Genome-wide association mapping including phenotypes from relatives without genotypes

H. Wang, I. Misztal, I. Aguilar, Andres Legarra, W.M. Muir

Genet. Res., Camb. (2012), **94**, *pp.* 73–83. © Cambridge University Press 2012
doi:10.1017/S0016672312000274

Genome-wide association mapping including phenotypes from relatives without genotypes

H. WANG[1]*, I. MISZTAL[1], I. AGUILAR[2], A. LEGARRA[3] AND W. M. MUIR[4]
[1] *Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602-2771, USA*
[2] *Instituto Nacional de Investigación Agropecuaria, INIA Las Brujas, 90200 Canelones, Uruguay*
[3] *INRA, UR631 Station d'Amélioration Génétique des Animaux (SAGA), BP 52627, 32326 Castanet-Tolosan, France*
[4] *Department of Animal Science, Purdue University, West Lafayette, IN 47907-1151, USA*

(*Received 19 September 2011; revised 8 December 2011, and 9 March 2012; accepted 13 March 2012*)

Summary

A common problem for genome-wide association analysis (GWAS) is lack of power for detection of quantitative trait loci (QTLs) and precision for fine mapping. Here, we present a statistical method, termed single-step GBLUP (ssGBLUP), which increases both power and precision without increasing genotyping costs by taking advantage of phenotypes from other related and unrelated subjects. The procedure achieves these goals by blending traditional pedigree relationships with those derived from genetic markers, and by conversion of estimated breeding values (EBVs) to marker effects and weights. Additionally, the application of mixed model approaches allow for both simple and complex analyses that involve multiple traits and confounding factors, such as environmental, epigenetic or maternal environmental effects. Efficiency of the method was examined using simulations with 15 800 subjects, of which 1500 were genotyped. Thirty QTLs were simulated across genome and assumed heritability was 0·5. Comparisons included ssGBLUP applied directly to phenotypes, BayesB and classical GWAS (CGWAS) with deregressed proofs. An average accuracy of prediction 0·89 was obtained by ssGBLUP after one iteration, which was 0·01 higher than by BayesB. Power and precision for GWAS applications were evaluated by the correlation between true QTL effects and the sum of m adjacent single nucleotide polymorphism (SNP) effects. The highest correlations were 0·82 and 0·74 for ssGBLUP and CGWAS with $m=8$, and 0·83 for BayesB with $m=16$. Standard deviations of the correlations across replicates were several times higher in BayesB than in ssGBLUP. The ssGBLUP method with marker weights is faster, more accurate and easier to implement for GWAS applications without computing pseudo-data.

1. Introduction

As a result of commercial availability of highly dense single nucleotide polymorphism (SNP) chips in humans, genome-wide association analysis (GWAS) has proven to be a powerful tool to identify genes for common diseases and complex traits (Hirschhorn & Daly, 2005; Visscher *et al.*, 2007). Similarly, GWAS has been applied to animals for the discovery of genes that are associated with disease and production traits (Karlsson *et al.*, 2007; Bennett *et al.*, 2010; Bolormaa *et al.*, 2010; Orr *et al.*, 2010; Pryce *et al.*, 2010). In animal breeding, a closely related procedure that makes use of the same SNP chips, but for an entirely different purpose, is the genomic estimation of breeding values (GEBVs) for genomic selection (GWMAS), a form of marker-assisted selection. GWMAS is often performed with procedures called BayesA or BayesB that consider all genetic associations derived from markers (Meuwissen *et al.*, 2001). Moreover, BayesA and BayesB solutions provide SNP effects; thus, these methods can be applied to GWAS (Goddard & Hayes, 2009; Sun *et al.*, 2011) with the additional advantage of accounting for population stratification and cryptic relatedness (Sillanpaa, 2011). The classical GWAS (CGWAS) is based on a test of a single marker, which treats each SNP marker as a covariate in the model (Hirschhorn & Daly, 2005). The main advantage of CGWAS is the ease of significance testing; however, it is likely to result in reduced fit to the data compared with methods where all SNPs are jointly considered. Additionally, neither

* Corresponding author: 425 River Road, Athens, GA, 30602-2771, USA. E-mail: huiyu@uga.edu

Bayesian methods nor single-marker analysis can directly include genetic association found in the pedigree of animals that have not been genotyped. Although such information can be considered indirectly in multiple-step procedures in which phenotypic data from relatives are summarized to create pseudo-data for genotyped individuals (VanRaden *et al*., 2009), new problems can arise, such as information loss, heterogeneity caused by different amounts of information in the original dataset and bias (Vitezica *et al*., 2011). Thus, multiple-step methods for computing genomic predictions are not only complicated but likely suboptimal for GWAS. This is particularly true in livestock species, where pedigrees are complex, and nuclear families are the exception rather than the rule. In contrast Misztal *et al*. (2009) and Christensen & Lund (2010) proposed a single-step GBLUP (ssGBLUP) that integrates phenotypes, genotypes and pedigree information. Such information can be combined with genomic data for greater detection power and estimation precision through a properly scaled and augmented relationship matrix (Legarra *et al*., 2009; Misztal *et al*., 2009). The ssGBLUP method has been shown to provide more consistent solutions and better accuracy than the multiple-step approach (Aguilar *et al*., 2010; Chen *et al*., 2011; Forni *et al*., 2011).

A limitation of the ssGBLUP methodology is that it is based on an infinitesimal model, which assumes equal variance for all SNP marker-QTL associated effects. An advantage of the infinitesimal model is that the resulting genomic relationship matrix is identical for all traits within a population (Aguilar *et al*., 2010). In contrast, although BayesA or BayesB is limited in that neither can include phenotypic information from non-genotyped individuals, they remove the assumption of equal variance for all SNP marker-QTL associated effects, which appears to be a more realistic situation. Unfortunately, relaxing this assumption comes at a cost of orders of magnitude more computing time in a Bayesian framework. Combining the strengths of both methods (i.e. allowing for unequal variances in an ssGBLUP context) could improve the accuracy of the estimation of GEBVs for breeding and selection applications, and precision for the estimation of SNP effects for GWAS applications.

Estimation of weights for SNP variances can be achieved without sampling. Zhang *et al*. (2010) derived SNP weights as functions of squares of SNP effects and incorporated those variances as weights in GBLUP. Sun *et al*. (2011) developed an iterative procedure for GBLUP, in which GEBVs were converted to SNP effects and weights were obtained similar to those in Zhang *et al*. (2010). However, neither study could directly utilize phenotypes of ungenotyped animals.

The objectives of this research were to investigate the optimal weights on marker variances for improving accuracy and precision in GWAS and GEBVs by ssGBLUP, and to compare results from ssGBLUP, CGWAS and the BayesB methods as described by Meuwissen *et al*. (2001).

