditional posterior distribution of $\mathbf{R}_e$ using the algorithm mentioned above.

Accuracy of genomic prediction

GEBVs for individuals in the candidate population were calculated as the sum of all marker effects according to their marker genotypes. For each trait, accuracy was measured as the correlation between TBV and GEBV ($r_{TBV,GEBV}$), and the regression of TBV on GEBV ($b_{TBV,GEBV}$) was also calculated for assessing the bias of genomic prediction. However, for the binary trait, the scale of GEBV was not the same as TBV due to the restriction of $\sigma_{e2}^2 = 1$. Thus, $b_{TBV,GEBV}$ must be rescaled back to the original scale by using $b_{TBV,GEBV}/\sqrt{v_{e2}}$, where v_{e2} is the true residual variance of the liabilities in the simulation. In addition, a t-test was carried out to investigate the differences in accuracy obtained from LT-BayesCπ and the single-trait methods BayesCπ or BayesTCπ.

Results

Analysis of simulated data

Estimates of SNP effects in the standard scenario. Fig 1 shows the simulated QTL effects and the estimated SNP effects by LT-BayesCπ, BayesCπ and BayesTCπ from a randomly selected replicate in the standard scenario. For the continuous trait (Trait A), the simulated absolute SNP effects ranged from 0–0.75, and the estimated absolute SNP effects ranged from 0–0.55 from BayesCπ and 0–0.70 from LT-BayesCπ, respetively. For the binary threshold trait (Trait B), the simulated absolute SNP effects ranged from 0–0.77, and the estimated absolute SNP effects ranged from 0–0.14 from BayesTCπ and 0–0.21 from LT-BayesCπ, respectively. Most of the segments containing QTL with large effects were mapped by all methods.

Accuracies of GEBVs in the standard scenario. Table 1 shows accuracies of GEBVs in terms of correlations between GEBVs and simulated true breeding values in generations 3–6 (candidate population) in the standard scenario. For all methods, accuracies of genomic prediction declined with generations as expected; the only exception is that the accuracy for trait
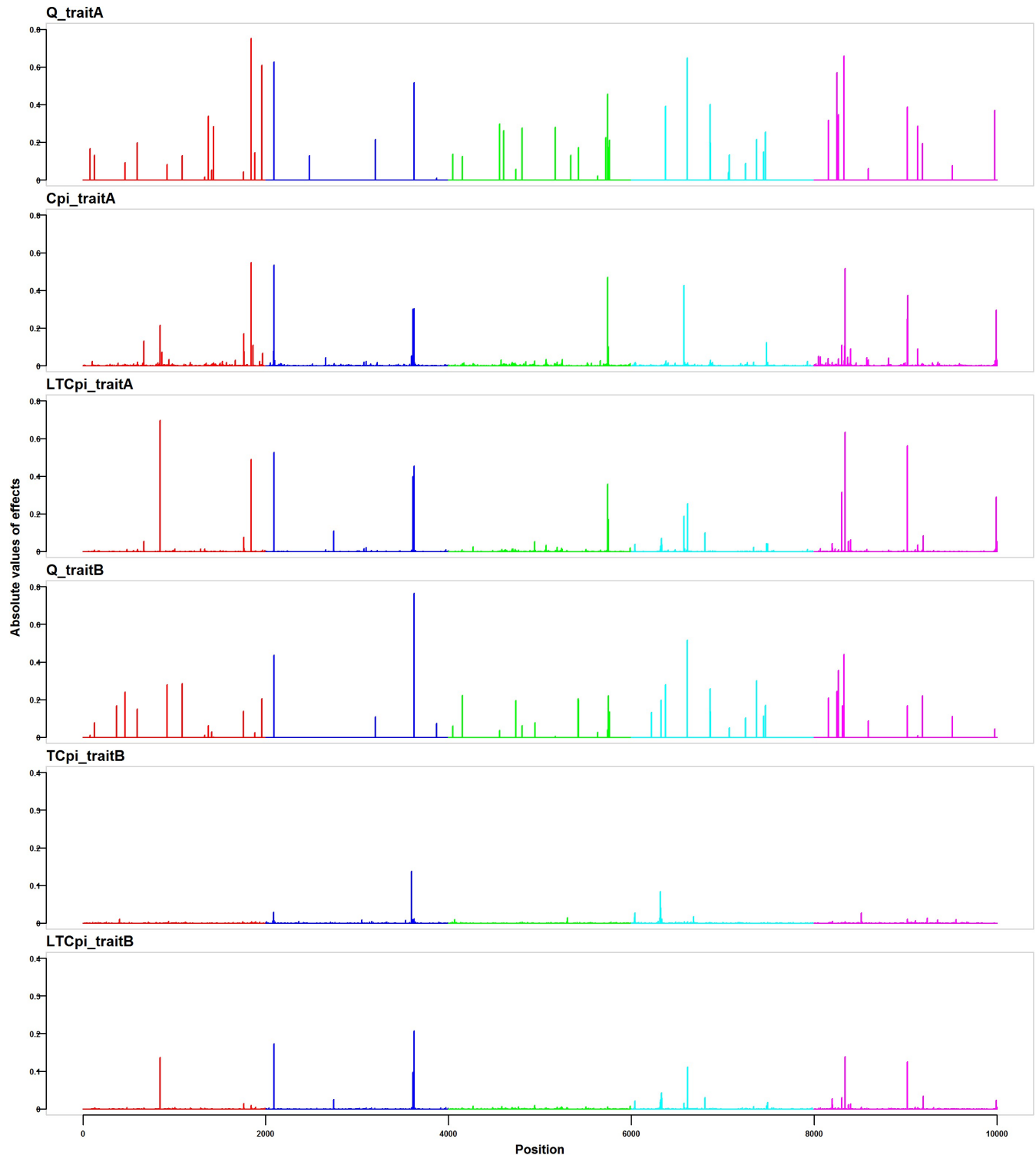
Fig 1. Simulated QTL effects and estimated SNP effects for the continuous trait (trait A) and the binary threshold trait (trait B) from a randomly selected replicate in the standard scenario. Panels Q_traitA and Q_traitB show the absolute values of the simulated true QTL effects. Panels Cpi_traitA, LTCpi_traitA, TCpi_traitB, and LTCpi_traitB show the absolute values of estimated SNP effects by BayesCπ for trait A, LT-BayesCπ for trait A, BayesTCπ for trait B, and LT-BayesCπ for trait B, respectively.

