Extension to Haplotypes of Genomic Evaluation Algorithms

P. Croiseau,[1] M.N. Fouilloux,[2] D. Jonas,[1,3,4] S. Fritz,[1,4] A. Baur,[1,4] V. Ducrocq,[1] F. Phocas,[1] and D. Boichard[1]
[1]INRA UMR1313 GABI Jouy en Josas, [2]Institut de l'Elevage, Paris, [3]AgroParisTech, Paris, [4]UNCEIA, Paris, France

ABSTRACT: Most genomic evaluation methods have been based on SNP. It has been shown that differences in accuracy of genomic breeding values were small between these methods. The use of a high density chip instead of a medium density one in within breed or multi-breed genomic evaluation leads to relatively small improvements. Haplotypes instead of SNP improve linkage disequilibrium between alleles and potential quantitative trait loci. Here, an extension to haplotypes of GS3, a free selection genomic software developed by Legarra et al. (2013) is proposed.

Keywords: genomic selection; Bayes Cπ; haplotype.

INTRODUCTION

Since the theoretical basis of the genomic selection proposed by Meuwissen et al. (2001), various statistical methods were developed and tested to improve the precision of genomic estimated breeding values (GEBV - Gianola et al. (2009); Habier et al. (2011); Gianola (2013); Meuwissen et al. (2001)). A large panel of these methods has been tested on French data for the three dairy cattle breeds with national genomic evaluations (Croiseau et al. (2012); see Ducrocq et al., this congress) and there was no clear advantage of one method over the others in term of accuracy of genomic selection.

The challenge of genomic selection approaches has moved to the use of higher density chips or to the possibility for breeds with small reference populations to benefit from the reference population of main breeds through multi-breed genomic selection. Again, no clear advantage of one approach over the others was evidenced in such contexts (Hayes et al. (2009); Erbe et al. (2012); Hozé et al. (in press)).

In this paper, we advocate the use of haplotypes instead of SNP to improve the accuracy of genomic selection approaches. Calus et al. (2007) showed on simulated data that for high heritability traits, the use of haplotypes led to higher accuracies of genomic estimated breeding values. In real situations, genomic evaluations were implemented in France using a marker assisted-BLUP approach where haplotypes of SNP were used to trace QTL. In this approach, in spite of a relatively limited number of haplotypes (around 800), accuracies of GEBV were as high as (or very close to) those obtained with other tested methods (see Ducrocq et al., this congress). Therefore the use of haplotypes is appealing to capture QTL effects and could improve GEBV accuracy. To address this idea, we formally derive a haplotypic version of the GS3 software developed by Legarra et al. (2013). We describe here this approach.

METHODS

General model. Here "haplotype" will refer to the locus analyzed, and "allele" to the allelic form at this locus. Following the original model proposed by Habier et al. (2011), the haplotypic model can be written as:

$$y_i = \mu + u_i + \sum_{j=1}^{K} \delta_j \left(h_{ij}^p + h_{ij}^m \right) + e_i$$

where μ is a mean (or a vector of nuisance effects), u_i is the random polygenic effect for animal i, K is the number of haplotypes, h_{ij}^p and h_{ij}^m are the random effects of the paternal and maternal alleles of haplotype j of animal i, e_i is a random residual for animal i, and δ_j is an indicator variable equal to 0 (no effect) with probability π and equal to 1 with probability (1-π). Note that δ_j is applied to the whole haplotype. So, for a given haplotype, all the allele effects are estimated or all of them are set to zero. All the alleles of a given haplotype have the same variance.

Haplotypic GBLUP.

In the haplotypic GBLUP, δ_j is equal to 1 (all haplotypes are included). GEBV are directly estimated using an extension of the mixed model equations proposed by Van Raden (2008). The genomic relationship matrix $\boldsymbol{G}$ is formed as follows (Legarra, personal communication):

$$\boldsymbol{G} = \frac{(\boldsymbol{Z} - \boldsymbol{P})(\boldsymbol{Z} - \boldsymbol{P})'}{2\sum_{k=1}^{K}\sum_{l=1}^{l_k}\sum_{j \neq i} p_{ki} p_{kj}}$$

where l_k is the number of alleles of haplotype k, and p_{ki} is the allelic frequency of its i[th] allele. $\boldsymbol{P}$ is the matrix of repeated rows of allelic frequencies and $\boldsymbol{Z}$ is the incidence matrix of haplotype effects on individuals, with terms $z_{n,ki}$ equal to 0, 1, or 2 according to the number of alleles i of haplotype k carried by individual n.

This approach is believed to be more accurate than the traditional GBLUP based on SNP because haplotypes are more informative. Indeed, the same alleles of a haplotype carried by two individuals are more likely to be identical by descent than identical by state.

Haplotypic BLUP

Using an equivalent model, the haplotypic effects can be estimated solving the standard mixed model equations (MME), with l_k levels per haplotype. The haplotypic part of the MME can be written as

$$\left[\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{I}\right]\boldsymbol{h} = \boldsymbol{Z}'\boldsymbol{y}$$

with $\lambda = \sigma_e^2 / \sigma_h^2$ and σ_h^2 equals σ_g^2 / K

This system of equations can be solved with conventional methods.