2. Materials and methods

(i) *Data simulation*

Data were simulated using QMSim (Sargolzaei & Schenkel, 2009) for an additive trait with a mean of 5·0, phenotypic variance 1·0 and heritability 0·5. Two 100 cM chromosomes were simulated, with each chromosome containing 15 uniformly distributed QTLs. For chromosome 1 and chromosome 2, on average 1552 and 1448 SNP markers, respectively, were evenly distributed. Both SNP markers and QTLs were assumed to be bi-allelic, and no marker loci overlapped with the QTLs. Minor allele frequencies were $>0{\cdot}05$. Effects of QTLs were randomly sampled from a Gamma distribution with a shape factor of 0·4 and a scale factor of 1·36. All additive genetic variance resulted from the QTLs. A simulated population started at generation 1001 (i.e. base population) and consisted of 100 individuals. For generations, 1001 to 1, mutation rate of 0·000025 was simulated for each locus of both QTLs and SNPs per generation, and non-overlapping generations were simulated with population size per generation increasing gradually from 100 to 2800. In generations 0–4, 80 randomly chosen males and 520 randomly chosen females were genotyped and produced 2600 progenies by random mating. The phenotypic information was recorded for all animals in generations 0–5. Genotypes were recorded for all parents in generations 3 and 4, and 300 random individuals in generation 5. For recent generations 0–5, the complete datasets contained 15 800 individuals in pedigree with records, of which 1500 individuals were genotyped. The simulation was replicated ten times. Some statistics of the simulated dataset are shown in Table 1.

(ii) *Model and methodology*

The single-trait model for ssGBLUP was

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_a a} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{y}$ is a vector of simulated observations (phenotypes), $\mathbf{1}$ is a vector of all ones, μ is the overall mean of phenotypic records, $\boldsymbol{Z_a}$ is an incidence matrix that relates individuals to phenotypes, $\boldsymbol{a}$ is a vector of individual animal effects and $\boldsymbol{e}$ is a vector of residuals. The variances of $\boldsymbol{a}$ and $\boldsymbol{e}$ are

$$\mathrm{var}\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}\boldsymbol{H}\sigma_a^2 & 0\\ 0 & \boldsymbol{I}\sigma_e^2\end{bmatrix}, \tag{2}$$

Table 1. *Description of genomic data from simulation*

Means (SDs*)		Chr1†	Chr2	Total
SNPs	Number	1552 (22)	1448 (22)	3000
	AvgMAF‡	0·28 (0·004)	0·28 (0·005)	0.28 (0·005)
QTLs	Number	16 (2)	14 (2)	30
	AvgEffect§	0·15 (0·04)	0·16 (0·04)	0·16 (0·04)

*SDs: standard deviations.
†Chr1 and Chr2: chromosome 1 and chromosome 2.
‡Average minor allele frequencies of SNPs.
§Average effects of QTLs.

Table 2. *Correlations* (*SDs*) *between TBVs from simulation with EBVs and DP from regular BLUP, GEBVs from ssGBLUP and from BayesB with non-weighted and weighted* (c = *0·1*) *DP*

	EBVs	DP						
BLUP	0·81 (0·01)	0·77 (0·01)						
	it1*	it2	it3	it4	it5	it6	it7	it8
SsGBLUP	0·87 (0·01)	0·89 (0·01)	0·88 (0·01)	0·88 (0·02)	0·88 (0·02)	0·87 (0·02)	0·87 (0·02)	0·87 (0·02)
	NW†	$c = 0{\cdot}1$						
BayesB_DP	0·88 (0·02)	0·88 (0·02)						

*GEBV solutions using ssGBLUP from iteration 1 (it1) to iteration 8 (it8).
†Non-weighted DP, and weighted DP with $c = 0{\cdot}1$.

where σ_a^2 and σ_e^2 are total genetic additive and residual variances, respectively, and $\boldsymbol{H}$ is a matrix that combines pedigree and genomic relationships as in Aguilar *et al.* (2010), and its inverse is

$$\boldsymbol{H}^{-1} = \boldsymbol{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{G}^{-1} - \boldsymbol{A}_{22}^{-1} \end{bmatrix}, \quad (3)$$

where $\boldsymbol{A}$ is a numerator (pedigree) relationship matrix for all animals; $\boldsymbol{A}_{22}$ is a numerator relationship matrix for genotyped animals; and $\boldsymbol{G}$ is a genomic relationship matrix. Matrix $\boldsymbol{G}$ was constructed based on VanRaden *et al.* (2009) that assumed allele frequencies of the current population and adjusted for compatibility with $\boldsymbol{A}_{22}$, which was applied in 'GC' and 'BLUP$_\alpha$' in Chen *et al.* (2011) and Vitezica *et al.* (2011).

(iii) *Derivation of SNP effects from breeding values*

Let the animal effects be decomposed into those for genotyped ($\boldsymbol{a}_g$) and ungenotyped ($\boldsymbol{a}_n$) animals. The animal effects of genotyped animals are a function of SNP effects:

$$\boldsymbol{a}_g = \boldsymbol{Zu}, \quad (4)$$

where $\boldsymbol{Z}$ is a matrix relating genotypes of each locus and $\boldsymbol{u}$ is a vector of SNP marker effects.

Thus, the variance of the animal effects is

$$\mathrm{var}(\boldsymbol{a}_g) = \mathrm{var}(\boldsymbol{Zu}) = \boldsymbol{ZDZ}'\sigma_u^2 = \boldsymbol{G}^*\sigma_a^2, \quad (5)$$

where $\boldsymbol{D}$ is a diagonal matrix of weights for variances of SNPs ($\boldsymbol{D} = I$ for GBLUP), σ_u^2 is the genetic additive variance captured by each SNP marker when no weights are present and $\boldsymbol{G}^*$ is the weighted genomic relationship matrix.

The joint (co)variance of animal effects ($\boldsymbol{a}_g$) and SNP effects ($\boldsymbol{u}$) is

$$\mathrm{var}\begin{bmatrix} \boldsymbol{a}_g \\ \boldsymbol{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{ZDZ}' & \boldsymbol{ZD}' \\ \boldsymbol{DZ}' & \boldsymbol{D} \end{bmatrix}\sigma_u^2, \quad (6)$$

subsequently

$$\boldsymbol{G}^* = \frac{\mathrm{var}(\boldsymbol{a}_g)}{\sigma_a^2} = \frac{\mathrm{var}(\boldsymbol{Zu})}{\sigma_a^2} = \boldsymbol{ZDZ}'\frac{\sigma_u^2}{\sigma_a^2} = \boldsymbol{ZDZ}'\lambda, \quad (7)$$

where λ is a variance ratio or a normalizing constant. According to VanRaden *et al.* (2009),

$$\lambda = \frac{\sigma_u^2}{\sigma_a^2} = \frac{1}{\sum_{i=1}^{M} 2p_i(1-p_i)},$$

Table 3. *Average correlations* (*SDs*) *between QTL effects and sum of cluster of* m *SNP effects using ssGBLUP*