https://doi.org/10.1371/journal.pone.0175448.g001

Table 1. Accuracies of GEBVs (mean±s.e. from 20 replicates) obtained from three methods in generations 3–6 in the standard scenario.

Method	Trait	Generation 3	Generation 4	Generation 5	Generation 6
BayesCπ	A	0.771±0.010	0.740±0.012	0.717±0.013	0.714±0.012
LT-BayesCπ	A	0.761±0.010	0.728±0.011	0.703±0.013	0.699±0.012
BayesTCπ	B	0.465±0.023	0.420±0.022	0.397±0.025	0.395±0.024
LT-BayesCπ	B	0.581±0.020	0.533±0.022	0.510±0.022	0.522±0.023

https://doi.org/10.1371/journal.pone.0175448.t001

B obtained from LT-BayesCπ in generation 5 is slight lower than that in generation 6, probably due to sampling error. For trait B, LT-BayesCπ performed much better than BayesTCπ consistently in all generations, and was about 11 percent higher in each generation (P<0.001). However, for trait A, no improvement was obtained from LT-BayesCπ compared with BayesCπ in all generations (P>0.10).

Impact of genetic correlation. Table 2 shows the accuracies of genomic prediction in generation 3 under different genetic correlations (0, 0.20, 0.50 and 0.80), while keeping the other parameters the same as in the standard scenario. For trait A, the accuracies of GEBV from LT-BayesCπ were nearly equal to those from BayesCπ regardless of the genetic correlation. It suggested that LT-BayesCπ performs comparably with BayesCπ for continuous traits. We also observed that differences in accuracies between LT-BayesCπ and BayesCπ were slightly decreased with increasing the genetic correlation between two traits. For trait B, in the case of no genetic correlation between the two traits, the accuracy of genomic prediction from LT-BayesCπ was slightly lower than that from BayesTCπ. However, the accuracies of genomic prediction obtained from LT-BayesCπ were dramatically increased with the increase of the genetic correlation. The improvement in the accuracy of LT-BayesCπ over BayesTCπ was consistently increased with the increase of the genetic correlation.

Besides the genomic breeding values, we also estimated the genetic and residual correlations between the continuous and the binary traits and the proportions of true QTL (π). As shown in Table 3, the estimates of π obtained by LT-BayesCπ were very close to the assigned

Table 2. Accuracies of GEBVs for the two traits in generation 3 in four scenarios of different genetic correlations.

Genetic correlation	Method	Accuracy	
		Trait A	**Trait B**
0.00	BayesCπ/TCπ	0.817±0.008	0.545±0.024
	LT-BayesCπ	0.800±0.010	0.492±0.026
	Increment	-0.017	-0.053
0.20	BayesCπ/TCπ	0.780±0.010	0.479±0.019
	LT-BayesCπ	0.770±0.010	0.525±0.017
	Increment	-0.010	0.046
0.50	BayesCπ/TCπ	0.771±0.010	0.465±0.023
	LT-BayesCπ	0.761±0.010	0.581±0.020
	Increment	-0.010	0.116[***]
0.80	BayesCπ/TCπ	0.762±0.008	0.473±0.019
	LT-BayesCπ	0.756±0.008	0.674±0.016
	Increment	-0.006	0.201[***]

*** P-value < 0.001

** P-value < 0.01

*P-value < 0.05

https://doi.org/10.1371/journal.pone.0175448.t002

Table 3. The estimated genetic correlations ($\widehat{r}_g$), residual correlations ($\widehat{r}_e$), and proportions of true QTL ($\widehat{\pi}$) from LT-BayesCπ in four scenarios of different genetic correlations.

