Estimation of genome-wide haplotype effects in half-sib designs

D. Kolbehdari, L.R. Schaeffer & J.A.B. Robinson

Center for Genetic Improvement of Livestock, Department of Animal and Poultry Science, University of Guelph, Guelph, Ontario, Canada

Keywords
BLUP; genome-wide selection; haplotypes; quantitative trait loci.

Introduction

Single nucleotide polymorphisms (SNP) are the most abundant form of DNA variation in the genome and are becoming preferred over other genetic markers because of their relatively low mutation rate and ease of genotyping (Romualdi et al. 2002; Youngerman et al. 2004). SNP are commonly used for detection and localization of quantitative trait loci (QTL) for complex traits. Several algorithms for identifying haplotypes using SNP have been developed (Daly et al. 2001; Patil et al. 2001; Zhang et al. 2004). Thousands of SNP can be used to cover the genome of an organism with markers that are no more than 1 cm apart throughout the entire genome. Following Meuwissen et al. (2001), best linear unbiased prediction (BLUP) or Bayesian methods can be used to simultaneously estimate the effects of each contiguous SNP interval (interval effects). The statistical model would allow the estimation of the effects of intervals on the trait of interest. Because SNP are generally biallelic, then for two SNP markers defining an interval, there would be four possible haplotypes an animal could inherit. With enough sons of

Summary

Genome-wide estimated breeding values can be computed from the simultaneous estimates of the effects of small intervals of DNA throughout the genome on a trait or traits of interest. Small intervals or segments of DNA can be created by the use of thousands of single nucleotide polymorphisms (SNP) available in panels of 10, 25 and 50 thousand SNP. A simulation study was conducted to compare factors that could influence the accuracy of genome-wide selection. Factors studied were the heritability of the trait, dispersion of quantitative trait loci (QTL) across the genome and size of the QTL effects. A 100-cM genome was assumed with 100 equally spaced SNP markers and 10 QTL. A granddaughter design was constructed with 20 sires and 100 sons per sire. Population-wide linkage disequilibrium was assumed to be sufficient after 25 generations of random mating starting with 30 sires and 400 dams. Best linear unbiased prediction was used to simultaneously estimate the effects of 99 SNP intervals, based on determining the SNP haplotype of each son inherited from the sire. Indicator variables were used in the model to indicate haplotype transmission. A genome-wide estimated breeding value was calculated as the sum of the appropriate haplotype interval estimates for each son. Correlations between estimated and true breeding values ranged from 0.60 to 0.79. Situations with unequally sized QTL effects and randomly dispersed QTL gave higher correlations. QTL positions could be estimated to within 2 cm or less.
sires, the effect of each of those four possible haplotypes could be estimated. With 10 intervals and 4 haplotypes per interval there would be a total of 40 interval effects to be estimated, for example.

Once interval effects have been estimated, then an estimated breeding value (EBV) would be the sum of the corresponding haplotype interval estimates over all intervals. This EBV will be denoted as GEBV for genome-wide estimated breeding value. Meuwissen et al. (2001) conducted a simulation study that showed the correlation between GEBV and an animal’s true BV to be 0.75 to 0.81 using a genome of 1000 cM, 1001 markers and 1000 QTL. The assumption was made that sufficient generations of random mating had passed such that within small regions of the chromosomes (i.e. 1 cM or less), the marker and QTL alleles tend to segregate together (i.e. sufficient linkage disequilibrium exists).

The sum of the absolute haplotype effects within each interval can be used to locate intervals that might contain QTL. If a QTL does not exist in an interval, then the estimates of haplotype effects should all be small. Intervals with the largest sum of absolute effects would likely contain a QTL or be adjacent to an interval having a QTL.

The purpose of this study was to verify the study of Meuwissen et al. (2001) in terms of accuracy of GEBV, but also in terms of location of QTL. In the current study, there were many sons of sires available such that the effect of each marker haplotype could be estimated accurately. Meuwissen et al. (2001) employed a randomly mated population with markers in linkage disequilibrium with the QTL, while in this study a half-sib population was considered. The current method is based on using extend linkage disequilibrium and the probability of transmission of haplotypes from sire to sons are considered. The effects of size of QTL effects and the dispersion of QTL effects across the genome were to be determined as well as the effect of heritability of the trait.

