Precision of estimated QTL positions in granddaughter designs using combined haplotype sharing TDT and linkage analysis

Davood Kolbehdari a,b,c,*, Gerald B. Jansen d, Ian McMillan a,1, L.R. Schaeffer a,2

a Center for Genetic Improvement of Livestock, Department of Animal and Poultry Science, University of Guelph, Guelph, Ontario, Canada N1G 2W1
b Department of Animal Science, Abureihan Higher Education Complex, University of Tehran, Iran
c 87 Terraview CR., Guelph, Ontario, Canada N1G 5E9
d Dekoppel Consulting, Guelph, Ontario, Canada N1G 2Y8

Received 4 April 2006; accepted 3 May 2006

Abstract

The aim of this study was to develop the linear haplotype sharing transmission disequilibrium test (LHS-TDT) method and combine this method with the simple regression method to estimate the precision of QTL positions in granddaughter designs. This precision was determined by Monte Carlo simulation in granddaughter designs. A single bi-allelic QTL at the midpoint of a linkage group and 26 markers with 1 cM intervals and with two alleles each were simulated. Three linear models, (i.e. the simple regression model, the linear haplotype sharing TDT method and the combination of these two models) were compared. The mean of absolute differences (A) between the estimated and true QTL position of each method was considered for six different scenarios consisting of combinations of a number of markers and the most frequent haplotypes. The mean of A, using the simple regression method, was 4.38 centimorgan (cM). The means of A using the LHS-TDT method were less than the simple regression method in all scenarios and ranged from 1.86 to 3.82 cM depending on the scenario. The mean of A using the combined method was more than the LHS-TDT method and less than the simple regression method. The means of A using the combined method ranged from 2.32 to 4.36 cM. Therefore, for populations similar to those population simulated in this study, the LHS-TDT was better than the simple regression method and the combined method for precision of estimated QTL position in granddaughter designs.

© 2006 Published by Elsevier B.V.

Keywords: Quantitative trait loci (QTL); Haplotype sharing transmission disequilibrium test (HS-TDT); Granddaughter designs

* Corresponding author. Center for Genetic Improvement of Livestock, Department of Animal and Poultry Science, University of Guelph, Guelph, Ontario, Canada N1G 2W1. Tel.: +1 519 824 9140 (Home), +1 519 824 4120x58353 (Office); fax: +1 519 837 8867.
E-mail addresses: dkolbeh@uoguelph.ca, dkolbeh@yahoo.ca (D. Kolbehdari), janroz@teksavvy.com (G.B. Jansen), imcmilla@uoguelph.ca (I. McMillan), lrs@uoguelph.ca (L.R. Schaeffer).
1 Tel.: +1 519 824 4120x53560 (Office); fax: +1 519 837 8867.
2 Tel.: +1 519 824 4120x53645 (Office); fax: +1 519 837 8867.
1871-1413/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.livsci.2006.05.003
1. Introduction

Mapping quantitative trait loci (QTL) in a very small region of a chromosome is very beneficial for efficient marker assisted selection (MAS) in livestock populations (Dekkers and Hospital, 2002). Linkage analysis of fine mapping of target genes in outbred populations is based on recombination fractions that can be observed in the data set. Linkage analysis using marker information in current generations and in outbred populations has low resolution and precision for estimating the position of QTL. In human genetics, linkage disequilibrium has been widely used for fine mapping of complex diseases (Hastbacka et al., 1992; Jorde, 1995; Terwilliger, 1995; Kaplan et al., 1995; McPeek and Strahs, 1999; Service et al., 1999). There have been several studies of livestock populations involving linkage disequilibrium analyses and combined linkage and linkage disequilibrium analyses for fine mapping of quantitative trait loci in a small region of the genome (Riquet et al., 1999; Farnir et al., 2000; Meuwissen and Goddard, 2000, 2001, 2004; Meuwissen et al., 2002; Du et al., 2002; Lund et al., 2003; Pérez-Enciso et al., 2003; Abdallah et al., 2003; Grapes et al., 2004; Lee and van der Werf, 2004).

