Multiple trait single step Bayesian GWAS on pooled data

J.P. Sánchez[1], *A. Legarra*[2] *& M. Piles*[1]

[1] *Institut de Reserca i Tecnologia Agroalimentaries (IRTA), Animal Breeding and Genetics*
Torre Marimon, 08140 Caldes de Montbui,Spain
juanpablo.sanchez@irta.es (Corresponding Author)
[2] *UMR 1388 GenPhySE, INRA. BP 52627, 31326 Castanet Tolosan, France.*

Summary

Socially affected traits must be recorded in group-housed animals. In these conditions sometimes it is not possible to obtain individual records of certain traits and in some species. Previous studies have addressed the issue on how to use group (i.e. pooled) data to estimate genetic parameters and to predict breeding values, but the value of this pooled data to conduct QTL mapping studies has not been assessed yet. . Our objective was to present a method, based on a Bayesian single step genomic evaluation, in which the SNPs effects were assumed to follow the prior of Bayesian Cπ model, allowing thus a variable selection approach to pinpointing the genome regions most likely harbouring QTLs. The method was applied to a multi-trait simulated data set, in which for one of the traits pooled records were generated summing groups of 10 individual records. Our results show that an important loss of power was observed when pooled data were used, but even though one of the true QTLs can be detected with a probability of being associated to the trait of 0.83. This QTL was associated to a mutation explaining 16% of the genetic variance and with a frequency of 0.46. Other mutation with an even greater effect (22%) but with lower frequency (0.38) could not be detected. It can be concluded that although the proposed model can be used for QTL mapping when grouped data are available its power is limited and only strongly associated regions are likely to be declared as QTLs.

Keywords: group data, multiple-regression Bayesian GWAS, single step, Bayes Cπ

Introduction

The single step approach (Aguilar *et al.*, 2010) for implementing genomic selection offers a excellent framework to properly consider complex models accounting for data from either genotyped or non-genotyped animals, without the need of defining pseudo-records that could lack accuracy or might introduce biases because of uncertainty associated to them is not properly defined. Examples of these models could be multi-trait, maternal or random regression models. Wang *et al.* (2012) proposed a method to retrieve SNP effects from single step approaches, integrating phenotypes from genotyped and non-genotyped individuals. They showed how higher prediction accuracy can be achieved by considering functions of the SNP effects to differential weight regions in the genome. Their implementation was done based on punctual predictions of genomic breeding values, so it does not account for the uncertainty of the genomic predictions, and also did not provide a measurement of the uncertainty of the estimated SNP effects. The Bayesian MCMC setting provides a scenario to characterize marginal posterior distributions of SNP effects or function of them, accounting for the uncertainty of the genomic breeding values predictions.

Sometimes recording data at group level is the best option. For example, in socially affected traits, it is necessary to keep animals housed in groups for these effects to be expressed. In those conditions, it is not possible to get individual records for some traits and species such as feed intake in rabbit or egg production in layers.

The study of a trait recorded at group level can be done using models in which the observation is explained by a function of the breeding values of the animals forming the group. This kind of models has been shown to be useful for the study of different traits in layers (Biscarini *et al.*, 2008), pigs (Sánchez *et al.*, 2014), or rabbits (Sánchez *et al.*, 2016). Obviously the accuracy of the predictions from these pooled data would be much lower than when records are taken at individual level, but the consideration of other correlated traits in a multivariate scenario would improve the accuracy of the genetic predictions (Sánchez *et al.*, 2014).

The objective of the present study was to explore the potential of models used to fit pooled data to directly explore the existence of genomic regions associated to some trait when only pooled records of this trait are available. To do this we extended a Gibbs Sampler implementation for conducting single step genomic evaluations.

