LINKAGE DISEQUILIBRIUM INTERVAL MAPPING OF QUANTITATIVE TRAIT LOCI

S. Boitard[1,2], J. Abdallah[3], H. de Rochambeau[4], C. Cierco-Ayrolles[1,2] and B. Mangin[1]

[1] Unité de Biométrie et Intelligence Artificielle, INRA, BP 52627, 31326 Castanet-Tolosan Cedex, France

[2] Laboratoire de Statistique et Probabilités, Université Paul Sabatier Toulouse III, 31400 Toulouse, France

[3] Laboratoire de Génétique Cellulaire, INRA, BP 52627, 31326 Castanet-Tolosan Cedex, France

[4] Station d'Amélioration Génétique des Animaux, INRA, BP 52627, 31326 Castanet-Tolosan Cedex, France

INTRODUCTION

The detection and mapping of loci affecting quantitative traits (QTLs) of interest in human, animal, and plant populations have attracted considerable research interest for several decades. This work has mainly concentrated on the use of pedigree or family data, especially in animal and plant populations where the structure of such experimental pedigrees can easily be planned and controlled. More recently, linkage disequilibrium (LD) methods based on the study of unrelated individuals from a given population have emerged as a promising tool for refining gene location estimates. In this paper we present a new interval mapping method using the LD information, denoted as HAPim, and provide a comparison with the identity by descent (IBD) method of Meuwissen and Goddard (2000).

THE HAPim METHOD

Mixture model. Let us assume that a quantitative trait locus (QTL) is segregating in a population with two alleles, Q and q. Consider that phenotypic measurements of the trait are available for N_s individuals from the population. Consider also that these individuals have been typed for M polymorphic markers in the region surrounding the QTL, and that for any individual the two haplotypes h_1 and h_2 composing its genotype have been identified. We assume that each phenotype measurement is a mixture of three independent normal distributions, corresponding respectively to the three genotypes at the QTL: QQ, Qq, and qq. Following Falconer and Mackay (1996), we assume that these distributions have the same variance σ^2 and expected values $\mu+a$, $\mu+d$, and $\mu-a$ respectively, with a and d defined as the additive and dominance deviations from the mean μ of the homozygote QTL genotypes. Assuming Hardy-Weinberg equilibrium, the probabilities $P(QQ|h_1,h_2)$, $P(Qq|h_1,h_2)$ and $P(qq|h_1,h_2)$ of being drawn from each of these distributions are directly deduced from the expected haplotype frequencies in the population under study. They can be derived under a few assumptions on the population model, as illustrated in the following sections.

Derivation of the expected haplotype frequencies. The derivation of the expected haplotype frequencies in the population models the decay in LD from an initial mutation to the present. The simultaneous use of more markers should increase the accuracy of the QTL location because the past recombination events can be identified more precisely. However, increasing the number of markers makes the computation of expected haplotype frequencies more complex. We previously (Abdallah *et al.* 2004) provided two mapping methods that required the computation of these expected frequencies for haplotypes comprising just one marker in addition to the trait locus. The first one was a single-marker method: for each postulated QTL position x on the map, only the closest marker was considered. The second method was a

composite likelihood method that used the set of L closest markers at each postulated position whilst assuming that these markers were associated with the QTL independently of each other. The derivation of the expected haplotype frequencies were thus the same as in the first method, but the results obtained with each of the L markers were then pooled together to compute the probabilities of bearing allele Q.

The above assumption of independence is clearly violated when markers are linked. To account for a correlation between close loci, we determined an expression for the expected frequency of haplotypes with one trait locus and two markers. This is an extension of the derivations in Xiong and Guo (1997), which were based on the hypothesis that the mutant allele is very rare. At each postulated position of the QTL, we can apply this formula to the haplotype comprising this locus and its two flanking markers. Let c_1 and c_2 be the recombination rates with the left- and right-side markers. Let π_{i1}, π_{i2} and $\pi_Q(0)$ be the frequencies of alleles *i1* of the left-side marker, *i2* of the right-side marker, and of allele Q at time *0* respectively. Let finally $\pi_{i1,Q,i2}(0)$, $\pi_{i1,Q}(0)$ and $\pi_{Q,i2}(0)$ denote the frequencies of haplotypes *(i1,Q,i2)*, *(i1,Q),* and *(Q,i2)* respectively. Assuming that allele frequencies at markers are deterministic, time invariant, and in equilibrium we proved that, under the three-locus Wright-Fisher model, the expected frequency of haplotype *(i1,Q,i2)* after t generations is given by

$$\begin{aligned}E[\pi_{i1,Q,i2}(t)] &= \pi_Q(0)\pi_{i1}\pi_{i2} + (1-c_1)^t(\pi_{i1,Q}(0) - \pi_Q(0)\pi_{i1})\pi_{i2} \\ &+ (1-c_2)^t(\pi_{Q,i2}(0) - \pi_Q(0)\pi_{i2})\pi_{i1} \\ &+ (1-c_1)^t(1-c_2)^t(\pi_{i1,Q,i2}(0) - \pi_{i1,Q}(0)\pi_{i2} - \pi_{Q,i2}(0)\pi_{i1} + \pi_Q(0)\pi_{i1}\pi_{i2})\end{aligned}$$

We can thus derive $P(Q|i1,i2)=E[\pi_{i1,Q,i2}(t)]/\pi_{i1,i2}$.

