Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Application of Whole-Genome Prediction Methods for Genome-Wide Association Studies: a Bayesian Approach

R.L. Fernando, A. Toosi, D.J. Garrick, and J.C.M. Dekkers.
Department of Animal Science, Iowa State University, Ames, Iowa, 50011-3150, USA

ABSTRACT: This paper discusses how Bayesian multiple-regression methods that are used for whole-genome prediction can be adapted for genome-wide association studies (GWAS). It is argued here that controlling the posterior type I error rate (PER) is more suitable than controlling the genomewise error rate (GER) for controlling false positives in GWAS. It is shown here that under ideal conditions, PER can be controlled by using Bayesian posterior probabilities that are easy to obtain. Computer simulation was used to examine the properties of this Bayesian approach when the ideal conditions were not met. Results indicate that useful inferences can be made using Bayesian posterior probabilities.
Keywords:
genome-wide association studies
Bayesian multiple-regression analyses

Introduction

Genomic prediction requires obtaining genotypes and phenotypes on several thousand animals in a training population to estimate effects of the SNP genotypes on the traits of interest. The estimated SNP effects are then used to predict the breeding values of selection candidates that may not have any phenotypes recorded but have been genotyped (Meuwissen et al. 2001). The genotype and phenotype data obtained for whole-genome prediction can also be used for genome-wide association studies (GWAS) to locate causal variants (QTL) for traits of economic importance.

Many GWAS for quantitative traits are based on testing one SNP at a time using simple regression models or using a mixed models with a fixed substitution effect of the SNP genotype included, along with a polygenic effect correlated according to pedigree relationships to capture the effects of all other genes. Such GWAS have been successful in detecting many associations, but the established associations typically explain only a small fraction of the genetic variability of quantitative traits (Visscher et al. 2010). On the other hand, in analyses that use whole-genome selection models that simultaneously fit all SNPs as random effects, the SNPs jointly explain a large proportion of the genetic variance (Onteru et al. 2010; Hayes et al. 2010; Fan et al. 2011). In these analyses, however, any given SNP may have only a weak association even with a closely linked QTL. The reason for this is that in a high-density SNP panel many SNP genotypes within a narrow genomic region are expected to be highly correlated with each other and with any QTL that are close to them. So, any one SNP may contribute only a little more to explain the variability of the QTL in addition to the other SNPs in the neighborhood. On the other hand, even if each SNP in a neighborhood is only weakly associated with a QTL, the SNPs in the neighborhood may jointly explain much more of the variability of a QTL than any SNP by itself. Therefore in multiple-regression models, SNPs in a genomic window should be used jointly to locate QTL (Onteru et al. 2010; Sahana et al. 2010; Hayes et al. 2010; Fan et al. 2011).

Inferences on genomic windows by frequentist methods, however, are computationally very challenging because they require repeated analyses of the data with bootstrap or permuted samples to obtain significance levels for tests (Onteru et al. 2010; Hayes et al. 2010; Fan et al. 2011). It can be shown that Bayesian posterior probabilities obtained from a single MCMC analysis can be used to make inferences on genomic windows. This approach to inference is related to the frequentist approach of controlling the conditional probability of a false positive (type I error) given a positive (significant) result, which is referred to as the the posterior type I error rate (PER) in the human linkage analysis literature (Morton 1955; Ott 1991; Risch 1991; Elston 1997). When multiple tests are involved, it has been shown that controlling the PER for a randomly chosen test is equivalent to controlling proportion of false positives (PFP) from the collection of all tests (Fernando et al. 2004). In contrast to controlling the genomewise error rate (GER), controlling PER or PFP has the property that the power

of detecting associations is not inversely related to the number of tests (Fernando et al. 2004; Stephens and Balding 2009). This property is especially attractive in GWAS, where the number of tests can be very large.

A requirement for controlling PER is knowledge of the distribution of the test statistic under the null hypothesis of no association, which is also required to control the usual type I error rate. In addition to this requirement, controlling PER requires knowing the proportion π of SNPs for which the null hypothesis is true and the average power of the test, which is the average probability of rejecting the null hypothesis when it is not true. These quantities are almost never known in a GWAS of a quantitative trait, and thus PER cannot be controlled in a manner that the usual type I error rate can be controlled (Elston 1997).