S1*	1†	2	4	8	16	40
it1	0·53 (0·07)	0·68 (0·05)	0·79 (0·03)	0·81 (0·02)	0·80 (0·03)	0·62 (0·08)
it2	0·46 (0·07)	0·66 (0·05)	0·78 (0·02)	0·82 (0·02)	0·81 (0·02)	0·63 (0·08)
it3	0·43 (0·07)	0·64 (0·05)	0·77 (0·02)	0·81 (0·02)	0·80 (0·02)	0·62 (0·08)
it4	0·42 (0·07)	0·63 (0·05)	0·77 (0·02)	0·81 (0·02)	0·80 (0·02)	0·62 (0·08)
it5	0·41 (0·07)	0·63 (0·05)	0·76 (0·02)	0·80 (0·02)	0·79 (0·02)	0·61 (0·08)
it6	0·41 (0·07)	0·62 (0·05)	0·75 (0·02)	0·80 (0·02)	0·79 (0·02)	0·61 (0·07)
it7	0·41 (0·07)	0·62 (0·05)	0·75 (0·02)	0·80 (0·02)	0·79 (0·02)	0·61 (0·07)
it8	0·41 (0·07)	0·62 (0·05)	0·75 (0·02)	0·80 (0·02)	0·79 (0·02)	0·60 (0·07)
S2	1	2	4	8	16	40
it1	0·53 (0·07)	0·68 (0·05)	0·79 (0·03)	0·81 (0·02)	0·80 (0·03)	0·62 (0·08)
it2	0·44 (0·09)	0·65 (0·06)	0·77 (0·03)	0·82 (0·03)	0·81 (0·02)	0·63 (0·06)
it3	0·41 (0·08)	0·62 (0·05)	0·75 (0·03)	0·79 (0·03)	0·79 (0·03)	0·65 (0·06)
it4	0·40 (0·07)	0·61 (0·05)	0·73 (0·03)	0·77 (0·03)	0·78 (0·03)	0·64 (0·06)
it5	0·40 (0·07)	0·60 (0·05)	0·72 (0·04)	0·76 (0·04)	0·77 (0·04)	0·64 (0·06)
it6	0·40 (0·07)	0·60 (0·05)	0·72 (0·04)	0·75 (0·04)	0·76 (0·04)	0·63 (0·06)
it7	0·40 (0·07)	0·60 (0·05)	0·72 (0·04)	0·75 (0·04)	0·76 (0·04)	0·63 (0·06)
it8	0·40 (0·07)	0·60 (0·05)	0·71 (0·04)	0·75 (0·04)	0·76 (0·04)	0·63 (0·06)

*S1: update weights for SNP effects but not for GEBVs; S2: update weights for both GEBVs and SNP effects in each iteration.
†Number of SNPs (i.e. *m* ranges from 1 to 40) in each cluster.

Table 4. *Average correlations* (*SDs*) *between QTL effects and sum of cluster of* m *SNP effects using BayesB and WOMBAT*

	BayesB		WOMBAT
Item*	NW†	$c=0{\cdot}1$	NW
1‡	0·48 (0·27)	0·47 (0·25)	0·57 (0·14)
2	0·65 (0·16)	0·64 (0·16)	0·68 (0·11)
4	0·78 (0·11)	0·78 (0·10)	0·73 (0·08)
8	0·82 (0·08)	0·82 (0·08)	0·74 (0·07)
16	0·82 (0·07)	0·83 (0·07)	0·73 (0·05)
40	0·66 (0·21)	0·67 (0·21)	0·63 (0·09)

*DP used as DV in BayesB and classical GWAS using WOMBAT.
†Non-weighted DP and weighted DP with $c=0{\cdot}1$.
‡Number of SNPs (i.e. *m* ranges from 1 to 40) in each cluster.

where M is the number of SNPs and p_i is the allele frequency of the second allele of the ith marker. Following Stranden & Garrick (2009) one can derive

$$\hat{\boldsymbol{u}} = \frac{\sigma_u^2}{\sigma_a^2}\boldsymbol{D}\boldsymbol{Z}'\boldsymbol{G}^{*-1}\hat{\boldsymbol{a}}_g. \quad (8)$$

Therefore, the equation for predicting SNP effects which uses weighted genomic relationship matrix $\boldsymbol{G}^*$ becomes

$$\hat{\boldsymbol{u}} = \lambda\boldsymbol{D}\boldsymbol{Z}'\boldsymbol{G}^{*-1}\hat{\boldsymbol{a}}_g = \boldsymbol{D}\boldsymbol{Z}'[\boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}']^{-1}\hat{\boldsymbol{a}}_g. \quad (9)$$

This is the best predictor of SNP effects given animal effects (Henderson, 1973). Estimates of SNP effects can be used to estimate individual variance of each SNP effect (Zhang *et al.,* 2010):

$$\hat{\sigma}^2_{u,i} = \hat{u}_i^2 2p_i(1-p_i). \quad (10)$$

(iv) *Computing algorithm*

The above formulae can be used to create an algorithm for estimation of $\boldsymbol{D}$ from ssGBLUP. Denote t as an iteration number and i as the ith SNP. The algorithm proceeds as follows:

1. $t=0$, $\boldsymbol{D}_{(t)}=\boldsymbol{I}$;$\boldsymbol{G}^*_{(t)}=\boldsymbol{Z}\boldsymbol{D}_{(t)}\boldsymbol{Z}'\,\lambda$.
2. Compute $\hat{\boldsymbol{a}}_g$ by ssGBLUP.
3. Calculate $\hat{\boldsymbol{u}}_{(t)}=\lambda\boldsymbol{D}_{(t)}\boldsymbol{Z}'\boldsymbol{G}^{*-1}_{(t)}\hat{\boldsymbol{a}}_{\mathbf{g}}$.
4. Calculate $d^*_{i_{(t+1)}}=\hat{u}^2_{i_{(t)}}2p_i(1-p_i)$ for all i as in Zhang *et al.* (2010).
5. Normalize $\boldsymbol{D}_{(t+1)}=\dfrac{\mathrm{tr}(\boldsymbol{D}_{(\mathbf{0})})}{\mathrm{tr}(\boldsymbol{D}^*_{(t+1)})}\boldsymbol{D}^*_{(t+1)}$.
6. Calculate $\boldsymbol{G}^*_{(t+1)}=\boldsymbol{Z}\boldsymbol{D}_{(t+1)}\boldsymbol{Z}'\lambda$.
7. $t=t+1$.
8. Exit, or loop to step 2 or 3.

In looping to step 3 (scenario S1), one applies the revised $\boldsymbol{G}^*$ only for the prediction of SNP effects, while calculating the animal effects only once, thus $\hat{\boldsymbol{a}}_g$ does not change during iterations. In looping to step 2 (scenario S2), both animal and SNP effects are recomputed. Whether scenario S1 is sufficient as opposed to scenario S2 and how much iterations are necessary is not clear and needs to be determined experimentally. In particular, scenario S1 is applicable to multiple-trait models where the relationship matrix needs to be identical for all traits.