Materials and methods

Data simulation
A genome of length 100 cM with 100 equally spaced SNP and 10 QTL was generated for each animal. The genome was assumed to consist of only one chromosome. An advantage of a 100-cM genome is the time saved for computations in the simulation study. Both SNP and QTL were assumed to be biallelic with equal allele frequency. Base population animals were generated based on Hardy-Weinberg equilibrium, with 30 sires and 400 dams followed by 25 generations of random mating. The initial alleles for each SNP were generated randomly. The paternal and maternal haplotypes for each animal were generated based on each parent using the Haldane mapping function (no interference) for generation of recombinant haplotypes. Sires and dams in the base generation were assumed to be unrelated. Twenty-five generations of random mating were made to generate linkage disequilibrium.

In the last generation, 20 sires were randomly selected from the previous generation, each with 100 sons. The 20 sire families were randomly split into two groups. In group 1, 10 sire families and their sons were chosen to be in a progeny test scheme each with 100 daughters giving an average daughter yield deviation (DYD). Group 1 animals were genotyped for the SNP and used to estimate the haplotype effects of the genome. In group 2, 10 sire families and their sons were also genotyped but did not have any DYD. Genome-wide estimated breeding values were calculated based on haplotype estimates from group 1 sire families. The DYD for group 1 sires were generated as:

\[ \text{DYD} = 0.5 \text{BV}_{\text{son}} + \delta, \]

where BV_{son} is the breeding value of the son and

\[ \delta = (0.5 \text{BV}_{\text{dams}} + MS + E)/n, \]

where BV_{dams} is the average breeding value of the dams of the progeny, MS is a Mendelian sampling effect, E is a residual effect, n is the number of offsprings per son (100 in this study), \( \delta \) has a normal distribution with mean zero and

\[ V(\delta) = \frac{3}{4h^2\sigma_p^2} + \frac{1 + h^2}{n^2}; \]

\( \sigma_p^2 \) is the total phenotypic variance and \( h^2 \) is the heritability.

Three heritabilities were compared (0.05, 0.30 and 0.50) which determined the average magnitude of the QTL effects. The total phenotypic standard deviation was always 100. The total genetic variance was 3000 in the scenario with heritability equal to 0.3. The 10 QTL covered all the genetic variance. Two types of QTL effects were compared, EQUAL and UNEQUAL. EQUAL QTL effects meant that 10 QTL all had the same size of effects. UNEQUAL QTL effects meant that one QTL had a large effect, two had medium effects and seven had small effects. Another variable was the location of the QTL in the genome, either evenly spaced (EVEN) or randomly spaced (RANDOM) across the genome. Together
these factors gave 12 different alternatives. Fifty replicates were made of each alternative and results were averaged across replicates.

An animal’s true breeding value was the sum of the 10 QTL effects. There were 99 pairs of marker haplotypes for the 100 SNP, giving a total of 396 haplotype interval effects. The haplotype effects were estimated from the 10 sire families in group 1. Only additive genetic effects were considered. The GEBV of sires in group 2 were calculated as the sum of the 99 estimated marker haplotype effects based on the sires’ genotypes.

**Linear haplotype model**

The linear model of analysis to estimate the haplotype effects was:

\[ y = Xb + Zh + e, \]  

(4)

where \( y \) is the vector of sons’ DYD, \( X \) is the incidence matrix for the fixed effects, \( b \) is a vector of fixed effects (in this study only the overall mean was included), \( Z \) is an incidence matrix of indicator variables that relate the inheritance of the son’s haplotype from its sire, \( h \) is the random vector of 396 haplotype interval effects, four haplotypes per interval and \( e \) is a \((n \times 1)\) residual vector.