In contrast to frequentist methods, when the Bayesian multiple-regression methods that are used for whole-genome prediction (Meuwissen et al. 2001) are applied to GWAS, posterior probabilities that are similar to PER can be obtained even without the requirement of knowing the null distribution of any test statistic. Further, in Bayesian analyses, π and the magnitude of the partial regression coefficients of markers, which determine average power, can be formally treated as unknowns such that their uncertainty is incorporated in the inference.

A posterior probability from a Bayesian analysis, however, is not a conditional probability in the frequentist sense. It is an expression of belief of some event of interest conditional on the observed data. In the multiple-regression models used here for GWAS, when the posteriors for π and the partial regression coefficients of markers are close to their true values, the posterior probability of an association is expected to be closely related to the frequentist conditional probability of a true association given the data. A computer simulation is used to examine this relationship between the posteror probability of association and the true frequency of association.

Methods

Controlling False Positives. The longstanding practice of using a lod score of three for declaring linkage between a monogenic disease locus and a random marker is based on control of the posterior type I error rate (PER) to about 0.05 (Elston 1997). In the multiple-test setting, it has been shown that the PER for a randomly chosen test from a family of k tests is equal to the level of PFP for the entire family of k tests (Fernando et al. 2004). So, it is clear that if PER for each test is controlled to a level γ, PFP for the entire set of tests will be also be controlled at γ. In a Bayesian analysis this can be achieved by declaring an association only when the posterior probability of association is greater than $(1-\gamma)$. In any such declaration, the posterior probability of no association will be $< \gamma$, resulting in γ being the upper bound for PFP.

Control of PFP is closely related to the control of the false discovery rate (FDR) (Benjamini and Hochberg 1995) and its close relative the positive false discovery rate (pFDR) (Storey 2002). Let V denote the number of false positive results and R the number of positives from a multiple-test experiment. Then, PFP is defined as

$$\mathrm{PFP} = \frac{\mathrm{E}(V)}{\mathrm{E}(R)},$$

FDR is defined as

$$\mathrm{FDR} = \mathrm{E}(\frac{V}{R}|R > 0)\,\Pr(R > 0),$$

and pFDR as

$$\mathrm{pFDR} = \mathrm{E}(\frac{V}{R}|R > 0)$$

Our justification for use of PFP to control false positives is that if PFP is controlled at say γ for each of n independent experiments, the proportion of false positives among significant results across all n experiments converges to γ as the number of experiments increases (Fernando et al. 2004). In general, this property does not hold for FDR or pFDR (Fernando et al. 2004).

The most widely used approach to control false positives is controlling the genomewise type I error rate (GER). If only one marker is tested, controlling GER to 0.05 will result in a much larger value for PER, i.e., among significant results a large proportion would be false positives (Fernando et al. 2004). Thus, to control PER to 0.05, a more stringent significance threshold has to be used. Now, suppose several markers are tested with the same stringent significance threshold each with the same prior probability of association and power of test. Then for each test, PER would be 0.05, even when the tests are not independent. So, it can be reasoned that if the proportion of errors among significant results from each test is 0.05, then among all the significant results the proportion of errors will also be 0.05. Thus, provided the prior probability of association and power are constant, the same significance threshold can be used to control the proportion of errors among significant results regardless of the number of tests (Fernando et al. 2004). In other words, when PER is used to control false positives there is no multiple-test penalty (Stephens and Balding 2009). This is not true for GER, which for a given significance threshold increases with the number of tests. So when the number of tests increases, to

maintain the same GER even more stringent significance thresholds need to be employed, incurring the multiple-test penalty of lower power to detect associations.

To control PER it is required to know the value of π, which is the probability that the null hypothesis is true for a test, and the average power of the test. On the other hand, in Bayesian analyses, quantities such as π can be treated as unknowns and an upper bound for PER can be obtained from Bayesian posterior probabilities. For example, the PER for the test of association for a genomic window, W, is obtained as 1 - WPPA, where WPPA is estimated by counting the number of MCMC samples in which α_j is non-zero for any SNP j in W_C, the central window of W (Figure 1). A Bayesian posterior probability, however, has a different meaning from a frequentist conditional probability. So, computer simulation was used to examine the relationship between these two probabilities.