MCMC Haplotypic BLUP

The haplotype effects of the previous model can be estimated by MCMC, with the following sampling process.

Let $\boldsymbol{y}_k^*$ denote the vector of phenotypes adjusted for all effects of the model except $\boldsymbol{h_k}$, $\boldsymbol{h_k}$ being the l_k-vector of allelic effects of haplotype k. $\boldsymbol{h_k}$ samples are generated as:

$$\boldsymbol{h_k} \sim MVN\left(\left[z_k' z_k + \lambda \boldsymbol{I}\right]^{-1} z_k' \boldsymbol{y}_k^*, \left[z_k' z_k + \lambda \boldsymbol{I}\right]^{-1} \sigma_e^2\right)$$

where MVN stands for the multivariate normal distribution.

Variance Estimation.

In the MCMC approach, a common haplotypic variance is assumed for all haplotypes. The residual and haplotypic variances are sampled as follows:

- $\sigma_e^2 \sim$ Inverted-χ^2 (*VarE, DfE*)

with $DfE = n+n_{ep}$ and $VarE = [\ \boldsymbol{e'e} / n + V_{ep}\, n_{ep}] / (n+n_{ep})$

V_{ep} is the a priori residual variance and n_{ep} the corresponding degree of belief.

- $\sigma_h^2 \sim$ Inverted-χ^2 (*VarH, DfH*)

with $DfH = K+n_{hp}$ and:

$$VarH = \frac{2\sum_{k=1}^{K}\sum_{i=1}^{l_k}\sum_{j\neq i} p_{ki}\left(h_{k,i} - \mu_k\right)^2 + V_{hp}\, n_{hp}}{K + n_{hp}}$$

μ_k is the allelic mean effect weighted by the allele frequencies. V_{hp} *is* the a priori haplotypic variance and n_{hp} *is* the corresponding degree of belief.

The additive genetic variance σ_g^2 can also be sampled as in the original version of GS3 based on SNP.

Haplotype Selection

Selection of a limited number of haplotypes, i.e. those with the largest prediction ability, is expected to be beneficial. Conceptually, the marker haplotypes with the highest association with QTL are picked. This strategy should provide a better persistency of predictions in case that close relationships between the reference population and candidates are lacking.

The selection step relies on the following process. For each haplotype k, two likelihoods are computed under model H_0 (δ_k =0, i.e., assuming no haplotypic effect) and H_1 (δ_k =1, i.e., assuming nonzero haplotypic effects with variance σ_h^2).

The log-likelihoods can be written as:

$$-2\ L_i = l_k \log(2\pi) + |\boldsymbol{V_i}| + \boldsymbol{y^{*\prime} Z_k\, V_i^{-1}\, Z_k' y^*}$$

with i=0 or 1, $V_0 = \boldsymbol{Z'_k Z_k}\, \sigma_e^2 + \log(\pi)$

$$V_1 = \boldsymbol{Z'_k Z_k}\, \sigma_e^2 + \boldsymbol{(Z'_k Z_k)'\ (Z'_k Z_k)}\ \sigma_a^2 + \log(1-\pi)$$

The probabilities of H_0 and H_1 are expressed as

$$\text{Prob}(H_0) = \frac{\exp(L_0 - c)}{\left[\exp(L_0 - c) + \exp(L1 - c)\right]}$$

and

$$\text{Prob}(H_1) = \frac{\exp(L_1 - c)}{\left[\exp(L_0 - c) + \exp(L1 - c)\right]}$$

where c is a position parameter (a constant) to avoid computing difficulties.

The selection can then be performed by sampling δ_k from a binomial distribution with probability Prob(H_1).

At each iteration, π, the proportion of haplotypes with no effect, is sampled from a Beta distribution as follow:

$$\pi \sim \text{Beta}(n+\alpha\ ;\ m+\beta)$$

where n and m correspond respectively to the total number of haplotypes retained (i.e., with an effect) or not retained in the model, α and β are parameters representing prior information. According to these prior values, π is either estimated by the iterative process or more or less constraint to a fixed preset value.

Practical implementation in GS3

GS3 is a widely spread software written by Legarra et al. (2013). It is able to handle high throughput SNP information to efficiently compute GBLUP in a direct way or via MCMC, with or without variance estimation, as well as the so-called BayesC and BayesCπ methods. It is a convenient starting point for the implementation of the haplotypic analysis as described above.

Data Format

Haplotypes are assumed to be defined before the analysis. SNP must be phased and without missing values. A first change compared with the SNP approaches involves the coding of haplotypic alleles, which are computed immediately after reading SNP data as a first step of GS3. The haplotype length is defined by a number of SNP provided by the user. Whereas GS3 analyses the complete genome without any reference to marker map or chromosomes, its haplotypic version must use files of phased genotypes per chromosome, with markers ordered according to their physical positions. Haplotypes are defined as sets of contiguous SNP that never overlap. Consequently, the last haplotype on a chromosome can be shorter than the others according to the number of markers on the chromosome. Summary statistics (number of alleles and allelic frequencies) about haplotypes are tabulated.