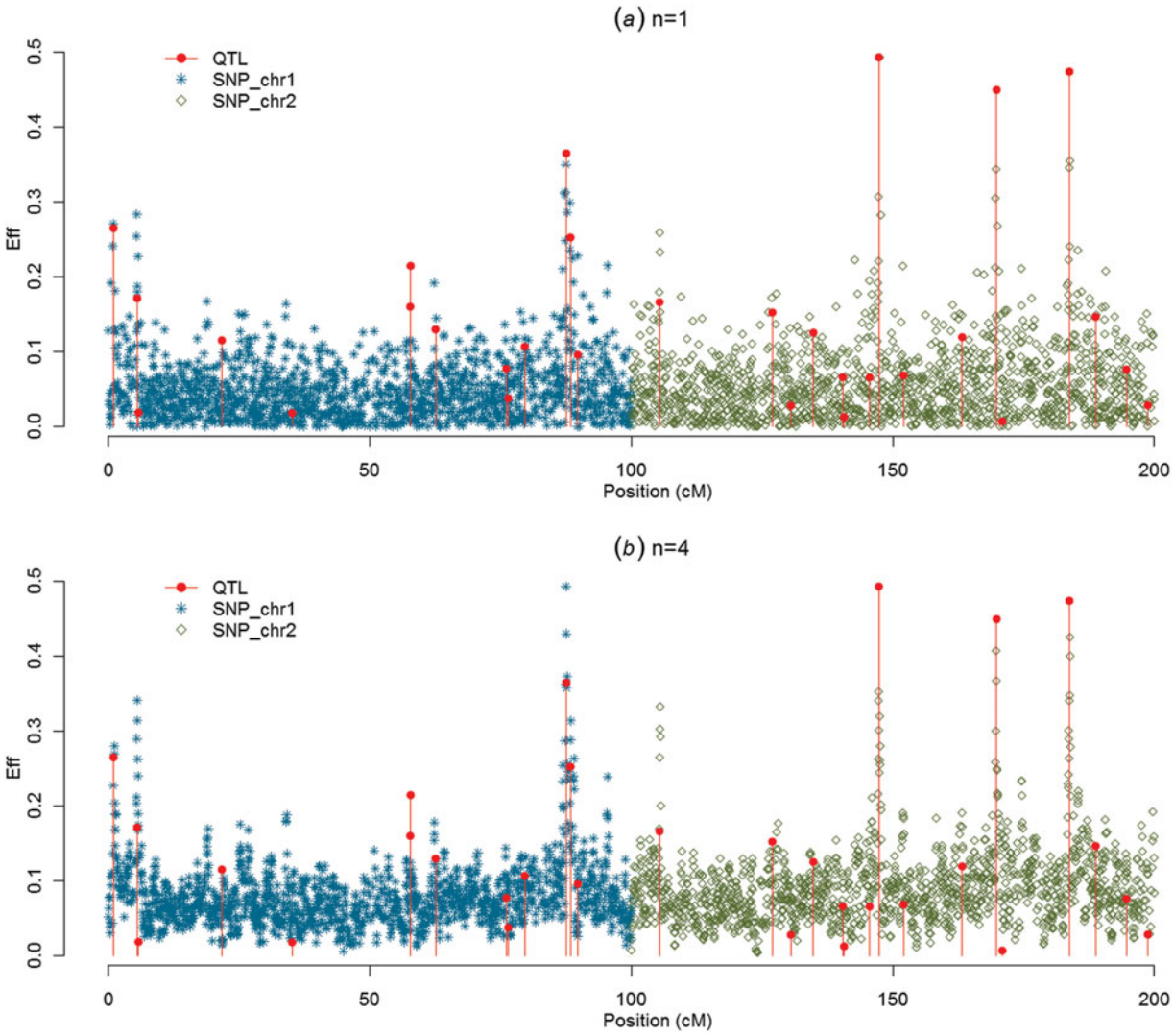

Fig. 1. SNP solutions and their four-point moving averages from ssGBLUP/S1 and ssGBLUP/S2 in the first iteration: (*a*) SNP solutions and (*b*) four-point moving average.

(v) *Computations*

Computations with ssGBLUP involved program BLUPF90 (Misztal *et al*., 2002) modified for genomic analyses (Aguilar *et al*., 2010), and used simulated parameters. Comparisons involved BayesB procedure as implemented in the GenSel package (Habier *et al*., 2010). These procedures used the model:

$$\tilde{\boldsymbol{y}} = \mathbf{1}\mu + \boldsymbol{Z_u u} + \boldsymbol{e}, \tag{11}$$

where $\tilde{\boldsymbol{y}}$ is a dependent variable (DV) for genotyped animals, with options being non-weighted deregressed proofs (DP) or weighted DP (Stranden & Garrick, 2009). For non-weighted DP, all weights were assumed equal to each other being 1; for weighted DP, the weight for *i*th individual was calculated as

$$w_i = \frac{(1-h^2)}{[c+(1-r_i^2)/r_i^2]h^2}$$

based on equation (10) in Garrick *et al*. (2009), where c is the fraction of the genetic variance not accounted for by SNPs, and was assumed to be 0·1 (Ostersen *et al*., 2011), h^2 is the heritability and r_i^2 is reliability of DP for the *i*th individual. Moreover, $\mathbf{1}$ is a vector of all ones, μ is the overall mean, $\boldsymbol{Z_u}$ is a matrix relating SNP marker effects to phenotypic information, $\boldsymbol{u}$ is a vector of SNP marker effects, $\boldsymbol{e}$ is a vector of residuals distributed as $N(0, \boldsymbol{D_b}\sigma_e^2)$, where $\boldsymbol{D_b}$ is a vector of weights as in Stranden & Garrick (2009). For BayesB, marker effects were assumed to be distributed as $u_j \sim N(0, \sigma^2_{u_j})$, where $\sigma^2_{u_j}$ is the variance of the *j*th SNP, and the proportion of SNPs with no effects ($\sigma^2_{u_j} = 0$) was set to 90%. As for ssGBLUP, the total genetic variance of BayesB methods was equal to the simulated value of 0·5. Priors for variances of SNP effects and residuals followed a scaled inverse Chi-square distribution with degrees of freedom 4 and 10, respectively. The Monte Carlo Markov Chain was run