Genetic correlation	$\widehat{r}_g$	$\widehat{r}_e$	$\widehat{\pi}$
0.00	0.026±0.030	-0.011±0.007	0.0051±0.0004
0.20	0.178±0.036	-0.006±0.007	0.0068±0.0003
0.50	0.471±0.036	0.005±0.008	0.0061±0.005
0.80	0.674±0.024	0.020±0.006	0.0060±0.0003

The assigned r_e and π are 0 and 0.006, respectively.

0.006 (60 QTL) in all scenarios with different genetic correlations. Meanwhile, the estimates of genetic and residual correlations were almost unbiased in all cases, except in the case of genetic correlation of 0.80, where the estimate of genetic correlation was biased downwards.

The regression coefficients of the simulated TBV on GEBV are presented in Table 4. For trait A, the regression coefficients from BayesCπ were all slightly lower than 1, while those from LT-BayesCπ were slightly greater than 1. For trait B, because the scale of GEBV is not the same as TBV due to the restriction of $\sigma_{e2}^2 = 1$, the regression coefficients must be rescaled back to original scale. After rescaling, both LT-BayesCπ and BayesTCπ generated nearly unbiased genomic prediction, i.e., the regression coefficients were closer to 1.0 in all situations except for LT-BayesCπ in the scenario of genetic correlation = 0.

Impact of number of QTL. As shown in Fig 2, BayesCπ, BayesTCπ and LT-BayesCπ all were sensitive to the number of QTL affecting traits of interest, and the accuracies of genomic prediction from them decreased rapidly with the increase of the number of QTL. When the number of QTL increased from 20 to 500, the accuracies of GEBVs were decreased by 0.141, 0.131, 0.156 and 0.169 for BayesCπ, BayesTCπ, LT-BayesCπ (Trait A) and LT-BayesCπ (Trait B), respectively. In the same scenario, LT-BayesCπ was equivalent to single trait method BayesCπ for the continuous trait A and performed better than BayesTCπ for the threshold trait B. For trait B, the accuracies from LT-BayesCπ were 0.113, 0.116, 0.095, 0.075 higher (P <0.001) than those from BayesTCπ in cases with 20, 60, 200 and 500 QTL, respectively.

Impact of heritability. Fig 3 shows the accuracies of GEBVs from the three methods in generation 3 under different heritabilities of one trait, while keeping the other parameters the same as in the standard scenario. By increasing the heritability of trait A from 0.3 to 0.8 and keeping the heritability of trait B unchanged, the accuracies of GEBVs for trait A from BayesCπ

Table 4. Regression coefficients of TBVs on GEBVs in generation 3 in four scenarios with different genetic correlations.

| Genetic correlation	Method	Regression coefficient	
		Trait A	Trait B*
0.00	BayesCπ/TCπ	0.996±0.018	0.962±0.062
	LT-BayesCπ	1.169±0.023	0.787±0.046
0.20	BayesCπ/TCπ	0.988±0.011	1.176±0.168
	LT-BayesCπ	1.153±0.014	0.874±0.047
0.50	BayesCπ/TCπ	0.982±0.017	1.093±0.156
	LT-BayesCπ	1.134±0.022	0.864±0.035
0.80	BayesCπ/TCπ	0.976±0.016	1.124±0.152
	LT-BayesCπ	1.140±0.023	0.918±0.038