Bayesian Regression. Following Meuwissen et al. (2001), consider the mixed linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{j=1}^{K} \boldsymbol{z}_j \alpha_j + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{y}$ is the vector of trait phenotypes, $\boldsymbol{X}$ is an incidence matrix relating the vector of non-genetic, fixed effects $\boldsymbol{\beta}$ to $\boldsymbol{y}$, $\boldsymbol{z}_j$ is a vector of genotype covariates (coded as 0, 1 or 2) for SNP j, α_j is the random, partial regression coefficient for SNP j, and $\boldsymbol{e}$ is a vector of residuals. In this model, the fixed effects are assumed to have a flat prior, and the α_j are a priori assumed independently distributed as

$$\alpha_j|\pi, \sigma^2_{\alpha_j} = \begin{cases} 0 & \text{with probability } \pi \\ \sim \mathrm{N}(0, \sigma^2_{\alpha_j}) & \text{with probability } (1-\pi), \end{cases} \tag{2}$$

where the $\sigma^2_{\alpha_j}$ are a priori assumed independently and identically distributed (iid) scaled inverse chi-square variables with scale S^2_α and degrees of freedom ν_α. The residuals are assumed iid normal with null mean and variance σ^2_e, with a scaled inverse chi-square prior for σ^2_e with scale S^2_e and degrees of freedom ν_e. Inferences on the unknowns in the model are made from their marginal posterior distributions, using Markov Chain Monte-Carlo (MCMC) methods (Meuwissen et al. 2001; Habier et al. 2010).

Although this model was first proposed for whole-genome prediction (Meuwissen et al. 2001), it can also be used to locate genomic regions that contain QTL (Yi et al. 2003; Sun et al. 2011; Fan et al. 2011). Consider a model where π is close to one, i.e., a model where most regions of the genome do not have markers that are associated with the trait. This is the model called BayesB by (Meuwissen et al. 2001). Given such a model, the posterior probability that α_j is non-zero for at least one SNP j in a window or region can be used to make inferences on the presence of QTL in that region. We will refer to this probability as the window posterior probability of association (WPPA). The underlying assumption here is that if a genomic window contains a QTL, one or more SNPs in that window will have non-zero α_j. Thus, WPPA, which is estimated by counting the number of MCMC samples in which α_j is non-zero for at least one SNP j in the window, can be used as a proxy for the posterior probability that the genomic window contains a QTL and is thus "associated with the trait".

It is possible, however, that SNPs in a window that does not contain any QTL are in association with a QTL outside the window, which is what has been called signal dependence (Chen and Storey 2006). Fortunately, for most populations, the linkage signal from LD extends over only short distances compared to that from cosegregation, and as described below, when all SNPs are fitted simultaneously, signal dependence is further reduced. Let W_C denote the window for which WPPA is estimated. Let W_L and W_R be windows of length k cM to the left and right of W_c as illustrated in Figure 1. A high WPPA for W_C is taken as evidence of a QTL in the "composite" window W comprised of W_L, W_C, and W_R. Because WPPA for W_C is a partial association conditional on all other SNPs in the model, including those in the flanking windows W_L and W_R, the influence of QTL from outside the composite window on the WPPA signal for W_C will be inversely related to the length k of the flanking windows. In other words, as the number of markers between a QTL and W_C increases, the influence of the QTL on the WPPA signal for W_C is expected to decrease. The computer simulation described next will serve to examine this expectation in addition to the relationship between WPPA and the true frequency of association

Computer Simulation. The simulation described here was used to test if WPPA can be used to control false positives in GWAS, where the tests are dependent. Actual SNP genotypes of purebred Angus bulls were used to simulate QTL and phenotypes as described in Kizilkaya et al. (2010). Exactly 100 data sets with 1,000 observations and another 100 with 3,570 observations were simulated, using genotypes at 52,910 SNP loci on 3,570 purebred Angus bulls. The 1,000 bulls were randomly sampled without replacement for inclusion in the data sets with 1,000 observations, whereas all bulls were used in the data sets with 3,570 observations.