Material and methods

Data

The simulated data set generated for the 16th European Workshop on QTL Mapping and Marker Assisted Selection was used (Usai *et al.*, 2012).It has genotypic and phenotypic information corresponding to 3 generations, involving a total of 3000 records from 1000 females. Twenty males per generation and 20 founders only had genotypes. In the present study only the first (T1) and the third (T3) traits were analysed. Heritabilities were 0.35 and 0.5 for T1 and T3, respectively, and both residual and genetic correlations were -0.45. In the simulated data set 50 QTLs were defined, but only 13 were detectable for T1 and T3, being 6 of them common to both traits. Further details of the data set and the simulation process can be found in the paper by Usai *et al.* (2012).

In a first analysis individual records of T3 were used. In a second analysis, 10 consecutive records were pooled (i.e. added up) to define a group performance record.
The original SNP panel was formed by 10000 SNPs, evenly distributed every 50Kb along 5 chromosomes. After quality control 9010 SNPs were retained for latter analyses.

Statistical Models

Two bivariate models were implemented. In the first one the following univariate model was fitted to both T1 and T3.

$$\mathbf{y}_{\mathrm{t}} = \mathbf{1}\mu_{\mathrm{t}} + \mathbf{Z}\mathbf{u}_{\mathrm{t}} + \mathbf{e}_{\mathrm{t}} \qquad (1)$$

where μ_{t} is an overall mean, $\mathbf{u}_{\mathrm{t}}$ is a vector with the genomic breeding values of the trait t, $\mathbf{Z}$ is a design matrix relating breeding values to observations and $\mathbf{e}_{\mathrm{t}}$ is a vector of residuals.

The conditional likelihood of the data was assumed to follow a normal distribution with mean $(\mathbf{1}\mu_1 + \mathbf{Z}\mathbf{u}_1, \mathbf{1}\mu_2 + \mathbf{Z}\mathbf{u}_2)'$ and variance $(\mathbf{r_0} \otimes \mathbf{I})$, where $\mathbf{r_0}$ is 2 x 2 matrix containing the residual (co)variances. Flat prior distributions were assumed for μ_1 and $\mathbf{r_0}$. The prior distribution of the genomic breeding value was the usual multivariate normal distribution under a single step approach (Aguilar *et al.*, 2010),

$$\mathrm{p}(\mathbf{u}|\mathbf{H},\mathbf{g}_0) = \mathrm{MVN}(\mathbf{0},\mathbf{g}_0 \otimes \mathbf{H}) \tag{2}$$

where $\mathbf{H}$ is a matrix containing either genomic or pedigree relationships depending on whether a particular animal was genotyped or not (Aguilar *et al.*, 2010). In this study, all animals in the pedigree were genotyped so $\mathbf{H}$ matrix was fully defined as the $\mathbf{G}$ matrix described in Vanraden *et al.* (2008), which is the one used by default in blupf90 family programs. Thus a GBLUP scenario was actually defined. An uniform prior distribution was assumed for elements of $\mathbf{g}_0$, genomic (co)variances.

In a second hierarchical level, parameters for $\mathbf{u}_\mathrm{t}$ were defined according to the following model:

$$\mathbf{u}_\mathrm{t} = \mathbf{M}\mathbf{a}_\mathrm{t} + \boldsymbol{\varepsilon}_\mathrm{t} \tag{3}$$

at this level the conditional distribution of the genomic breeding values was defined as a multivariate normal density with vector of means $\mathbf{M}\mathbf{a}_\mathrm{t}$ and covariance matrix $(\mathbf{I}\sigma^2_{\varepsilon_\mathrm{t}})$. Where $\mathbf{M}$ is the matrix of standardized genotypes (the one used for obtaining $\mathbf{G}$), $\mathbf{a}_\mathrm{t}$ is a vector of SNP additive effects, and $\sigma^2_{\varepsilon_\mathrm{t}}$ is the residual variance at this level.

The adopted prior distribution for $\mathbf{a}_\mathrm{t}$ was that of Bayes Cπ (Habier *et al.*, 2011), assuming with probability π that the SNP effect is zero, and with probability $(1-\pi)$ that its effect come from a normal distribution with variance following a priori an inverted Chi-square distribution. A normal density was adopted as prior for the residual distribution in this second level of hierarchy, and an inverted Chi-square distribution as prior density for its variance $(\sigma^2_{\varepsilon_\mathrm{t}})$. π was assumed to be known and equal to 0.995. Genomic breeding values were assumed to be independent across traits, thus the residual covariance matrix between traits could be defined as $\mathbf{r}_{\boldsymbol{\varepsilon}}$, a diagonal matrix containing $\sigma^2_{\varepsilon_1}$ and $\sigma^2_{\varepsilon_2}$.