Meuwissen and Goddard (2000, 2001) developed a method to estimate the positions of QTL using linkage disequilibrium between several closely linked markers and QTL. They used recombination information and pedigrees in the data set from current generations, base generations and historical recombinations. This method was implemented for detecting twinning rates in dairy cattle data (Meuwissen et al., 2002). The results of this study showed that mapping of QTL with combined linkage and linkage disequilibrium analyses were more accurate than by only using either linkage analysis or linkage disequilibrium analysis. Combined linkage and linkage disequilibrium method were developed for multiple QTL with posterior probability (Meuwissen and Goddard, 2004). The multiple QTL method gave a much sharper QTL peak than with single QTL mapping. Du et al. (2002) developed the Meuwissen and Goddard (2001) approach and included the Markov Chain Monte Carlo method to estimate the IBD probability when the linkage phase of animals is unknown.

The results of several studies show that the transmission disequilibrium test (TDT) approach and the regression method are appropriate for fine mapping QTL in outbred livestock populations (Knott et al., 1996; Hernandez-Sanchez et al., 2002). The simple regression method is based on linkage analysis and using multiple linear regression models with multiple linked markers to detect QTL. The transmission disequilibrium test is based on linkage disequilibrium analysis. The TDT approaches using haplotypes of multiple tightly linked markers are more informative than those using single markers. Several studies in human genetics showed the haplotype transmission disequilibrium test is a precise method for estimating the position of the target gene (Clayton, 1999; Zhao et al., 2000; Zhang et al., 2003). Therefore, the objective of this study was to combine linkage analysis using the simple regression method with a haplotype sharing transmission disequilibrium test to estimate the precision of estimated QTL positions in granddaughter designs.

2. Methods

Three linear models based on the objective of this study were compared. The simple regression method was used to capture the linkage information. The linear haplotype sharing TDT method was used to capture the linkage disequilibrium information and a combination of the simple regression method with LHS-TDT was used to capture the combination linkage and linkage disequilibrium information.

2.1. The Simple Regression Model (MLR)

The simple regression method used in this study is based on the multiple linked marker approach described by Knott et al. (1996). This is the analysis of individual linkage groups across several half-sib families. The probability of inheritance of the first haplotype of the parent of a linkage group at every 1 cM interval conditional to its marker genotype is computed for each offspring. The quantitative trait loci are fitted at each fixed interval along the linkage group by regression of a phenotypic trait on the conditional probability. The analysis is based on the nested designs within families and the residuals are pooled across families to calculate a test statistic. The multiple linear regression model...
(MLR) with multiple linked markers for every linkage group is as follows:

\[ y_{ij} = x_{ij} \beta_i + s_i + e_{ij} \]  

(1)

where \( y_{ij} \) is the daughter yield deviations (DYD) of son \( j \) from sire \( i \), \( x_{ij} \) is the conditional probability of individual \( j \) inheriting the first haplotype of sire \( i \), \( \beta_i \) is the regression coefficient within family \( i \), \( s_i \) is the polygenic effect of sire \( i \) and \( e_{ij} \) is the residual effect. Sires could be heterozygous for at least one marker locus. If all markers in a linkage group are uninformative for an individual, the conditional probability would be equal to 0.5. If the paternal allele of the sire is transmitted to his son, the conditional probability would be equal to 1 and otherwise is equal to zero. The null hypothesis in this model was no QTL segregation in this chromosome region. The test statistic was an \( F \)-test and \( F \)-ratios were computed at every map position (Knott et al., 1996). The position with the largest \( F \)-ratio was used to estimate the position of the QTL.

2.2. Linear haplotype sharing TDT (LHS-TDT)

The mixed linear haplotype sharing TDT (LHS-TDT) model for analysing the quantitative trait data is:

\[ y_{ij} = h_{ij} d_k + s_i + e_{ij}. \]  

(2)

where \( y_{ij} \) is the daughter yield deviations (DYD) of son \( j \) from sire \( i \), \( h_{ij} \) is the haplotype sharing coefficient of haplotype \( k \) from sire \( i \) transmission to son \( j \), \( d_k \) is the regression coefficient of haplotype \( k \), \( s_i \) is the polygenic effect of sire \( i \), \( e_{ij} \) is the residual effect. The next section describes how to compute the haplotype sharing coefficient TDT.