In each of the 200 data sets, SNP effects of markers were sampled according to the prior of the BayesC model of Habier et al. (2011) with $\pi = 0.995$, where a

proportion π of the loci have null effects and the remaining loci have normally distributed effects with null mean and common variance σ_α^2 of SNP effects. The value of the common variance of SNP effects was chosen as in (Kizilkaya et al. 2010) such that the additive genetic variance for the trait was 0.9. The average number of QTL in the data sets was about 260. The residual variance for the trait was set at 0.1 to give a heritability of 0.9, reflecting the heritability of progeny means. Both data were analyzed without including the SNPs that represent the QTL in the marker panel. Posterior inferences were based on 10,000 MCMC samples after a burn-in of 1,000 samples.

In all analyses, the genome was divided into 2,676 one cM intervals according to the bovine map. The WPPA was computed for each such window, W_C, as explained previously. When the QTL are not included in the marker panel, it is not straightforward to determine if W_C is or is not associated with the trait. In this study, W_C was defined to have an association if it, or windows W_L or W_R, flanking W_C, contained one or more QTL. In order to study the relationship between WPPA and the true frequency of association, each genomic window, W_C, was classified into one of 10 WPPA classes of length 0.1 between 0 and 1. For example, all windows with WPPA between 0 and 0.1 were be classified into the first class, and those with WPA between 0.1 and 0.2 to the second class. The true frequency of association for a WPPA class j was estimated as the frequency of the total number of composite windows belonging to class j that contained at least 1 QTL relative to the total number of windows belonging to that class.

Recall that the prior of BayesC with $\pi = 0.995$ was used in the simulation of SNP effects. Thus, WPPA from a BayesC analysis with $\pi = 0.995$ is expected to agree well with the actual frequency of the QTL. BayesC and BayesB with $\pi = 0.995$, and BayesCπ, where π is treated as an unknown (Habier et al. 2011), were used to analyze the data sets with 1,000 observations without including the QTL in the marker panel.

Results and Discussion

Figure 2 presents results from three analyses of the 100 data sets with 1,000 observations, and Figure 3 gives results for the 100 data sets with 3,570 observations. In these analyses, which did not have the QTL included in the marker panels, in genomic windows of 1cM ($k = 0$), WPPA for W_C substantially overestimated the frequency of association when WPPA was greater than about 0.15. For example, in plot B of Figure 2, which shows the relationship between WPPA and the frequency of association for BayesC with $\pi = 0.995$, in genomic windows of 1cM with WPPA between 0.9 and 1.0, the frequency of association was about 0.72 and in genomic windows of 1cM with WPPA between 0.8 and 0.9, the frequency of association was only about 0.5. When the QTL were included in the analysis, the comparable QTL frequencies were 0.97 and 0.81 (results not shown). Thus, when the QTL were not in the panel, WPPA overestimated the frequency of association for W_C. Following are two possible reasons for this. The first is that the prior used for marker effects does not agree with the actual distribution of effects. When the QTL are not included in the marker panel, only markers that are in complete LD with the QTL will have effects that are distributed as the QTL. In Angus, the average LD between adjacent markers for the 50k SNP panel is about 0.2. Thus, the distribution of marker effects may be quite different from that of the QTL and this may have an impact on the relationship between WPPA for a genomic interval and the frequency of association for that interval even when the distribution used to generate the QTL effects is used as the prior for marker effects as in the BayesC analysis with $\pi = 0.995$. The second reason is violation of the assumption that WPPA is equivalent to the posterior probability that W_C contains a QTL (WPPQ), which is our definition of a true association when $k = 0$. Recall that WPPA is the posterior probability that a marker in window W_C has a non-zero effect on the trait. When the QTL are included in the panel, WPPA is also the posterior probability of a QTL in W_C because QTL by definition have non-zero effects on the trait. However, when the QTL are not included in the panel, WPPA is not equivalent to probability of a QTL in W_C. A marker in W_C may have a non-zero effect even when W_C does not contain any QTL due to it being in LD with a QTL in an adjacent window. This would cause WPPA to be higher than WPPQ, which is consistent with our results.