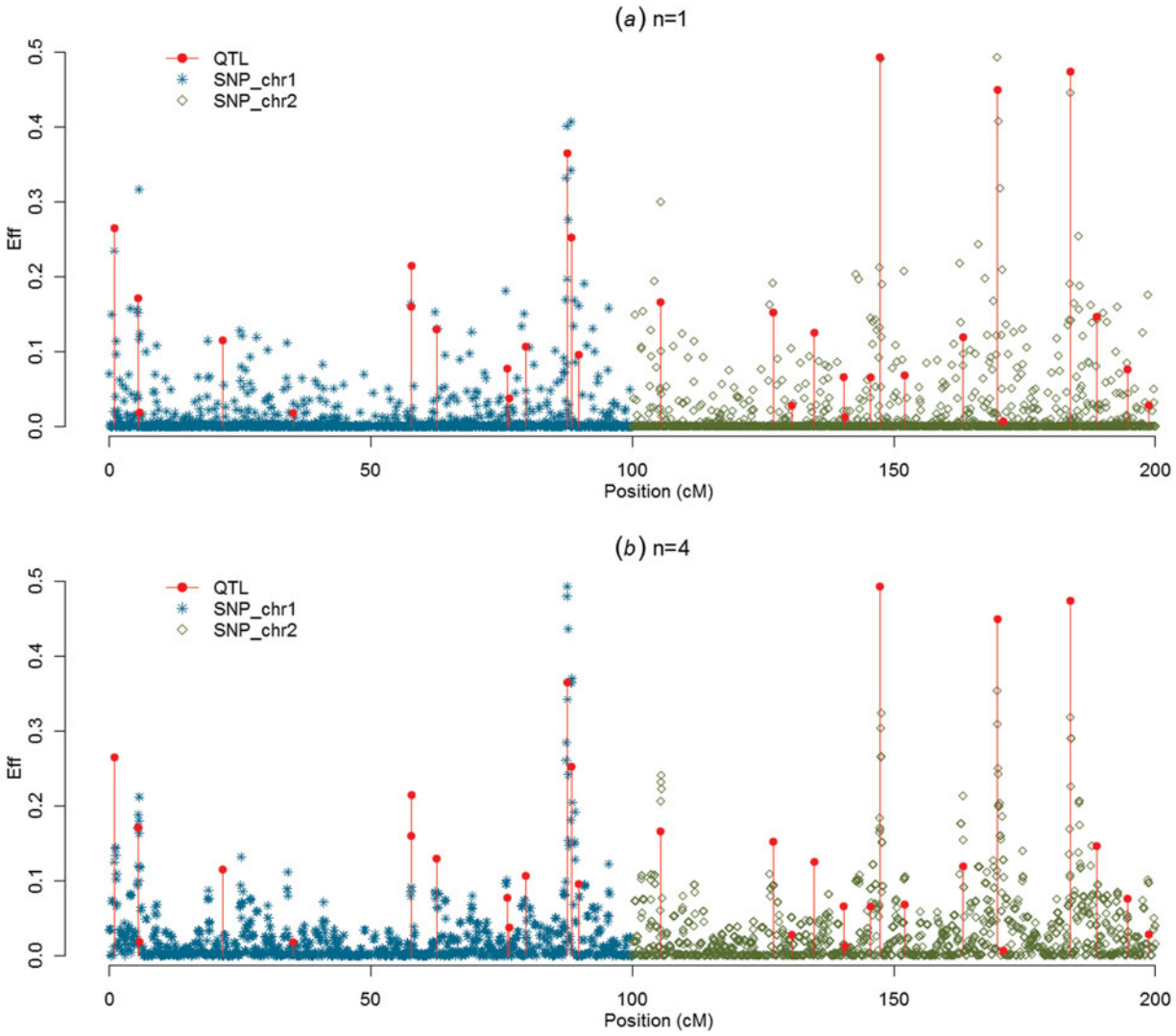

Fig. 2. SNP solutions and their four-point moving averages from ssGBLUP/S1 in the third iteration: (*a*) SNP solutions and (*b*) four-point moving average.

for 100 000 iterations (first 10 000 rounds were discarded as burn-in) with Gibbs sampling, with 100 of Metropolis–Hastings sampling within each Gibbs sampling cycle. Estimates of GEBVs and SNP effects were based on the posterior means according to the remaining 90 000 iterations. Accuracies of genotyped animals were defined as correlations between true breeding values (TBVs) and GEBVs. Accuracy of GWAS was determined by correlations of QTL effects with the sum of m SNP solutions adjacent to each QTL, where m varied from 1 to 40. We did not attempt to declare detection thresholds, or P-values, because they are difficult to define and compare with classical frequentist test of hypothesis, in the context of shrunken or Bayesian estimators, as is the case here (Servin & Stephens, 2007; Wakefield, 2009).

For comparisons, SNP solutions were also estimated by CGWAS using a 'Snappy' approach implemented in WOMBAT (Meyer & Tier, 2012). When CGWAS analyses are repeated for a large number of SNPs, the computing time can be large, especially for large SNP panels. In 'Snappy', matrices common to all SNPs are pre-computed, greatly reducing the computation time for the complete scan.

3. Results and discussion

(i) *Accuracy of estimated breeding values* (*EBVs*)

EBVs had been obtained through regular BLUP, ssGBLUP and Bayesian methods (BayesB using non-weighted or weighted DP), respectively. Accuracies of genotyped animals are shown in Table 2, and defined as correlations between TBVs and EBVs: EBVs for regular BLUP and GEBVs for other approaches. Accuracies of ssGBLUP ranged from 0·87 (0·02) to 0·89 (0·01) depending on iterations, and they were always higher than EBVs for BLUP. For S1, accuracies of GEBVs remained 0·87 (0·01). This result occurred because GEBVs were not recomputed (only SNP effects). For S2, however, the accuracy increased

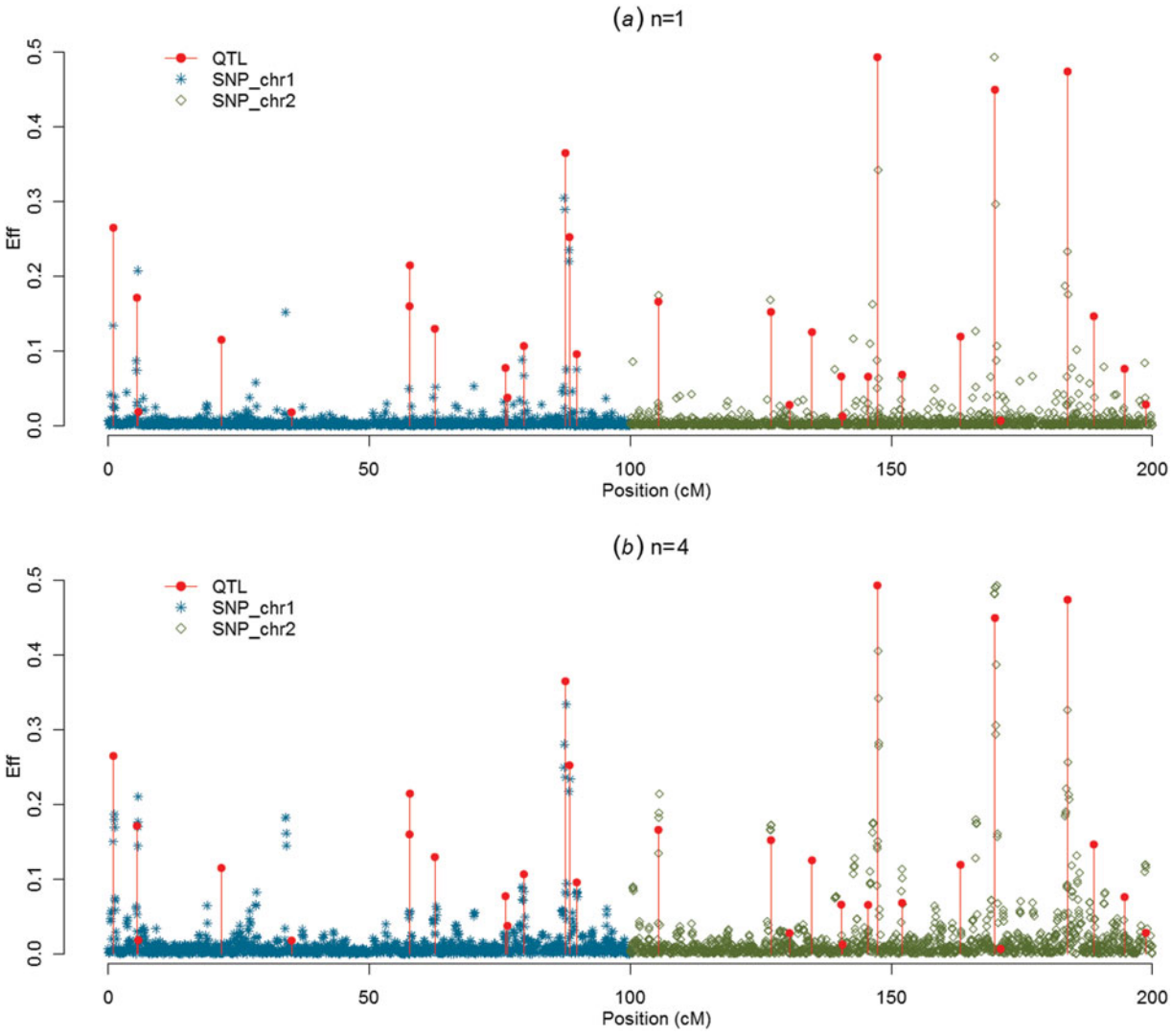

Fig. 3. SNP solutions and their four-point moving averages from BayesB with weighted DP ($c=0{\cdot}1$) as the DV: (*a*) SNP solutions and (*b*) four-point moving average.

to 0·89 (0·01) by the second round, then dropped to 0·88 (0·01 or 0·02) until the sixth round, and then dropped to 0·87 (0·02). The slight decrease of accuracy in the later rounds could be due to excessive weights given to SNPs associated with few QTLs with larger effects, and reduced weights for numerous QTLs with smaller effects.