The key point in this study was the element of the \( Z \) matrix. The marker haplotypes for contiguous pairs of SNP were assumed to be known for sire and son. The indicator variables for all possible combinations of sire and son haplotypes are given in Table 1. The incidence matrix has four columns for every pair of markers. There are four possible haplotypes for markers with two alleles, namely, \( H_1H_1, H_1H_2, H_2H_1 \) and \( H_2H_2 \). In the first row of Table 1, the two sire haplotypes are both \( H_1H_1 \). The son therefore must inherit \( H_1H_1 \). The elements of Table 1 became the elements of \( Z \) depending on the haplotypes of the sire and son.

Also,

\[ E(y) = Xb, \]  

(5)

and

\[ \text{Var}(y) = ZGZ^T + \sigma^2_e, \]  

(6)

where

\[ G = \sum_{i=1}^{99} I_i^2, \]  

(7)

each \( I \) is of order 4, and \( \sigma^2_i \) is the variance of the haplotype effects of the \( i \)th interval. For intervals not containing a QTL, \( \sigma^2_i \) should probably be zero, or at least very small. For intervals with QTL, \( \sigma^2_i \) will be

<table>
<thead>
<tr>
<th>Sire haplotypes</th>
<th>Sons’ haplotypes</th>
<th>Four combination haplotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>One</td>
<td>Two</td>
<td>( H_1H_1 )</td>
</tr>
<tr>
<td>( H_1H_1 )</td>
<td>( H_1H_2 )</td>
<td>1</td>
</tr>
<tr>
<td>( H_1H_1 )</td>
<td>( H_1H_2 )</td>
<td>1</td>
</tr>
<tr>
<td>( H_1H_1 )</td>
<td>( H_2H_1 )</td>
<td>0</td>
</tr>
<tr>
<td>( H_1H_1 )</td>
<td>( H_2H_1 )</td>
<td>0</td>
</tr>
<tr>
<td>( H_1H_1 )</td>
<td>( H_2H_2 )</td>
<td>0</td>
</tr>
<tr>
<td>( H_1H_2 )</td>
<td>( H_1H_2 )</td>
<td>1</td>
</tr>
<tr>
<td>( H_1H_2 )</td>
<td>( H_2H_2 )</td>
<td>0</td>
</tr>
<tr>
<td>( H_2H_1 )</td>
<td>( H_2H_2 )</td>
<td>0</td>
</tr>
<tr>
<td>( H_2H_1 )</td>
<td>( H_2H_2 )</td>
<td>0</td>
</tr>
<tr>
<td>( H_2H_1 )</td>
<td>( H_2H_2 )</td>
<td>0</td>
</tr>
<tr>
<td>( H_2H_2 )</td>
<td>( H_2H_2 )</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1. Scoring values for the incidence matrix to indicate inheritance of haplotype from the sire for pair of single nucleotide polymorphism markers.
larger depending on the magnitude of the QTL effect. In this study a common variance was assumed for all intervals. In addition, the ratio of $\sigma_e^2$ to $\sigma_i^2$ was assumed to be 1 for all analyses regardless of heritability, and no attempts were made to estimate variances towards an optimal analysis.

**QTL position**

To recognize a plausible QTL within each interval, the sum of the absolute values of the haplotype estimated effects was calculated, referred to as $S_i$ for interval $i$. Proceed sequentially through the genome one interval at a time. If $S_i > S_{i-1}$ then go to the next interval. If $S_i < S_{i-1}$, then interval ($i-1$) is a peak and may be the location of a QTL. Proceed until the next peak is found. The majority of $S_i$ should have small values, but there could be a base level of ‘noise’. Only $S_i$ values above the base level need to be considered for QTL. Let $Q$ be the true position of a QTL and $\hat{Q}$ be the estimated position, then bias of estimation will be