It can be argued that both of the reasons given above played a role in the observed over estimation WPPQ by WPPA. Violation of the assumption that WPPA is equivalent to WPPQ, however, seems to have played a greater role. The three plots in Figure 2 were obtained using three different priors. Plot A is from a BayesB analysis with $\pi = 0.995$, where a central t distribution with four degrees of freedom was used as the prior for marker effects. Plot B is from BayesC with $\pi = 0.995$, where a normal distribution is used for marker effects, and plot C is from BayesCπ, where π is treated as unknown with a uniform prior between 0 and 1 and a normal prior for marker effects. The fact that the results from these three analyses were very similar indicates that with 1,000 observations these differences in priors had a negligible effect on the relationships between WPPA and QTL frequencies. Further, if the overestimation of WPPQ by WPPA was due to the prior for

marker effects not being appropriate, then better results would be expected in the data sets with 3,570 observations. However, this was not the case. Overestimation was even greater with the bigger data sets (Figure 3). On the other hand, if the observed overestimation of WPPQ was due to markers in W_C being in LD with QTL in adjacent windows, it is possible that with more data associations with even more distant QTL could further inflate WPPA. Comparison of true frequencies of association in plot C of Figure 2 with those in Figure 3 for genomic windows with WPPA between 0.8 and 0.9 and $k = 0, 1,$ and 2 suggests that with the bigger data sets more distant QTL contributed to the WPPA value calculated for W_C. In these analyses that did not include the QTL in the marker panel, there was good agreement between WPPA and the true frequency of association (WPPQ) for the composite window W with $k = 2$ when WPPA was larger than 0.8.

Acknowledgements

This work was supported by the US Department of Agriculture, Agriculture and Food Research Initiative National Institute of Food and Agriculture Competitive grant no. 2012-67015-19420 and by National Institutes of Health grant R01GM099992.

Literature Cited

Benjamini, Y. and Hochberg, Y. (1995) *J. R. Statist. Soc. B* 57:289–300.

Chen, L. and Storey, J. D. (2006) *Genetics* 173(4):2371–2381.

Elston, R. C. (1997) *Am. J. Hum. Genet.* 60:225–262.

Fan, B., Onteru, S. K., Du, Z.-Q., Garrick, D. J., Stalder, K. J. and Rothschild, M. F. (2011) *PLoS ONE* 6(2):e14726.

Fernando, R. L., Nettleton, D., Southey, B., Dekkers, J., Rothschild, M. and Soller, M. (2004) *Genetics* 1666;11–619.

Habier, D., Fernando, R., Kizilkaya, K. and J., G. D. (2010) Proc. *9th Wld. Congr. Genet. Appl. Livest. Prod.* 9:468.

Habier, D., Fernando, R. L., Kizilkaya, K. and Garrick, D. (2011) *BMC Bioinformatics* 12:186.

Hayes, B. J., Pryce, J., Chamberlain, A. J., Bowman, P. J. and Goddard, M. E. (2010) *PLoS Genet.* 6(9):e1001139.

Kizilkaya, K., Fernando, R. L. and Garrick, D. J. (2010) *J. Anim. Sci.* 88(2):544–551.

Meuwissen, T. H. E., Hayes, B. J. and Goddard, M. E. (2001) *Genetics* 157:1819–1829.

Morton, N. (1955) *Am. J. Hum. Genet.* 7:277–318.

Onteru, S. K., Fan, B., Nikkilä, M. T., Garrick, D. J., Stalder, K. J. and Rothschild, M. F. (2010) *J. Anim. Sci.* 89(4):988–995.

Ott, J. (1991) Analysis of Human Genetic Linkage. Johns Hopkins University Press, Baltimore.

Risch, N. (1991) *Am. J. Hum. Genet.* 48:1058–1064.

Sahana, G., Guldbrandtsen, B., Janss, L. and Lund, M. S. (2010) *Genet. Epid.* 34:455–462.

Stephens, M. and Balding, D. J. (2009) *Nat. Rev. Genet.* 10(10):681–690.

Storey, J. D. (2002) *J. R. Statist. Soc. B* 64:479–498.

Sun, X., D, H., R.L, F., Garrick, D. and J.C.M., D. (2011) *BMC proc.* 5(Suppl 3):S13.

Visscher, P. M., Yang, J. and Goddard, M. E. (2010) *Twin Res. Hum. Genet.* 13(6):517–524.

Yi, N., George, V. and Allison, D. B. (2003) *Genetics* 164:1129–1138.