For BayesB methods, the accuracies of non-weighted DP were 0·88 (0·02) and were the same as the result of weighted DP ($c=0{\cdot}1$). As using DP as DV yields more reliable breeding value solutions than using EBVs in genetic evaluation (Ostersen *et al.*, 2011), other types of DV (e.g. phenotypic records and EBVs) were not considered in this study. For both scenarios of using non-weighted and weighted DP as DV, accuracies from BayesB methods were similar to ssGBLUP with slightly larger standard deviations (SDs) across replications. Although the Bayesian methods lose accuracies when pseudo-data are used (Vitezica *et al.*, 2011) that loss of accuracy seems to be similar to the loss of accuracy in ssGBLUP by assuming variances of all SNPs are equal. In the work of Vitezica *et al.* (2011), genotyped animals do not have observations of their own, whereas here genotyped animals do have associated phenotypes. Therefore information from related animals added little to EBV accuracies.

(ii) *Accuracy of QTL estimates*

Table 3 presents accuracies of ssGBLUP for QTLs defined as correlations between QTL effects and the sum of m adjacent SNP marker effects, where m varied from 1 to 40. SNP effects under scenarios of S1 and S2 were updated iteratively resulting in similar results. For both S1 and S2, and all iterations, accuracies of QTLs increased from $m=1$ to $m=8$, and decreased sharply for $m=40$. Iterations improved the

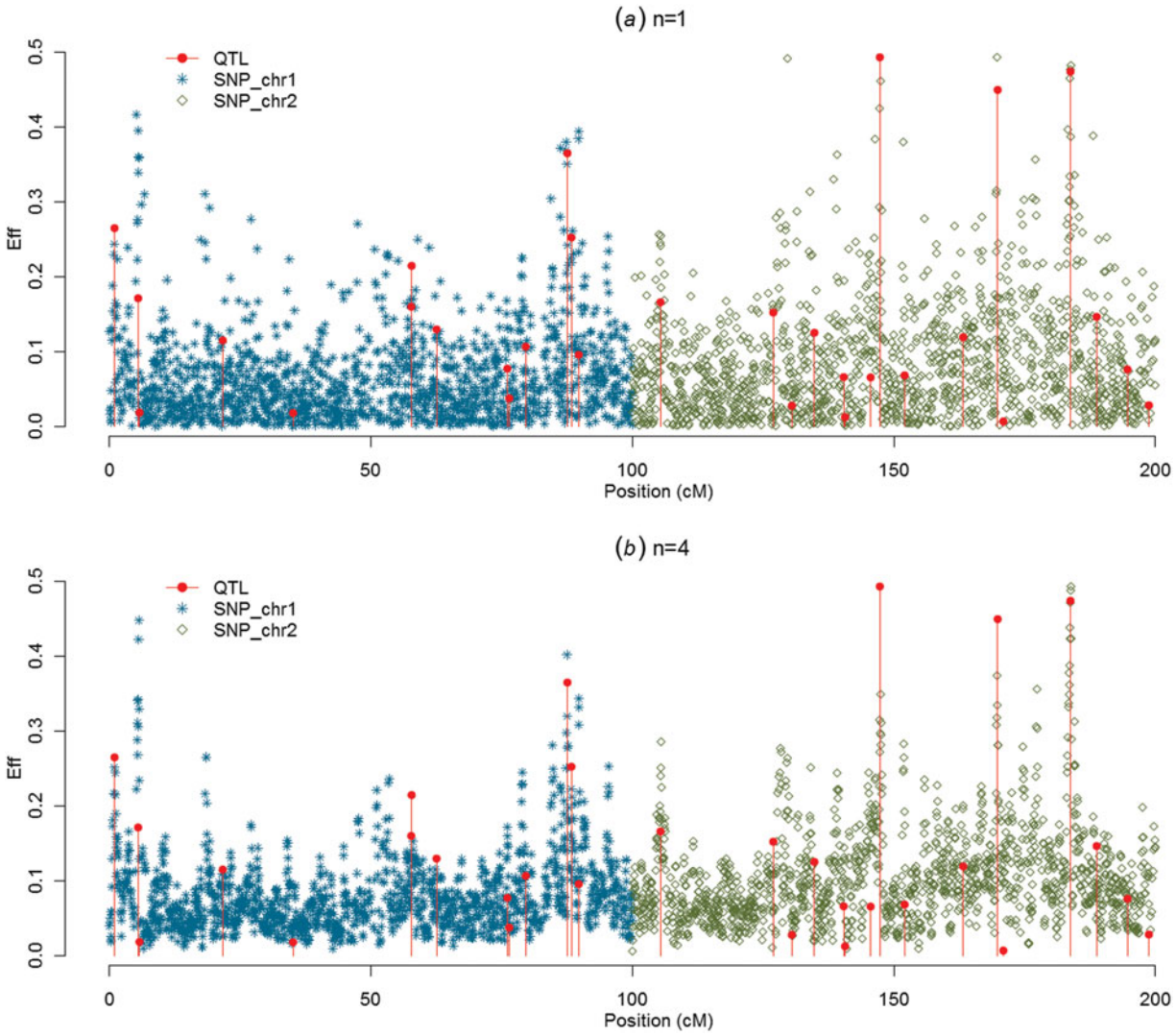

Fig. 4. SNP solutions and their four-point moving averages from r with non-weighted DP as the DV: (*a*) SNP solutions and (*b*) four-point moving average.

accuracy of the S1 and S2 options but only for $m=8$ and $m=16$. With iteration and subsequent re-computation of SNP weights, small SNP effects were reduced every round while the large effects became even larger. Iteration for new GEBVs (S2) allowed corrections to SNP with small effects. The highest correlation at $m=1$, 2 and 4 was after the first iteration. The highest correlation was 0·82 with $m=8$ and the second iteration. In both S1 and S2, iteration for GEBVs maximized the accuracy of GEBVs given weights. The highest accuracy was achieved by having a combination of weights that minimized estimation errors but reflected the reality that SNPs adjacent to a QTL contribute to estimation of that QTL.

The advantage of S2 over S1 is dependent on the number and distribution of QTL effects. With many QTL effects and relatively equal distribution, assigning differential weights to SNPs does not greatly improve the accuracy of GEBVs, and therefore little is gained by iteration on GEBVs. Greater improvements with S2 are expected when differential weights on SNP improve accuracy to a greater degree. In a separate study (results not reported), the realized accuracy of S2 improved up to the third iteration for some traits, while deteriorating for other traits in subsequent round. Further research may establish an optimum number of rounds for each particular situation.