$$\text{Bias} = \frac{\sum(Q_i - \hat{Q}_i)}{n},$$

where $n$ is the number of QTL, and the mean squared error (MSE) is

$$\text{MSE} = \frac{\sum(\hat{Q}_i - Q_i)^2}{n}.$$  

**Precision ($P$)** is

$$P = \sqrt{\text{MSE} - \text{Bias}^2}.$$ 

**Genomic estimated breeding values**

The GEBV of animal $k$ was calculated as:

$$\text{GEBV} = z_k' \hat{h},$$

where $z_k'$ is a row of $Z$ for animal $k$. This row indicates the haplotypes of animal $k$ inherited from its sire. Meuwissen et al. (2001) found that using only the intervals that contain possible QTL gave much lower correlations of GEBV with true BV. Hence all 99 intervals were used in this study. The correlation between GEBV and true BV was calculated for an estimate of accuracy. Statistics was averaged over the 50 replicates for each alternative.

**Results and discussion**

**QTL positions**

Figures 1 and 2 present the $S_i$ values for each interval for the alternatives with EQUAL and UNEQUAL QTL size, respectively, and EVEN distribution of QTL effects. Figure 3 shows the $S_i$ values for EQUAL QTL size and RANDOM distribution of QTL over the genome. The base noise level seems to be about 1.75. Values of $S_i$ above 1.75 would be candidates for possible QTL locations. Peaks were quite easy to pick out in Figures 1 and 2, but not as easy in Figure 3.

The statistics for precision of estimating QTL position are given in Table 2. The best situation was when $h^2 = 0.5$, UNEQUAL QTL effects and RANDOM QTL distribution along the genome. Precision in this case was 1.30 cM. UNEQUAL QTL effects and RANDOM QTL distribution was the best alternative within each heritability level. Precision degraded as heritability decreased.
The worst alternative was EQUAL QTL effects and RANDOM QTL distribution within each heritability level. The other two alternatives were intermediate in precision. EVEN QTL distribution was not preferred although this alternative was easier to simulate, and has been used in many simulation studies.

In all alternatives the QTL position could be located within a 2-cM region. This method was useful for detecting QTL in that all QTL (of any size for any trait) were located simultaneously with one statistical analysis. Having more SNP marker within 1-cM region would give more intervals of smaller size and might increase the precision of locating QTL. Given the rate of discovery of new SNP, a very dense marker map will be possible in the near future. A simulation of denser marker maps is needed. In the very least, these analyses should help to locate major QTL more quickly.

Besides locating QTL of interest, the analyses could provide an estimate of the number of QTL affecting different traits. The estimated number, of course, could not be greater than the number of intervals produced by the SNP markers. Two or more QTL within 1-cM region would appear as one QTL. Using the $S_i$ values, the distribution of QTL effects could be plotted. Some traits may follow the assumptions of the infinitesimal model while others may have fewer QTL as well as some with large effects. No epistatic effects were simulated between QTL loci, but this should probably be considered in future studies (Hayes & Goddard 2001).

**Table 2** Comparison of 12 alternative scenarios using linear mixed haplotype model for estimating quantitative trait loci (QTL) position and genomic-estimated breeding values representing three heritability levels, size of QTL effects and dispersion of QTL in the genome