* Rescaled regression coefficients of TBVs on GEBVs
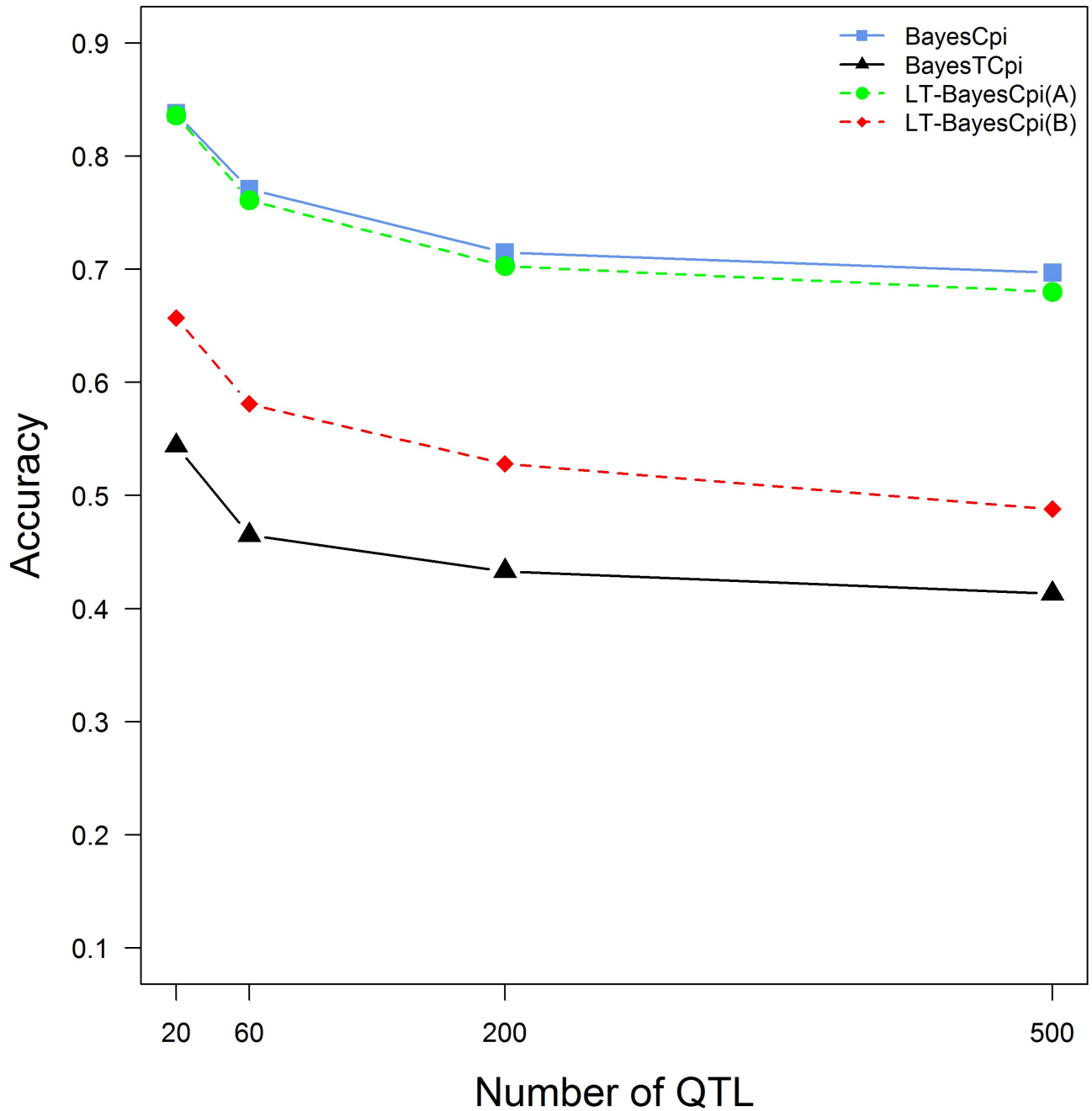
Fig 2. Accuracies of GEBVs from three methods in generation 3 when the number of simulated true QTL changed from 20 to 500.

https://doi.org/10.1371/journal.pone.0175448.g002

and LT-BayesCπ consistently increased as expected. In addition, for trait B, the accuracies from LT-BayesCπ also increased from 0.581 to 0.632 (the accuracy from BayesTCπ was 0.465). On the other hand, by increasing the heritability of trait B from 0.1 to 0.5 and keeping the heritability of trait A unchanged, the accuracies of GEBVs for trait B from BayesTCπ and LT-BayesCπ consistently increased as expected. In addition, for trait A, the accuracies from LT-BayesCπ also increased from 0.761 to 0.777 (the accuracy from BayesCπ was 0.771). These results indicate that increasing heritability of one trait is helpful to improve the accuracy of
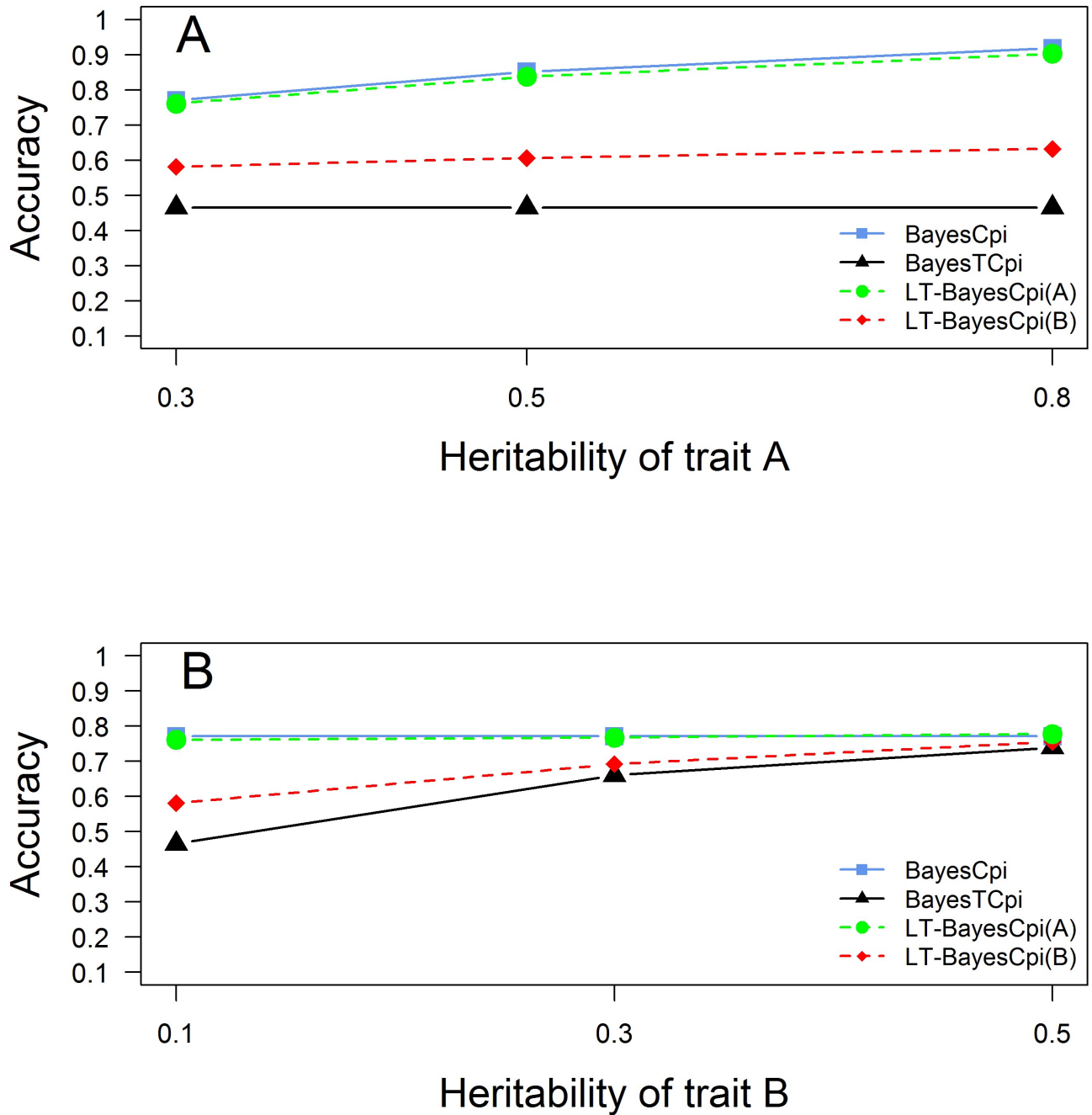
Fig 3. Accuracies of GEBVs from three methods in generation 3 with different heritabilities. A: heritability of the continuous trait A changing from 0.3 to 0.8, while keeping the heritability of the binary threshold trait constant (0.1); **B:** heritability of the binary threshold trait B changing from 0.1 to 0.5, while keeping the heritability of the continuous trait constant (0.3).