In the second analysis, T3 was considered to be the sum of 10 consecutive individual records. In this case, the model was defined as before except for the first stage, which was defined as:

$$\begin{aligned}\mathbf{y}_1 &= \mathbf{1}\mu_1 + \mathbf{Z}\mathbf{u}_1 + \mathbf{Z}\mathbf{d}_1 + \mathbf{e}^*_1\\ \mathbf{y}^*_3 &= \mathbf{1}\mu^*_3 + \mathbf{Z}_\mathrm{g}\mathbf{u}^*_3 + \mathbf{Z}_\mathrm{g}\mathbf{d}_3 + \mathbf{e}^*_3\end{aligned} \tag{4}$$

where $\mathbf{y}^*_3$ represents the mean of 10 measurements, $\mathbf{d}$ would be a vector of dummy effects, specific for each animal and assumed to be normally distributed and independent between animals within a trait, but correlated across traits. This effect is actually part of the residual, and it is considered just to take into account the residual correlation between traits, because in this second analysis actual residuals, $\mathbf{e}^*_1$ and $\mathbf{e}^*_3$, has to be assumed to be independent. The rest of component of the models, as well as the prior assumption and the structure of the second hierarchical level is the same as that described for the first analysis.

In both analyses windows of 5 consecutive SNPs (200 Kb) were defined and for each one of these segments the window posterior probability of association (WPPA) was computed, this was defined as in Fernando *et al.*, (2014) i.e. computing the number of MCMC rounds in which at least 1 SNP in that particular window had non null effect in and them dividing by the total number of runs.

In the two analyses 0.5 million gibbs sampler rounds were obtained, discarding the first 100,000 as burning period.

Results and discussion

Figures 1 and 2 shows the Manhattan plots obtained for the WPPA when the two traits were recorded at individual level, or when one of the traits was recorded as the sum of 10 consecutive individual records. When the analysis was conducted using individual records, 2 and 7 QTLs were respectively declared for T1 and T3, respectively. In the analysis using pooled records of T3 3 QTLs were declared for T1 while only 1 for T3.

A window was declared to be associated to the trait when its WPPA was higher than 0.75. We chose this threshold because 0.75 seems a reasonable minimum value to interpret results as likely associated, being these results the base for further analyses and experiments focusing in that particular region of the genome. In addition, Fernando *et al.*, (2014) showed that for values of WPPA greater than 0.7, a good agreement with WPPQ (actual frequency of association) was observed as far as the region actually harbouring the mutation included flanking windows of twice the length of the focus windows. This could be interpreted as a measurement of the length of the uncertainty region around an associated window, i.e. the uncertainty region would be twice as large as the focus window at both sides of this.