<table>
<thead>
<tr>
<th>Scenarios $H^2$</th>
<th>QTL size</th>
<th>QTL dispersion (cM)</th>
<th>Bias MSE (cM$^2$)</th>
<th>Precision</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 0.5 EQUAL EVEN</td>
<td>1.86 5.63</td>
<td>1.37</td>
<td>$0.63 \pm 0.08$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 0.5 EQUAL RANDOM</td>
<td>2.34 7.36</td>
<td>1.32</td>
<td>$0.77 \pm 0.06$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 0.5 UNEQUAL EVEN</td>
<td>1.28 3.21</td>
<td>1.24</td>
<td>$0.62 \pm 0.08$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 0.5 UNEQUAL RANDOM</td>
<td>1.56 4.43</td>
<td>1.30</td>
<td>$0.79 \pm 0.05$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 0.3 EQUAL EVEN</td>
<td>1.61 4.44</td>
<td>1.43</td>
<td>$0.60 \pm 0.07$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 0.3 EQUAL RANDOM</td>
<td>2.66 9.38</td>
<td>1.40</td>
<td>$0.76 \pm 0.06$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 0.3 UNEQUAL EVEN</td>
<td>1.62 4.90</td>
<td>1.51</td>
<td>$0.61 \pm 0.09$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 0.3 UNEQUAL RANDOM</td>
<td>1.52 4.24</td>
<td>1.38</td>
<td>$0.77 \pm 0.08$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 0.05 EQUAL EVEN</td>
<td>2.04 6.88</td>
<td>1.64</td>
<td>$0.60 \pm 0.08$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 0.05 EQUAL RANDOM</td>
<td>2.72 9.26</td>
<td>1.51</td>
<td>$0.74 \pm 0.07$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 0.05 UNEQUAL EVEN</td>
<td>1.88 6.11</td>
<td>1.57</td>
<td>$0.60 \pm 0.09$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 0.05 UNEQUAL RANDOM</td>
<td>1.16 3.04</td>
<td>1.41</td>
<td>$0.75 \pm 0.08$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The worst alternative was EQUAL QTL effects and RANDOM QTL distribution within each heritability level. The other two alternatives were intermediate in precision. EVEN QTL distribution was not preferred although this alternative was easier to simulate, and has been used in many simulation studies.

**Figure 3** Sum of absolute haplotype effects per interval for heritability of 0.5, EQUAL quantitative trait loci (QTL) size of effects and RANDOM QTL distribution (one replicate). (Actual QTL positions were: 8, 11, 25, 28, 41, 45, 60, 67, 82 and 95).

**GEBV**

Correlations (i.e. accuracy) of GEBV with true BV for group 2 sire families are given in Table 2. The range was from 0.60 to 0.79. Correlations above 0.70 were obtained for all situations with RANDOM QTL distribution. In this situation the number of markers between two QTL locations is different. Segments with more markers between QTL have more information to estimate haplotype effects. Correlations increased with an increase in heritability, but the range was only from 0.75 to 0.79. EQUAL QTL effects gave slightly higher (0.01) correlations than UNEQUAL QTL effects, but the difference was not significant.

EVEN QTL distribution generally gave correlations below 0.65. In real life, QTL are most likely to be randomly dispersed through the genome, and therefore, accuracy of GEBV should be better than if they were evenly distributed. Also, the majority of QTL will have small and equal effects with only a few QTL having very large effects.

Meuwissen et al. (2001) presented the difficulties of estimating a different variance for each interval. In their study assuming a common variance, $\sigma^2_e$ for all intervals did not seriously affect the accuracy of GEBV, and was much easier to implement. Hence for this study, a common variance was assumed for all intervals. In addition, the ratio of $\sigma^2_e$ to $\sigma^2$ was assumed to be 1 for all analyses, and no attempts were made to estimate variances. Thus, the results of this study could possibly be made better (more optimistic) if different variances were estimated for each interval, and if different variance ratios were used in the mixed model equations for each interval.
The accuracy of GEBV was reasonably high. Most dairy cows, e.g. never achieve accuracy greater than 0.50 in their lifetime (after having one or more progeny and several lactation records). A GEBV can be obtained when the heifer is born and would have an accuracy of at least 0.70. Female selection could become much more important in future breeding programmes because accuracy of selection would be increased and generation interval would be decreased. Genetic change can possibly be doubled over current progeny testing strategies (Schaeffer 2006).

Conclusions

Dense marker maps in the future will allow animals to be evaluated very accurately at birth through simple genotyping using a SNP array with thousands of markers. The accuracy of genomic estimated breeding values will cause breeding strategies in dairy cattle and most species to be radically altered. Finding QTL with major effects will become easier. Marker-assisted selection will not really be necessary because all QTL can be selected simultaneously using GEBV.

Acknowledgements

The authors acknowledge the financial contributions of Semex Alliance, the Ontario Research and Development Challenge Fund and the Ontario Centre for Agricultural Genomics.

References