genomic prediction of the genetic correlated traits when jointly analyzing them. It should be noted that, for trait A, when the heritability of the continuous trait A was 0.3 and the heritability of the threshold trait B was 0.5, the accuracies from LT-BayesCπ was slightly higher than BayesCπ, while in all other cases, they were slightly lower.

Impact of incidence of the threshold trait. The accuracies of GEBVs for different incidences of the threshold trait B are presented in Fig 4 (the other parameters were the same as in
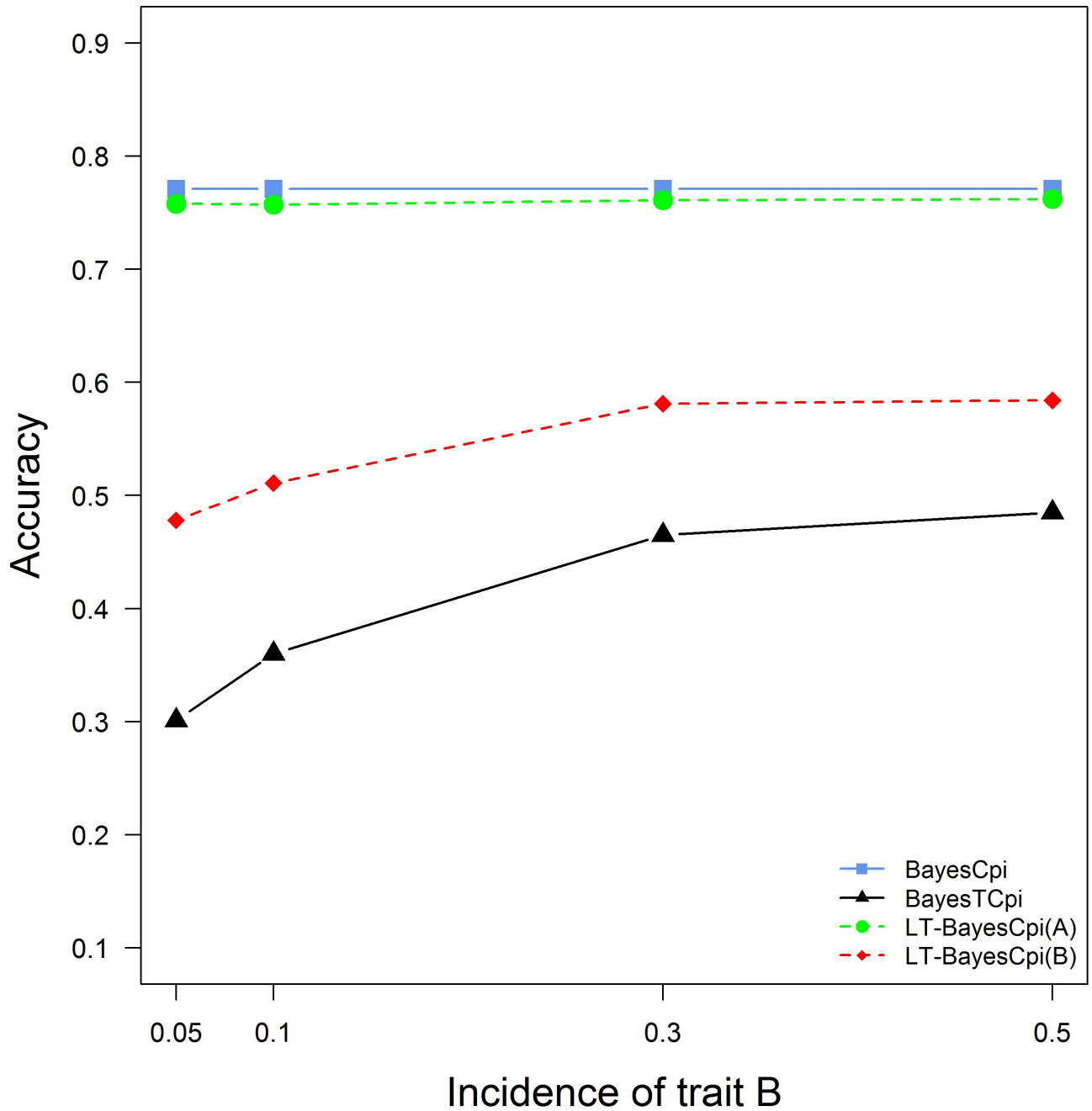
Fig 4. Accuracies of GEBVs from three methods in generation 3 when the incidence of the binary threshold trait increased from 0.05 to 0.5.