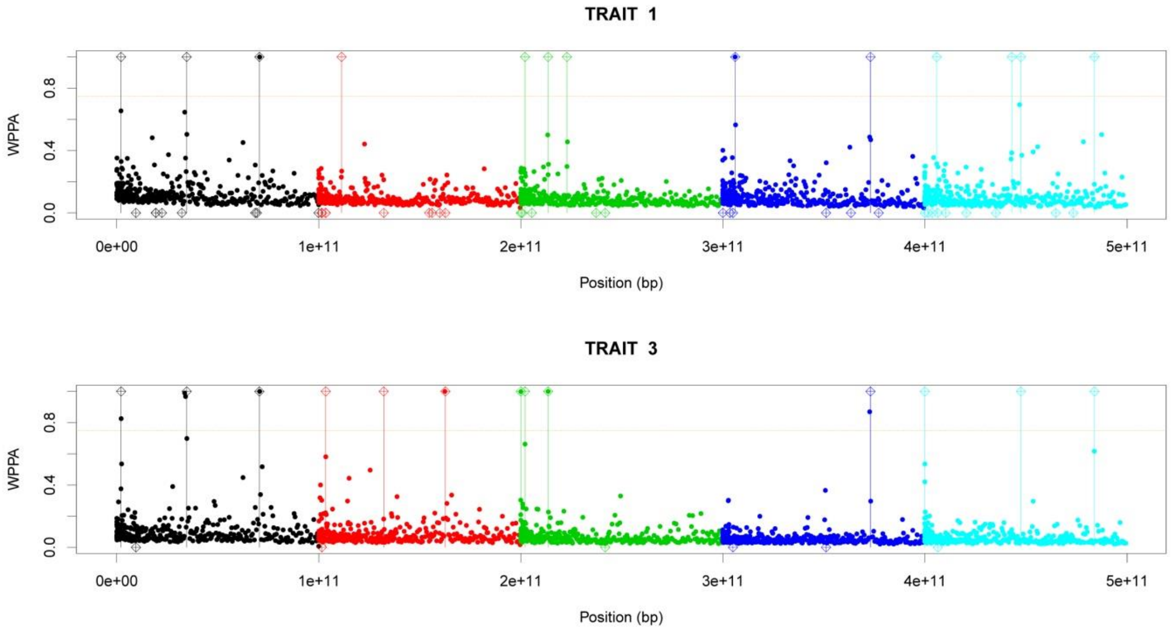

Figure 1.- Manhattan plot based on WPPA (200Kb Windows Posterior Probability of Association) obtained with Trait 3 defined as individual data. Vertical bars represent the position of QTLs declared to be detectable; points at zero WPPA represent no detectable QTLs.

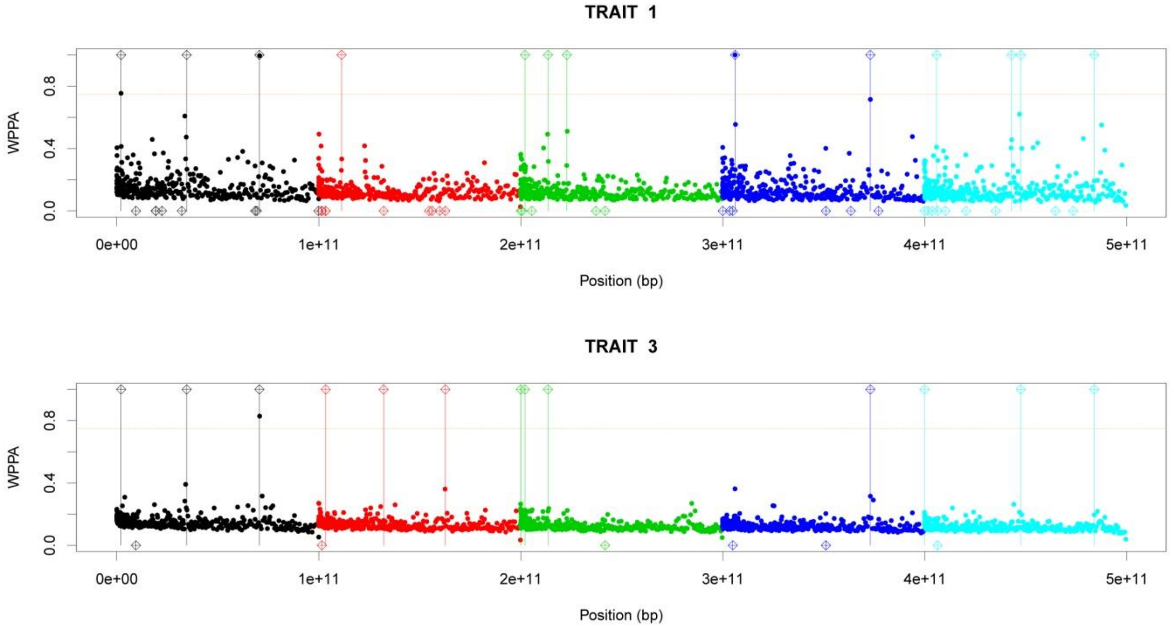

Figure 2.- Manhattan plot based on WPPA (200Kb Windows Posterior Probability of Association) obtained with Trait 3 defined as pooled data. Vertical bars represent the position of QTLs declared to be detectable; points at zero WPPA represent no detectable QTLs.