Relatively lower correlations are not unexpected at low *m*. Zondervan & Cardon (2004) have found that the closest SNP marker is not always the best predictor of its neighbouring QTL. There should be an 'optimal haplotype length' according to the marker density and extent of linkage disequilibrium in the population (Villumsen *et al.*, 2009). Density of SNP markers and QTLs in the simulated genome was, on average, 0·067 and 6·06 cM, respectively. With each QTL distributed approximately every 90

SNP markers, the QTL effects could be best approximated by the sum of the adjacent 90 SNP effects. However, due to recombination and mutation for 1000 generations, the best haplotype length can be much shorter than expected. In this study, approximations with eight SNPs were the most accurate, while those with close to 40 or more were not. Decreases of accuracies in later iterations can be explained by excessive weights on larger SNPs in later iterations. A different algorithm to calculate weights of SNP effects, e.g. with a lower bound similar to Sun *et al*. (2011), may improve accuracies in later rounds. The form of constructing weights used here is indeed suboptimal, because it considers that the estimate of the *j*th SNP effect $\hat{u}_j$ is the true value, whereas in fact it is a regressed value. An optimal procedure would consider the uncertainty in the estimation of SNP effects by expectation–maximization (EM) or by Bayesian procedures (Xu, 2010; Legarra *et al*., 2011).

Table 4 shows the correlation between QTL effects and the sum of *m* adjacent SNP solutions for BayesB using non-weighted or weighted DP and for CGWAS using non-weighted DP. When BayesB was applied, weighting DP had little effect on the correlations, which most likely was due to the simple population structure in our simulated study and subsequently similar weights for most genotyped animals. Compared with ssGBLUP and iteration 1, the correlations resulting from application of BayesB were smaller for $m \leqslant 4$ and slightly higher for $m \geqslant 16$. Although the average correlations using BayesB were the same as, or even slightly better than when ssGBLUP was used, the SDs calculated over 10 replications were much higher for BayesB than ssGBLUP. Even in the best situation, the SDs were 0·07 for BayesB with $m = 16$, as compared to 0·02 (or 0·03) for ssGBLUP/S1 (or S2) with $m = 8$. For other *m*, SDs ranged from 0·08 to 0·27 for BayesB, and from 0·02 to 0·09 for ssGBLUP. Larger SDs with BayesB could be due to its sampling structure (Gianola *et al*., 2009), which also made BayesB less robust than ssGBLUP. With CGWAS, the correlations were higher than any other methods with $m = 1$, matched ssGBLUP in iteration 1 with $m = 2$, and were lower than the other methods with $m \geqslant 4$. Due to fitting a single SNP as a fixed effect, CGWAS is best for identifying a single causative SNP, but seems less efficient in identifying regions containing the QTLs. In general, SDs with CGWAS were lower than with BayesB, but higher than with ssGBLUP.

(iii) *Graphs of SNP solutions and their moving averages*

Figures 1–4 present SNP solutions or their four-point moving averages for several methods. The graphs of SNP solutions are the least noisy for BayesB, and the noisiest for CGWAS, with ssGBLUP in between. While most SNP solutions in BayesB are set to 0, lack of shrinkage in CGWAS results in solutions with more noise. Solutions from the third iteration of ssGBLUP/S1 were more similar to those of BayesB, as each round of ssGBLUP shrinks smaller solutions. With averaging, graphs from all the methods were more similar, with closest similarity between BayesB and ssGBLUP/S1 in iteration 3, and CGWAS and ssGBLUP in iteration 1. The similarities confirm that for this particular dataset, most QTLs cannot be located with a single SNP accurately; however, all of the methods are similar in identifying regions containing large QTLs.

(iv) *Computing considerations*

In terms of computing time, one round of ssGBLUP required about 2 min, a run of BayesB required about 5 h, and a run of WOMBAT only required 13 s. Long running time in BayesB is due to long sampling. The extraordinarily fast run in WOMBAT is due to an ingenious algorithm; in testing, WOMBAT was over 100 times faster than previous CGWAS approaches. However, the timing analyses were not fully comparable. Both BayesB and ssGBLUP are useful for creating prediction equations based on computed SNP effects, while CGWAS is only useful for GWAS. Comparisons based on computing times are not complete, as BayesB and CGWAS require a BLUP run to create DP, but no such step is required with ssGBLUP.

When implemented efficiently, the cost of BayesB is linear for the number of SNPs and the number of subjects (Legarra & Misztal, 2008). As currently implemented, the creation of $\boldsymbol{G}^{-1}$ in ssGBLUP is linear with respect to the number of SNPs and cubic with respect to the number of subjects (Aguilar *et al*., 2011). With efficient implementation, the time to create $\boldsymbol{G}^{-1}$ is about 1 min for 7k genotypes and 1 h for 30k genotypes (Aguilar *et al*., 2011). The ssGBLUP method has a potential of smaller than cubic cost with respect to the number of genotypes with non-symmetric mixed model equations and preconditioned conjugate gradients (PCG) iteration (Misztal *et al*., 2009; Legarra & Ducrocq, submitted).

(v) *Additional considerations*

In practice, GWAS (as practiced in humans) seeks to find loci strongly associated across 'unrelated' individuals. Genomic selection works with closely related populations, and this relation generates strong linkage (disequilibrium) within the sample that cannot be ignored. As results from the three methods are similar, none of the methods do a particularly good job of distinguishing associations from that due to linkage disequilibrium. Additional analyses are

required to determine whether markers with large effects are due to associated loci or to linkage disequilibrium.

For the datasets in this study, in the best case, ssGBLUP delivered more accurate GEBVs than the best-case BayesB. All the methods delivered similar predictions of QTL effects based on the sum of 2-SNP effects. The ssGBLUP/S1 method is still relatively new and can benefit from further refinements. In particular, the refinements would involve more accurate sampling of SNP variances as discussed before, and a determination of the optimum number of rounds in ssGBLUP/S2 for maximum accuracy of GEBVs and GWAS. Another needed refinement for ssGBLUP is methodology for significance testing. Without such testing, the use of ssGBLUP for GWAS is limited to identifying SNPs or regions of SNPs with very large effects.

In general, ssGBLUP/S1 seems to provide more consistent estimates than either BayesB or CGWAS using DP. The ssGBLUP/S1 method is also much simpler and therefore more robust to run as: (i) no pseudo-data are required and (ii) no sampling is used. Mrode *et al*. (2010) found large differences regarding results and computing time among various implementations of BayesB.

Models used in this study were very simple with a relatively balanced population structure. For complicated models, such as a multi-trait, maternal effect, random regression or reaction norm models, DP are hard or near impossible to create. Even if they can be created, approximations of DP (Vitezica *et al*., 2011) would reduce accuracy. The performance of ssGBLUP is likely to improve with field data and more complex models with additional refinements.

4. Conclusions

The ssGBLUP method can be modified to compute SNP effects and estimate variances of SNP effects. Such modifications allow for increased accuracy of GEBVs and enable GWAS. The main advantage of ssGBLUP for GWAS is the ability to incorporate phenotypes of ungenotyped animals directly in a BLUP-like approach, without computing pseudo-data. Modified ssGBLUP may become the method of choice for GWAS in the case where merely a fraction of the population with phenotypes is genotyped. In which case, the model for analysis is too complex for use of other methods, and pseudo-data, such as DP, for use with method BayesB and CGWAS, cannot be obtained with sufficient accuracy. In addition, ssGBLUP has the advantages of fast computing, robust estimates and simplicity.

We acknowledge helpful discussions and pointing to the Zhang *et al*. (2010) study by R. L. Fernando. We used moving averages of SNP solutions following examples by J. Dekkers. This study was partially funded by the Holstein Association and Agriculture and Food Research Initiative grants 2009-65205-05665 and 2010-65205-20366 from the USDA National Institute of Food and Agriculture Animal Genome Program.