The key technique in this study was to calculate a haplotype sharing transmission disequilibrium test based on linkage disequilibrium analyses. The number of haplotypes increases with increasing number of markers, \( H_i = \{M_{1i}, M_{2i}, \ldots, M_{ki}, \ldots, M_{mi}\} \), where \( H_i \) is the specific haplotype with marker alleles \( M_{1i} \) to \( M_{mi} \). For example, for 6 bi-allelic markers there are 64 different haplotypes. A large number of haplotypes would require additional degrees of freedom for the test and thus reduce the power. To reduce the number of degrees of freedom the most frequent haplotypes in the data set were identified. In this study 2, 4 or 6 most frequent haplotypes were identified. The literature and simulated data showed a few haplotypes that are the most frequent haplotypes in the populations.

The individuals with the less frequent haplotypes were included to a new group in the analysis. Table 1 shows the structure of haplotype sharing TDT indicator values in this study. The haplotype sharing TDT indicator values were calculated based on transmission or non-transmission of specific marker haplotypes from sire to progeny. These indicator values were introduced based on Sun et al. (2000) when a parent genotype is available. In this study it was assumed that the marker haplotypes of sires and progeny were correctly known.

The null hypothesis tested by haplotype sharing TDT is \( H_0 \): no linkage and no association (\( \theta = 0.5, \delta = 0 \)), and the alternative hypothesis is \( H_1 \): linkage and association (\( \theta < 0.5, \delta \neq 0 \)), where \( \theta \) is the recombination fraction between the marker and the hypothesized QTL and \( \delta \) is the disequilibrium coefficient between them. When the recombination fraction (\( \theta \)) is equal to 0.5, the marker and QTL alleles are segregating independently. Recombination fractions less than 0.5 mean the marker and QTL alleles are linked and the frequencies of gametes depend on the value of \( \theta \). The disequilibrium coefficient (\( \delta \)) shows the association between markers and QTL. With \( \delta = 0 \), the markers and QTL are not associated. The test statistic was an \( F \)-test and \( F \)-ratios were computed at every map position.

2.3. Combined model (ComB)

The combined model was based on a combination of two regression coefficients of the simple regression

<table>
<thead>
<tr>
<th>Table 1</th>
<th>LHS-TDT indicator values for transmission ( H_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Progeny genotypes</td>
<td>Sire genotypes</td>
</tr>
<tr>
<td>( H_{ij} )</td>
<td>( H_{ij} )</td>
</tr>
<tr>
<td>( H_{ij} )</td>
<td>( H_{ij} )</td>
</tr>
<tr>
<td>( H_{ij} )</td>
<td>( H_{ij} )</td>
</tr>
<tr>
<td>( H_{ij} )</td>
<td>( H_{ij} )</td>
</tr>
</tbody>
</table>

* Uninformative haplotype.
* Informative haplotype and \( H_{ij} \) transmission from sire to progeny.
* Informative haplotype and \( H_{ij} \) non-transmission from sire to progeny.

D. Kolbehdari et al. / Livestock Science 105 (2006) 137–143
method (MLR) and linear haplotype sharing TDT (LHS-TDT) in one mixed linear model as follows:

\[ y_{ij} = x_{ij}\beta_i + h_{ij}\delta_k + s_i + e_{ij}, \]  

where \( y_{ij} \) is the daughter yield deviations (DYD) of son \( j \) from sire \( i \), \( x_{ij} \) is the conditional probability of individual \( j \) inheriting the first haplotype of family \( i \), \( \beta_i \) is the regression coefficient within family \( i \), \( h_{ij} \) is the haplotype sharing coefficient of haplotype \( k \) from sire \( i \) transmission to son \( j \), \( \delta_k \) is the regression coefficient of haplotype \( k \), \( s_i \) is the polygenic effect of sire \( i \), \( e_{ij} \) is the residual effect. This model was compared to MLR and the LHSTDT. The hypothesis test was the same of the previous model and the test statistic was an \( F \)-test and \( F \)-ratios were computed at every map position.

2.4. Parameter estimation and test statistic

The test statistic was a standard \( F \)-test. The positions of the marker locus with the largest \( F \)-statistic was used to estimate the position of the QTL. The ability of the three methods to estimate the QTL position was evaluated with simulated data. The absolute differences between the estimated QTL position using these three methods and the true QTL position obtained for each replicate of a simulation is as follows:

\[ A_i = |P_i - P_o| \]  

where \( A_i \) is the absolute difference between the estimated QTL position, \( P_i \), and the true position, \( P_o \). The mean of \( A_i \) was used to measure the ability of each method to estimate the position of the QTL. The method with a smaller mean of \( A \) would be preferable.