the standard scenario). With increased incidence from 5% to 50%, the accuracies of GEBVs for trait B from both BayesTCπ and LT-BayesCπ increased consistently as expected. However, the superiority of LT-BayesCπ over BayesTCπ decreased as the incidence increased. The accuracies of genomic prediction for trait A from LT-BayesCπ were not influenced by the incidence of trait B.

Table 5. Accuracies and bias of GEBVs from three methods for the common dataset from the 14th QTL-MAS workshop.

Trait	Methods	Accuracy	Regression coefficient
Q	BayesCπ	0.677	0.955
	LT-BayesCπ	0.681	0.933
B	BayesTCπ	0.829	1.228[*]
	LT-BayesCπ	0.867	1.055[*]

* Rescaled regression coefficients, i.e., $b_{TBV,GEBV}/\sqrt{V_{e2}}$, where v_{e2} = 18.16 is the true residual variance for the threshold trait B in the simulation.

https://doi.org/10.1371/journal.pone.0175448.t005

Analysis of the common dataset of the 14th QTL-MAS workshop

For each trait and each Bayesian method, the accuracy and bias of GEBVs in the candidate population are shown in Table 5. For the quantitative trait Q, again the accuracy of genomic prediction from LT-BayesCπ was comparable with that from BayesCπ as they did in our simulation study. For the threshold trait B, LT-BayesCπ generated higher accuracy and less bias of GEBV than BayesTCπ, which was consistent with the results from the analysis of the simulated data as well.

Discussion

In traditional genetic evaluation, a multiple-trait model was proved to be able to increase the accuracy of the estimated breeding values by making use of information from genetically correlated traits[22], and has been widely implemented in practical breeding value estimation. Since the concept of genomic selection was proposed in 2001[1], many models, such as GBLUP, BayesA, BayesB, BayesCπ, have been developed for genomic prediction, and most studies focused on genomic prediction for a single continuous trait. A few extensions of BayesA, BayesB and BayesCπ were proposed for a single threshold trait recently [8,23,24]. Recently, some investigations took the correlation structure between traits into account for joint genomic prediction of multiple continuous traits, showing increased accuracy of genomic prediction[10–14]. However, researches on joint genomic prediction of continuous traits and threshold/binary traits are still scarce. In this study, we developed a novel method, LT-BayesCπ, to deal with the joint genomic prediction of one continuous and one threshold trait that are genetically correlated. The results from our simulation study and the common dataset of the 14th QTL-MAS workshop indicated that, in all scenarios considered, when analyzing a continuous and a binary trait jointly using LT-BayesCπ and, both the accuracy and the unbiasedness of GEBV for the binary trait could be remarkably improved in comparison with that from single trait analysis using BayesTCπ, while for the continuous trait the accuracy and the unbiasedness were comparable with that from single trait analysis using BayesCπ.

Genetic correlation between traits is essential for getting benefit from multiple trait analysis. Genetic correlation between two traits arises from pleiotropic effects of common QTL affecting both traits and/or from linkage disequilibrium between QTL affecting different traits. In our simulation study, the genetic correlation between traits resulted mainly from pleiotropic effects of common QTL. The results from the analysis of the simulated data indicated that, the larger the genetic correlation was, the more benefit would be obtained from a joint analysis (Table 2). However, in the case of a zero genetic correlation, the accuracies of joint genomic prediction were lower than that from separate single trait analysis for both traits. This phenomenon was also reported by Jia and Jannink[11] who observed that two continuous trait model performed worse than single trait model if no genetic correlation existed between the two continuous traits. The reason may be that, in such situation, sampling from multiple trait

model leads to nonzero estimates of correlation, and then to erroneous information sharing across traits.