References

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. *Journal of Dairy Science* **93**, 743–752.

Aguilar, I., Misztal, I., Legarra, A. & Tsuruta, S. (2011). Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. *Journal of Animal Breeding and Genetics* **128**, 422–428.

Bennett, B. J., Farber, C. R., Orozco, L., Kang, H. M., Ghazalpour, A., Siemers, N., Neubauer, M., Neuhaus, I., Yordanova, R., Guan, B., Truong, A., Yang, W. P., He, A., Kayne, P., Gargalovic, P., Kirchgessner, T., Pan, C., Castellani, L. W., Kostem, E., Furlotte, N., Drake, T. A., Eskin, E. & Lusis, A. J. (2010). A high-resolution association mapping panel for the dissection of complex traits in mice. *Genome Research* **20**, 281–290.

Bolormaa, S., Pryce, J. E., Hayes, B. J. & Goddard, M. E. (2010). Multivariate analysis of a genome-wide association study in dairy cattle. *Journal of Dairy Science* **93**, 3818–3833.

Chen, C. Y., Misztal, I., Aguilar, I., Legarra, A. & Muir, W. M. (2011). Effect of different genomic relationship matrices on accuracy and scale. *Journal of Animal Science* **89**, 2673–2679.

Christensen, O. F. & Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. *Genetics Selection Evolution* **42**, 2.

Forni, S., Aguilar, I. & Misztal, I. (2011). Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genetics Selection Evolution* **43**, 1.

Garrick, D. J., Taylor, J. F. & Fernando, R. L. (2009). Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genetics Selection Evolution* **41**, 55.

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E. & Fernando, R. L. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics* **183**, 347–363.

Goddard, M. E. & Hayes, B. J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews Genetics* **10**, 381–391.

Habier, D., Fernando, R. L., Kizilkaya, K. & Garrick, D. J. (2010). Extension of the Bayesian alphabet for genomic selection. In The 9th World Congress on Genetics Applied to Livestock Production, p. 468. German Society for Animal Science, Leipzig, Germany.

Henderson, C. R. (1973). Sire evaluation and genetic trends. *Journal of Animal Science* **1973**, 10–41.

Hirschhorn, J. N. & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6**, 95–108.

Karlsson, E. K., Baranowska, I., Wade, C. M., Salmon Hillbertz, N. H., Zody, M. C., Anderson, N., Biagi, T. M., Patterson, N., Pielberg, G. R., Kulbokas, E. J., Comstock, K. E., Keller, E. T., Mesirov, J. P., von Euler, H., Kampe, O., Hedhammar, A., Lander, E. S., Andersson, G., Andersson, L. & Lindblad-Toh, K.

(2007). Efficient mapping of mendelian traits in dogs through genome-wide association. *Nature Genetics* **39**, 1321–1328.

Legarra, A., Aguilar, I. & Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science* **92**, 4656–4663.

Legarra, A. & Ducrocq, V. (submitted). Computational strategies for national integration of phenotypic, genomic and pedigree data in a single-step BLUP. *Journal of Dairy Science*.

Legarra, A. & Misztal, I. (2008). Technical note: Computing strategies in genome-wide selection. *Journal of Dairy Science* **91**, 360–366.

Legarra, A., Robert-Granie, C., Croiseau, P., Guillaume, F. & Fritz, S. (2011). Improved Lasso for genomic selection. *Genetics Research* **93**, 77–87.

Meuwissen, T. H., Hayes, B. J. & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829.

Meyer, K. & Tier, B. (2012). ''SNP Snappy'': a strategy for fast genome-wide association studies fitting a full mixed model. *Genetics* **190**, 275–277.

Misztal, I., Legarra, A. & Aguilar, I. (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *Journal of Dairy Science* **92**, 4648–4655.

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T. & Lee, D. H. (2002). BLUPF90 and related programs (BGF90). In The 7th World Congress Genetics Application Livestock Production, pp. 28, Montpellier, France.

Mrode, R., Coffey, M. P., Stradén, I., Meuwissen, T. H. E., Kaam, J. B. C. H. M. v., Kearney, J. F. & Berr, D. P. (2010). A comparison of various methods for the computation of genomic breeding values of dairy cattle using software at genomicselection.net. In the 9th World Congress on Genetics Applied to Livestock Production, Leipzig, Germany. Available at http://www.icbf.com/publications/files/WCGALP2010_RMrode.pdf

Orr, N., Back, W., Gu, J., Leegwater, P., Govindarajan, P., Conroy, J., Ducro, B., Van Arendonk, J. A., MacHugh, D. E., Ennis, S., Hill, E. W. & Brama, P. A. (2010). Genome-wide SNP association-based localization of a dwarfism gene in Friesian dwarf horses. *Animal Genetics* **41** (Suppl. 2), 2–7.

Ostersen, T., Christensen, O. F., Henryon, M., Nielsen, B., Su, G. & Madsen, P. (2011). Degressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs. *Genetics Selection Evolution* **43**(1), 38.

Pryce, J. E., Bolormaa, S., Chamberlain, A. J., Bowman, P. J., Savin, K., Goddard, M. E. & Hayes, B. J. (2010). A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. *Journal of Dairy Science* **93**, 3331–3345.

Sargolzaei, M. & Schenkel, F. S. (2009). QMSim: a large-scale genome simulator for livestock. *Bioinformatics* **25**, 680–681.

Servin, B. & Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genetics* **3**, e114.

Sillanpaa, M. J. (2011). Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. *Heredity* **106**, 511–519.

Stranden, I. & Garrick, D. J. (2009). Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *Journal of Dairy Science* **92**, 2971–2975.

Sun, X., Fernando, R. L., Garrick, D. J. & Dekkers, J. C. M. (2011). An iterative approach for efficient calculation of breeding values and genome-wide association analysis using weighted genomic BLUP. *Journal of Animal Science* **89** (E-Suppl 2), e11.

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F. & Schenkel, F. S. (2009). Invited review: Reliability of genomic predictions for North American Holstein bulls. *Journal of Dairy Science* **92**, 16–24.

Villumsen, T. M., Janss, L. & Lund, M. S. (2009). The importance of haplotype length and heritability using genomic selection in dairy cattle. *Journal of Animal Breeding and Genetics* **126**, 3–13.

Visscher, P. M., Macgregor, S., Benyamin, B., Zhu, G., Gordon, S., Medland, S., Hill, W. G., Hottenga, J. J., Willemsen, G., Boomsma, D. I., Liu, Y. Z., Deng, H. W., Montgomery, G. W. & Martin, N. G. (2007). Genome partitioning of genetic variation for height from 11,214 sibling pairs. *American Journal of Human Genetics* **81**, 1104–1110.

Vitezica, Z. G., Aguilar, I., Misztal, I. & Legarra, A. (2011). Bias in genomic predictions for populations under selection. *Genetics Research* **93**, 357–366.

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with *P*-values. *Genetic Epidemiology* **33**, 79–86.

Xu, S. (2010). An expectation–maximization algorithm for the Lasso estimation of quantitative trait locus effects. *Heredity* **105**, 483–494.

Zhang, Z., Liu, J., Ding, X., Bijma, P., de Koning, D. J. & Zhang, Q. (2010). Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PLoS One* **5**, e12648.

Zondervan, K. T. & Cardon, L. R. (2004). The complex interplay among factors that influence allelic association. *Nature Reviews Genetics* **5**, 89–100.