The bias of each method was estimated by

\[ \text{Bias} = \Sigma(P_i - P_o)/n \]  

where \( n \) is the number of replicates of simulations (Grapes et al., 2004).

3. Data simulations

Half-sib data, which is commonly found in dairy cattle populations, were simulated. The single bi-allelic QTL at the midpoints of the marker intervals was simulated. Twenty six marker loci each with two alleles and equal intervals were generated. The initial allele frequency in the QTL and markers was equal to 0.5. The intervals between the markers were the same and equal at 1 cM. The Haldane function (no interference) was used to generate these markers and QTL.

The half-sib family generated was based on granddaughter designs in outbred livestock populations. The simulation in this study was divided into three steps. First, the base population was simulated based on linkage equilibrium. The markers and QTL genotypes were generated randomly for 30 sires and 400 dams in the base population and for the next 100 generations. The effective population size was 111 and fixed for 100 generations. Second, after 100 generations of random mating 10 grandsires were selected based on daughter yield deviations (DYD). Third, the 10 grandsires were randomly mated to 1000 dams from the previous generation to generate 100 sons per grandsire. The DYD were estimated for all sons, assuming 100 progeny per sons. All scenarios were simulated using 500 replicates.

The polygenic heritability of the quantitative trait was 0.3 and the total phenotypic standard deviation was assumed to be 100. The total genetic variance was 3000 and the QTL variance relative to total genetic variance was 0.1. The daughter yield deviations (DYD) created for each son were based on 100 daughter’s progeny tests. The same DYD were used in three models analysis. DYD is an average of the phenotypes of the daughters adjusted for systematic environmental effects and genetic values of the dams.

The combination of the number of markers in haplotypes and the number of most frequent haplotypes were considered in six different scenarios. The number of markers used in the haplotypes were two, four and six. The linkage group of markers with equal intervals and midpoint QTL were separated into five
segments. The QTL in all scenarios was simulated at the third segment. The number of most frequent haplotypes used were two, four and six. Table 2 shows the different combinations of these scenarios.

### 4. Results

The means of absolute differences (A) for each method for different scenarios were estimated. Table 3 shows the means of A overall for the three linear models. The average mean of A across the scenarios using LHS-TDT method was 2.96 cM. The mapping resolution using LHS-TDT was more accurate than MLR. The means of A using LHS-TDT were 2.48, 1.86, 3.05, 3.02, 3.81 and 3.56 cM for scenarios one to six, respectively. In scenario one using two markers and two of the most frequent haplotypes, using paired t-test, the mean of A using LHS-TDT was significantly different from MLR ($P < 0.05$). When the number of most frequent haplotypes was increased to 4, in scenario two, the mean of A using LHS-TDT significantly decreased to 1.86 cM. For haplotypes with 4 markers the mean of A increased significantly in scenarios three and four using LHS-TDT. When the number of most frequent haplotypes increased from 4 to 6, the mean of A using LHS-TDT of two scenarios were very close. For haplotypes with 6 markers the means of A increased in scenarios five and six using LHS-TDT.

The mean of A using ComB was larger than the mean of A using LHS-TDT but smaller than the mean of A using MLR ($P < 0.05$). The average of A using ComB across the scenarios was 3.53 cM, which differs significantly from the average of A using LHS-TDT. The means of A using ComB were 2.91, 2.32, 3.71, 3.75, 4.36 and 4.14 cM for scenarios one to six, respectively. ComB was not a successful approach for precise estimation of QTL positions compared with LHS-TDT. The mean of A in all scenarios was larger than for LHS-TDT. Scenario two produced the smallest mean of A using ComB, which is the same result as using LHS-TDT.

The bias of the estimated QTL positions was expected to be zero, because the QTL was simulated at the midpoint of the linkage group. Table 4 shows the bias for the three linear models. The bias was not exactly zero, but they were very small. The differences between bias values and zero in different scenarios and models were tested. The statistical tests across the scenarios and models showed the differences were not significant. Thus the bias values were produced by chance across the replicates.