Table 1 shows the windows declared to be QTLs regions, as well as their closest actual mutation. In the analysis of individual records, 3 out 9 declared QTLs, actually contained the simulated mutation. For the other 6 the distance to the actual mutation from the closest bound of the window was always lower that 1Mb and in 3 cases this distance was only 25 Kb. Given that the focus windows are 200-300 Kb length (some SNPs were removed during quality control), the uncertain region would cover up to 500 Kb on each side of a focus window. Thus, only in the worst case, the actual mutation being 875 Kb apart from the window bounds, it can be said that the uncertainty region of this QTL does not include the actual mutation.

In general, results based on individual records can be said to be similar, in terms of detection capability, to the average performance of all the methods presented in the conference for which the dataset was generated (Usai *et al.*, 2012). For T3 we were able to detect mutations with an effect as low as 2.57% genetic variance, although others with larger effect could not been detected. For T1 (with a lower heritability) the detection capability was lower, and the lowest detected effect was 7.6% genetic variance, but similarly to T3, other mutations with larger effect were not detected. One important feature of the approach presented here is that a nearly null false discovery rate was actually observed. The only detected signal that could be said to be a false positive it was less than 1 Mb apart from the actual mutation.

When T3 was analysed as pooled data, the power of detection drastically drops. Nevertheless, even in these conditions 1 QTL region was declared with a WPPA equal to 0.83 (Table 1). The mutation associated to this region was not included within the 200 Kb window declared as associated, but it was 25 Kb apart from the lower bound of the window. For this particular region the same occurred when data were analysed as individual records. This mutation was the second strongest mutation for T3, explaining 16.21 % genetic variance and with a frequency of 0.46. The strongest mutation explained 22% of the genetic variance but its frequency was a bit lower (0.38).

Table 1.- Declared QTL regions, and position and effects of the actual mutations. Includes results for the analysis when both traits were individual records, and those obtained when T3 data corresponded to pooled records.

MUTATION LOCATION AND EFFECT					**DECLARED QTLs**			
CHR	**TRAIT**	**Location(Mb)**	**Freq**	**Effect**[1]	**W. ini. (Mb)** [2]	**W. fin. (Mb)** [2]	**WPPA**[2]	**Min. D. from W. (Kb)** [2]
		T1 Individual records - T3 Individual records						
1	1	84.025	0.46	7.61	84.05	84.25	1.00	25
4	1	24.925	0.47	26.66	24.70	24.90	1.00	225
1	3	14.625	0.52	3.84	14.65	14.85	0.82	25
1	3	58.825	0.38	22.00	[3]57.95	58.5	0.99	875
1	3	84.025	0.46	16.21	84.05	84.25	1.00	25
2	3	79.175	0.74	15.42	78.95	79.25	1.00	WITHIN
3	3	2.175	0.24	6.49	2.00	2.20	1.00	WITHIN
3	3	36.825	0.73	13.52	36.75	36.95	1.00	WITHIN
4	3	85.525	0.71	2.57	85.25	85.50	0.85	275
		T1 Individual records - T3 pooled records						
1	1	14.625	0.52	8.69	14.45	14.70	0.75	WITHIN
1	1	84.025	0.46	7.61	84.05	84.25	1.00	25
4	1	24.925	0.47	26.66	24.75	24.95	1.00	WITHIN
1	3	84.025	0.46	16.21	84.05	84.25	0.83	25

[1] % mutation additive variance with respect to total additive variance
[2] W. ini.= Window initial bound; W. fin.= Window final bound; WPPA= Window Posterior Probability of Association; Min. D. from W.= Minimal distance from window bounds to the mutation position.
[3] Defined after merging two consecutives windows having both a WPPA of nearly 1.