In the simulation study, no benefit was obtained from the joint analysis for the continuous trait in most scenarios considered. One reason is that the threshold trait with low heritability of 0.1 cannot provide enough information to help improving the accuracy of the continuous trait. Similar results were also obtained in other researches on continuous traits[15,16]. Jia and Jannink[11] simulated two continuous traits with heritability of 0.1 and 0.5, respectively, and the results showed that no increase in the accuracy for the trait with heritability of 0.5, while significant improvement was obtained for the trait with heritability of 0.1. In the present study, in the analysis of the scenario of $h_A^2 = 0.3$ and $h_B^2 = 0.5$ and of the common data set of the 14th QTL-MAS workshop, where both traits had a heritability of around 0.5, improvement in accuracy were obtained for both the continuous and the threshold trait from the joint analysis using LT-BayesCπ in comparison with single trait analysis, indicating that the proposed method has the potential to improve the accuracy of genomic prediction for a continuous trait, in addition to improving the accuracy for the threshold trait with a high heritability.

It has been generally accepted that the number of QTL controlling traits of interest affects the accuracy of genomic prediction by Bayesian methods [8,11,25,26]. This was also confirmed by our results. The prediction accuracies of all methods declined with an increase in the number of QTL. When the number of QTL increases, the effect of a QTL on average should become smaller, given a fixed total genetic variance, which will also decrease the accuracy of estimating SNP effects in a given training population.

When the heritability of the continuous trait increased, the accuracies of LT-BayesCπ increased not only for the continuous trait as expected, but also for the threshold trait. Meanwhile, when the heritability of the threshold trait increased, the accuracies of LT-BayesCπ increased not only for thethreshold trait as expected, but also for the continuous trait (Fig 3). These results imply that low-heritability traits can borrow information from correlated high-heritability traits, and consequently, achieve higher prediction accuracy as also observed by Jia and Jannink[11] and Guo et al.[12]. This is also in accordance with the findings in traditional genetic evaluation that the benefit of using a multiple-trait model will be more profound for traits with lower heritability.

The accuracy for the threshold trait increased as the incidence of the threshold trait approached to 0.5. The reason is that traits with a small incidence need larger training populations to estimate variance components and thus to achieve sufficient accuracies of GEBVs [8,27]. On the other hand, the accuracy of genomic prediction for the continuous trait was not affected by the change of the incidence of the threshold trait (Fig 4). This might be due to that the variance-covariance matrix did not change in LT-BayesCπ, resulting in negligible influence for the continuous trait.

In the analysis of the simulated data, we set the values of the hyper-parameters v_g, $\mathbf{V}_g$, v_e and $\mathbf{V}_e$ of the prior distributions of $\mathbf{G}_0$ and $\mathbf{R}_e$ to be -3, $\mathbf{0}$, -3, and $\mathbf{0}$, respectively, such that the priors reduced to flat priors. This may have an impact on the biased estimates. An alternative way to define them is to draw them from a flat distribution or other particular distributions. We will evaluate this to see their influence on the estimates of parameters in the near future.

Conclusions

Our work indicates that the linear-threshold model based method LT-BayesCπ is useful for predicting GEBVs of a continuous and a threshold trait jointly. In particular, a joint analysis using LT-BayesCπ significantly improved the accuracy for the threshold trait compared with single trait analysis. The larger the genetic correlation between the two traits is, the more benefit would be obtained. Increasing the heritability of the continuous or/and the threshold traits

is helpful to improve the genomic accuracy for both traits, particularly for the threshold trait. The incidence of the threshold trait affected the prediction accuracy only for the threshold trait. LT-BayesCπ could be a method of choice for a joint analysis of a continuous and a threshold trait.

Supporting information

S1 Appendix. The simulation data (one replication for one case) and compiled programs.
(BZ)

Author Contributions

Conceptualization: QZ XD.

Data curation: CW XL RQ.

Formal analysis: CW XL.

Funding acquisition: CW XL QZ XD.

Investigation: CW XL RQ.

Methodology: CW XL QZ XD.

Project administration: CW XL QZ XD.

Resources: XL RQ.

Software: CW XL QZ XD.

Supervision: QZ XD.

Validation: CW XL GS QZ XD.

Visualization: CW XL.

Writing – original draft: XL.

Writing – review & editing: CW GS QZ XD.