Haplotype Handling

Whereas SNP-based methods estimate only one effect per SNP, haplotypic methods have to deal with a varying number l_k of alleles per haplotype. This requires indexing the first position and the number of alleles of each haplotype in vector $\boldsymbol{h}$ and the corresponding columns of $\boldsymbol{Z}$.

Computations by Block

SNP-based methods use only scalar algebra whereas the haplotypic approach requires dealing with all alleles of each haplotype at a time. The successive left-hand sides are therefore l_k x l_k matrices which need to be built and inverted. The situation is similar for the log-likelihoods computation. For most methods implemented in GS3, these computations must be done repeatedly, with iterative solving methods if matrices cannot be stored, and especially for MCMC approaches because then these matrices are not constant. The sampling of haplotypic effects was also modified because it involves multivariate normal distributions. All these parts related to matrix algebra are very time consuming computationally.

Substitution Effects

GS3 estimates allelic substitution effects, therefore only one effect per SNP. In order to maintain this interesting feature and to replicate what is done in case of haplotypes of one SNP, it was decided to also estimate the substitution effects in the haplotypic implementation. Therefore, the most frequent allele was chosen as a basis of comparison for the other ones.

Accordingly, the $\boldsymbol{Z}$ matrix was adapted in the following way:

$$\boldsymbol{Z}_{n,k_i} = \frac{\left(l_k m_{n,k_i} - 2\right)}{l_k}$$

with m_{n,k_i} the number (0, 1 or 2) of alleles i of haplotype k carried by individual n.

Computing Time.

As mentioned before, the haplotypic algorithm involves matrix computations instead of scalar operations for SNP. This evolution is quite expensive in terms of computing time. The overall cost is a function of the average allele number per haplotype.

A simulation study was carried out on a Montbéliarde cattle training population of 1701 bulls genotyped with the Illumina Bovine SNP50 Beadchip® (50K). Each genotyped animal was phased and missing genotypes at each particular SNP were imputed using the DAGPHASE software (Druet and Georges (2010).

Table1 shows the total number of haplotypes, the total number of estimated effects, the average number of alleles per haplotype. Haplotype size varied from 2 to 5 SNP. With the 50K chip, the number of observed allelic combination was high due to the modest linkage disequilibrium between successive SNP. With haplotypes of 5 SNP, i.e., with 32 potential alleles, about half of them were actually encountered on average. Consequently, the time to carry out a large number of iterations was highly impacted. Practical strategies must be envisaged to reduce the number of haplotypes and, therefore, limit computing time.

Table 1. Total number of haplotypes, total number of estimated effects, mean number of alleles per haplotype in the Montbéliarde test.

Size of haplotypes	Total number of haplotypes	Total number of alleles	Average number of alleles
2	21,892	82,820	3.78
3	14,599	97,694	6.69
4	10,956	120,222	10.97
5	8,768	147,950	16.87

Haplotype Size

The main benefit of haplotypes is to increase linkage disequilibrium with QTL. Indeed, for QTL with 2 alleles with very unbalanced frequencies, it is quite unlikely to observe strong linkage disequilibrium between one SNP and the QTL. On the other hand, a large increase in number of alleles is likely to lead to an over-parameterization of the model. Therefore, there is an optimum to be determined. From our experience with the French genomic evaluation model (Ducrocq et al., 2014), haplotypes with 8-10 alleles appear to be a good trade-off. This value is obtained with a small number of SNP of the 50K (4 on average). Due to much stronger linkage disequilibrium, more SNP can be included to reach this goal with the high density chip.

ACKNOWLEDGMENT

This work was performed within the framework of the GEMBAL project, funded by the French Agence Nationale de la Recherche (ANR-10-GENM-0014).

LITERATURE CITED

Calus, M.P.L., Meuwissen, T.H.E., de Roos, A.P.W. et al. (2007). Genetics. 178: 553–561.

Croiseau P., Guillaume, F., and Fritz, S. (2012). Interbull Bulletin.

Ducrocq, V., Croiseau, P., Baur, A. et al. (2014). WCGALP, this congress.

Druet, T., and Georges, M. (2009). Genetics. 184 : 189-198

Erbe, M., Hayes, B.J., Matukumalli, L.K. et al. (2012). J. Dairy Sci. 95: 4114-4129.

Gianola, D., de los Campos, G., Hill, W.G. et al. (2009). Genetics 183: 347-363.

Gianola, D., (2013). Genetics. 194: 573-96.

Habier, D., Fernando, R.L., Kizilkaya, K. et al. (2011). BMC Bioinformatics. 12: 186.

Hayes, B., Bowman P., Chamberlain, A. et al. (2009). Genet. Sel. Evol. 41: 51.

Hozé, C., Fritz, S., Phocas, F. et al. (2014), J. Dairy Sci., in press.

Legarra, A., Ricard, A., and Filangi, O. (2013). http://snp.toulouse.inra.fr/~alegarra.

Meuwissen, T., Hayes, B., and Goddard, M. (2001). Genetics 157: 1819-29.

VanRaden, P. (2008). J. Dairy Sci. 91: 4414-23.