In the proposed model, genomic breeding values are described by two different conditional densities $\mathrm{p}(\mathbf{u}|\mathbf{H},\mathbf{g}_0)\sim\mathrm{MNV}(\mathbf{0},\mathbf{g}_0\otimes\mathbf{H})$ and $\mathrm{p}(\mathbf{u}|\mathbf{M},\mathbf{a},\mathbf{r}_{\boldsymbol{\varepsilon}})\sim\mathrm{MNV}\left(\begin{matrix}\mathbf{Ma_1}\\\mathbf{Ma_3}\end{matrix},\mathbf{r}_{\boldsymbol{\varepsilon}}\otimes\mathbf{I}\right)$. For these two being compatible it has to be shown that after integrating with respect to the prior distributions of their respective parameters, the same marginal distributions would be kept. In the first case the marginal will be a multivariate student’s t distribution (Sorensen & Gianola, 2002). In the second, the integration with respect to $\mathbf{r}_{\boldsymbol{\varepsilon}}$ yield also a multivariate student’s t, while the integration of this with respect to a mixture of multivariate student’s t distribution (marginal prior of $\mathbf{a}$ in Bayes Cπ) (Habier *et al.*, 2011) will also result in a multivariate student’s t density. So, both conditional distributions of the genomic breeding values effects result in the same marginal, but hyper-parameters for both marginal has to be granted to be the same. In our case this is not the case because in the second hierarchical level genomic breeding values are assumed to be independent across traits, while in the first they are assumed to be correlated. Future work will explore the effect of this model incongruence.

Conclusions

The proposed multiple regression model for implementing genome-wise association studies performed satisfactory in terms of power of detection and control of false discoveries, i.e.

within the range of other proposed approaches considering individual records. An important power loose was observed when the model was applied to pooled data, but even in this situation one of the strongest mutation can still be detectable. Further research is needed in order to properly address some theoretical issues that might affect our results, as well as to explore the performance of the approach with real pooled data.

Acknowledgements

This study has been financed by the Spanish research project RTA2014-00015-C02-01 and Feed-a-Gene Project; the last one funded from the European Union's H2020 Programme under grant agreement no 633531.

List of References

<1 line>

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., Lawlor, T.J., 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93, 743–752.

Biscarini, F., Bovenhuis, H., and van Arendonk, J. A. M., 2008. Estimation of variance components and prediction of breeding values using pooled data. J. Anim. Sci. 86:2845-2852.

Fernando R.L., Toosi A., Garrick D.J. and Dekkers J.C.M, 2014. Application of whole-genome prediction methods for genome-wise association studies: a Bayesian approach. Proceedings of the 10th world congress on genetics applied to livestock production. Vancouver, Canada.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12: 186.

Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee DH, 2002. BLUPF90 and Related Programs (BGF90). Proceedings of the 7th world congress on genetics applied to livestock production. Montpellier, France.

Sánchez J.P, Ramon J., Rafel, O., Ragab M., Piles M., 2016. Using collective feed intake data to select for feed efficiency on full or restricted feeding regimen. Proceedings of the 11th world rabbit congress. Qingao, China.

Sánchez J.P, Piles M, Tulsà M, Reixach J, Quintanilla R., 2014. The Value of group records in predicting breeding values for individual feed intake in pigs. Proceedings of the 10th world congress on genetics applied to livestock production. Vancouver, Canada.

Sorensen D. and Gianola D., 2002. Likelihood, Bayesian and MCM Methods in Quantitative Genetics. Springer-Verlag New York.

Usai M.G., Gaspa G. , Macciotta N.P.P. , Carta A. , Casu S., 2014. XVIth QTLMAS: Simulated dataset and comparative analysis of submitted results for QTL mapping and genomic evaluation. BMC Proceedings 8(Suppl 5):S1

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423

Wang H, Misztal I, Aguilar I, Legarra A, Muir W. 2012.Genome-wide association mapping including phenotypes from relatives without genotypes. Genet Res (Camb) 94:73–83.