Figure 1. Illustration of composite genomic window W consisting of central window W_C and flanking windows W_L and W_R. To test the null hypothesis of no QTL in W, window PPA (WPPA) is computed by counting the number of MCMC samples in which α_j is non-zero for at least one SNP in the central window W_C.

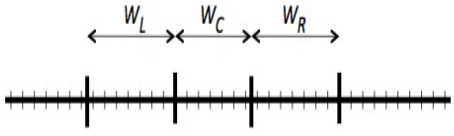

Figure 2. Relationship between window posterior probability of association (WPPA) and the actual frequency of association (WPPQ). WPPA was computed for each 1cM window (W_C) of the genome and grouped into 10 WPPA classes (x-axis). For each WPPA class, the actual frequency of simulated QTL in the composite window consisting of W_C and the flanking windows of k cM ($k = 0, 1,$ or 2) in length is given in the y-axis as the frequency of association. Results are for BayesB with $\pi = 0.995$ (plot A), BayesC with $\pi = 0.995$ (plot B), and BayesCπ (plot C) from from 100 data sets each with 1,000 observations. The QTL were not included in the marker panel.

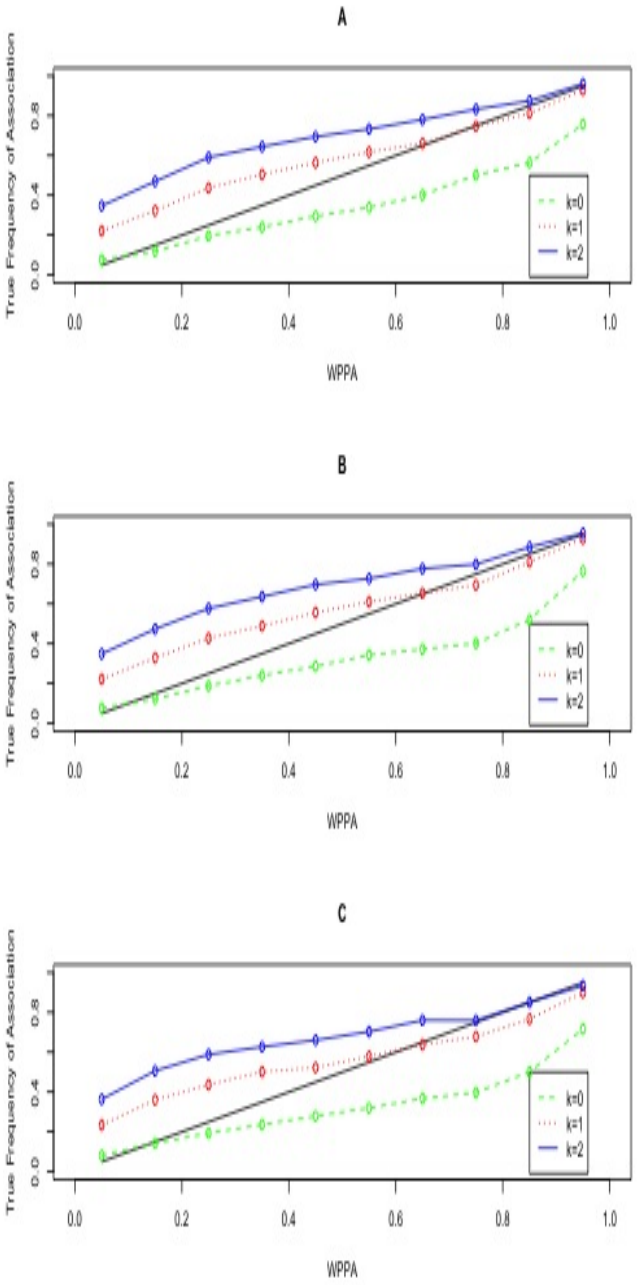

Figure 3. Relationship between window posterior probability of association (WPPA) and the actual frequency of association (WPPQ). WPPA was computed for each 1cM window (W_C) of the genome and grouped into 10 WPPA classes (x-axis). For each WPPA class, the actual frequency of simulated QTL in the composite window consisting of W_C and the flanking windows of k cM ($k = 0, 1,$ or 2) in length is given in the y-axis as the frequency of association. Results are for BayesCπ from 100 data sets each with 3,570 observations. The QTL were not included in the marker panel.