References

1. Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001; 157: 1819–1829. PMID: 11290733

2. Meuwissen THE, Solberg TR, Shepherd R, Woolliams J a. A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. Genet Sel Evol. 2009; 41: 2. https://doi.org/10.1186/1297-9686-41-2 PMID: 19284681

3. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: Genomic selection in dairy cattle: progress and challenges. J Dairy Sci. 2009; 92: 433–43. https://doi.org/10.3168/jds.2008-1646 PMID: 19164653

4. Su G, Guldbrandtsen B, Gregersen VR, Lund MS. Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. J Dairy Sci. 2010; 93: 1175–83. https://doi.org/10.3168/jds.2009-2192 PMID: 20172238

5. Ding X, Zhang Z, Li X, Wang S, Wu X, Sun D, et al. Accuracy of genomic prediction for milk production traits in the Chinese Holstein population using a reference population consisting of cows. J Dairy Sci. 2013; 96: 5315–23. https://doi.org/10.3168/jds.2012-6194 PMID: 23746588

6. Gianola D, de Los Campos G, Hill WG, Manfredi E, Fernando R. Additive genetic variability and the Bayesian alphabet. Genetics. 2009; 183: 347–363. https://doi.org/10.1534/genetics.109.103952 PMID: 19620397

7. Habier D, Fernando RL, Kizilkaya K, Garrick DJ. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics. 2011; 12: 186. https://doi.org/10.1186/1471-2105-12-186 PMID: 21605355

8. Wang CL, Ding XD, Wang JY, Liu JF, Fu WX, Zhang Z, et al. Bayesian methods for estimating GEBVs of threshold traits. Heredity (Edinb). 2012; 110: 213–219.

9. Iwata H, Ebana K, Fukuoka S, Jannink JL, Hayashi T. Bayesian multilocus association mapping on ordinal and censored traits and its application to the analysis of genetic variation among Oryza sativa L. germplasms. Theor Appl Genet. 2009; 118: 865–880. https://doi.org/10.1007/s00122-008-0945-6 PMID: 19132337

10. Calus MPL, Veerkamp RF. Accuracy of multi-trait genomic selection using different methods. Genet Sel Evol. 2011; 43: 1–14.

11. Jia Y, Jannink J-L. Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics. 2012; 192: 1513–1522. https://doi.org/10.1534/genetics.112.144246 PMID: 23086217

12. Guo G, Zhao F, Wang Y, Zhang Y, Du L, Su G. Comparison of single-trait and multiple-trait genomic prediction models. BMC Genet. 2014; 15: 30. https://doi.org/10.1186/1471-2156-15-30 PMID: 24593261

13. Jiang J, Zhang Q, Ma L, Li J, Wang Z, Liu J-F. Joint prediction of multiple quantitative traits using a Bayesian multivariate antedependence model. Heredity (Edinb). 2015; 115: 29–36.

14. Hayashi T, Iwata H. A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. BMC Bioinformatics. 2013; 14: 34. https://doi.org/10.1186/1471-2105-14-34 PMID: 23363272

15. Wang CS, EJ Q. Bayesian analysis of calving ease scores. 1997;

16. Varona L, Misztal I, Bertrand JK. Threshold-linear versus linear-linear analysis of birth weight and calving ease using an animal model: I. Variance component estimation. J Anim Sci. 1999; 77: 1994–2002. PMID: 10461973

17. Sorensen D, Gianola D. Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer; 2002.

18. Korsgaard IR, Lund MS, Sorensen D, Gianola D, Madsen P, Jensen J. Multivariate Bayesian analysis of Gaussian, right censored Gaussian, ordered categorical and binary traits using Gibbs sampling. Genet Sel Evol. 2003; 35: 159–184. https://doi.org/10.1186/1297-9686-35-2-159 PMID: 12633531

19. Szydlowski M, Paczyńska P. QTLMAS 2010: simulated dataset. BMC Proc. 2011; 5 Suppl 3: S3.

20. Korsgaard IR, Andersen AH, Sorensen D. A useful reparameterisation to obtain samples from conditional inverse Wishart distributions. Genet Sel Evol. 1999; 31: 177–181

21. Zhang Z, Li X, Ding X, Li J, Zhang Q. GPOPSIM: a simulation tool for whole-genome genetic data. BMC Genet. 2015; 16: 10. https://doi.org/10.1186/s12863-015-0173-4 PMID: 25652552

22. Henderson CR, Quaas RL. Multiple trait evaluation using relatives records. J Anim Sci. 1976; 43: 1188–1197.

23. González-Recio O, Forni S. Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. Genet Sel Evol. 2011; 43: 7. https://doi.org/10.1186/1297-9686-43-7 PMID: 21329522

24. Villanueva B, Fernández J, García-Cortés L a, Varona L, Daetwyler HD, Toro M a. Accuracy of genome-wide evaluation for disease resistance in aquaculture breeding programs. J Anim Sci. 2011; 89: 3433–42. https://doi.org/10.2527/jas.2010-3814 PMID: 21742941

25. Coster A, Bastiaansen JWM, Calus MPL, van Arendonk JAM, Bovenhuis H. Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. Genet Sel Evol. 2010; 42: 9. https://doi.org/10.1186/1297-9686-42-9 PMID: 20302681

26. Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams J a. The impact of genetic architecture on genome-wide evaluation methods. Genetics. 2010; 185: 1021–1031. https://doi.org/10.1534/genetics.110.116855 PMID: 20407128

27. Gilmour AR, Anderson RD, Rae AL. The analysis of binomial data by a generalized linear mixed model. Biometrika. 1985; 72: 593–599.