### 5. Discussion

TDT is widely used to test the linkage and associations between marker and target gene, which is based on linkage disequilibrium, and has been shown to be a powerful approach. When TDT uses multiple tightly linked markers the statistical power may be increased to compare with the single marker

---

**Table 3**
The means of absolute difference (centimorgans) of estimated QTL positions for three mapping methods under the six scenarios (regression method=4.38 (4.15a) cM)

<table>
<thead>
<tr>
<th># Scenarios</th>
<th># Markersb</th>
<th># Haplotypesc</th>
<th>LHS-TDT</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2.48 (3.738)*</td>
<td>2.91 (3.875)*</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
<td>1.86 (3.124)**</td>
<td>2.32 (3.498)*</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3.05 (3.934)*</td>
<td>3.71 (4.125)</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>6</td>
<td>3.02 (3.874)*</td>
<td>3.75 (3.952)</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>4</td>
<td>3.81 (4.107)</td>
<td>4.36 (4.077)</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>6</td>
<td>3.56 (4.007)</td>
<td>4.14 (3.935)</td>
</tr>
</tbody>
</table>

* The standard error of estimation.

** Table 4**
The bias of QTL position estimates for three mapping methods under the six scenarios (bias of regression method was 0.28)

<table>
<thead>
<tr>
<th># Scenarios</th>
<th># Markersb</th>
<th># Haplotypesc</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>LHS-TDT</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0.02</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>4</td>
<td>0.4</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>6</td>
<td>−0.13</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>4</td>
<td>0.11</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>6</td>
<td>0.06</td>
</tr>
</tbody>
</table>

* The number of markers in haplotype.

** The number of most frequent haplotypes.

---

*a The number of markers in haplotype.

*b The number of most frequent haplotypes.

** The differences between the simple regression method and this method are significant at level $P < 0.05$.

*** The differences between the simple regression method and this method are significant at level $P < 0.001$. 

---

D. Kolbehdari et al. / Livestock Science 105 (2006) 137–143
TDT method. The single marker TDT is usually used to detect disease loci and QTL. Multiple tightly linked markers are used as haplotypes in TDT to fine map the disease loci and QTL. The results of Zhang et al. (2003) and the current study show that the haplotype sharing TDT is a powerful approach for fine mapping genes.

TDT using multiple tightly linked markers as haplotypes produces some difficulties in the statistical approach. First, increasing the number of markers increases the number of haplotypes and the additional degrees of freedom required for the test. Therefore the power of the test decreases. Second, the haplotypes must be reconstructed from multiple tightly linked marker genotypes. Several methods have been used to reduce the number of degrees of freedom. The first method groups the haplotypes: Seltman et al. (2001) used evolutionary relationships; Li et al. (2001) used the clustering method. The second approach uses maximum identity length contrast. Bourgain et al. (2000) used this approach and the simulation results showed that this method may be more powerful than single marker TDT. Finally, Zhang et al. (2003) have used haplotype sharing to analyze multiple tightly linked markers. The method used in this study was based on the most frequent haplotypes in the data set. The literature and simulated data in the current study showed that after 100 generations of random mating, the most frequent haplotypes in the populations are few. The most frequent haplotype approach proved to be a powerful approach in LHS-TDT method. In scenarios with large numbers of the most frequent haplotypes, some replicates in the simulation did not produce enough haplotypes and the replicate was removed from the analysis. For example in scenario six with six most frequent haplotypes 1650 replicates were filtered for 500 replicate observations. Several algorithms have been introduced for haplotype reconstruction from population data (Stephens et al., 2001). The error rates of haplotype reconstruction are small, especially for the case of a restricted number of haplotypes in the sample. The current study used simulated data and the error rates of haplotype reconstruction were zero, because the haplotype structure was known and the probability of correct phase construction with 100 sons was very high, even with very simple methods.

The three linear models were considered for precision of estimated QTL positions in granddaughter designs. The simple regression method based on linkage analysis was not a successful method. The linkage analysis method was an appropriate approach for genome scans and to identify the region of the genome that has an association with QTL. LHS-TDT provided better results in the scenarios which were simulated for fine mapping. This method gave small error rates when estimating the QTL positions. The combined model has both the regression coefficients of the simple regression method and the LHS-TDT method, but was not a suitable approach. In some situations these two regression coefficients were equal and created a linear dependence between the columns of the $X$ matrix. Further studies are needed to extend this method and investigate the properties of this approach for fine mapping QTL in livestock populations.

Acknowledgements

The authors are very grateful to the Dairy Cattle Genetics Research Council (DairyGen) and the Matching Investment Initiative of Agriculture and Agri-food Canada and the University of Tehran, Islamic Republic of Iran.

References