Initial creation of LD. Our method relies on the assumption that the haplotype frequencies in the population were in equilibrium until a genetic or demographic event suddenly created LD between the QTL and a unique marker haplotype at time *0*. Classical examples of such events are the introduction of a favorable allele Q into an isolated population, by mutation or migration. After this event, haplotype frequencies evolve along generations as described earlier until the present generation denoted as *t.* This model allows us to reduce the number of parameters used to describe haplotype frequencies at time *0*. Indeed, following Terwilliger (1995) and Xiong and Guo (1997), we introduce a heterogeneity parameter α in addition to allele frequencies π_{i1}, π_{i2} and $\pi_Q(0)$. This parameter represents the proportion of new copies of allele Q introduced at time *0* into the population. Note that $\alpha=1$ if Q did not exist previously in the population. Assuming that new alleles Q are associated with allele *1* of both markers, the initial haplotype frequencies can be expressed as

$$\begin{aligned}\pi_{i1,Q}(0) &= (1-\alpha)\pi_{i1}\pi_Q(0) + \alpha\pi_Q(0)\delta_{i1=1} \\ \pi_{Q,i2}(0) &= (1-\alpha)\pi_{i2}\pi_Q(0) + \alpha\pi_Q(0)\delta_{i2=1} \\ \pi_{i1,Q,i2}(0) &= (1-\alpha)\pi_{i1}\pi_{i2}\pi_Q(0) + \alpha\pi_Q(0)\delta_{i1=1}\delta_{i2=1}\end{aligned}$$

where $\delta_{x=y}$ is the Kronecker delta operator (equal to 1 if $x=y$ and 0 otherwise).

Linear model. At the first order, our method is finally equivalent to fitting a linear model $Y=X\theta+\varepsilon$, where Y is the vector of phenotype records, θ is the vector of diplotype effects, ε is a vector of independent random noises with variance σ^2 and X is a design matrix of size N_s x D,

D being the number of possible diplotypes. Each component of θ is a known function of the QTL location x and of the other following parameters: μ, a, d, α, $\pi_Q(0)$, t and the marker haplotype initially associated with Q. Each component of θ is also assumed to fit the phenotype mean observed for one particular diplotype, so that each diplotype provides one equation. Our aim is to identify the parameter values that are optimal with respect to the whole set of equations. For that purpose we use a likelihood maximization procedure.

COMPARISON WITH IBD METHODS

Simulation results. Using forward simulations as described in Abdallah et al (2004), we duplicated the simulation scenarios described in Table 2 in Meuwissen and Goddard (2000): 50 population replicates with biallelic markers initially at equal frequencies with spacings of 0.25, 0.5, and 1.0 centi-Morgan (cM), an effective population size and a sample size of $N=N_s=100$, and a time $t=100$ since the initial mutation. We applied HAPim on this simulated data. The distribution of the deviations (in marker intervals) in the QTL location estimates from the correct bracket are presented in Table 1. They can be directly compared with those of Table 3 in Meuwissen and Goddard (2000), which were obtained with their IBD method. A chi-square test of equality between the deviation distributions of HAPim and the IBD method revealed no significant difference (the smallest p value was 0.08). We also tested our method under the simulation scenarios used by Grapes *et al.* (2004, 2006), who compared single- and two-marker regression analysis with an IBD method very similar to that in Meuwissen and Goddard (2000). The results we obtained with HAPim were similar to the ones given by their IBD method using two-marker haplotypes: least-square mean absolute differences (LSMDs) of 1.36, 0.71, and 0.39 for marker spacings of 1.0, 0.5, and 0.25cM, respectively.

Discussion. There are fundamental differences between HAPim and the IBD method. First, haplotype effects are modeled as fixed effects in the former and as random effects in the latter. While it is well-known that location parameters are easier to estimate than dispersion parameters, it is not clear whether this has a significant effect on the estimation of the QTL position. Second, the IBD method does not include dominance effects, while HAPim handles that very efficiently. Third, the time t since the initial creation of LD and the effective population size N have to be known before using the IBD method. Some simulation results in Meuwissen and Goddard (2000) suggested that the default choice of $t=100$ and $N=100$ was almost optimal, whatever the true value of these parameters. However the comparison of tables V and VII in Meuwissen and Goddard (2001) indicates that the IBD matrix with $N=1000$ is really different from the one with $N=100$. Thus it is not obvious why the IBD method assuming $N_s=100$ should be accurate for a population of actual effective size $N=1000$. On the other hand, neither t nor N are required for the use of HAPim. Consequently this method can be used in a wider range of populations.

A nice advantage of the IBD method is its ability to deal with haplotypes composed of more than two markers. If used with caution, this can provide more accuracy in location estimates (Grapes *et al.* 2006). HAPim could also offer this possibility in the future, as flanking markers can be replaced by flanking haplotypes. At present, the several differences highlighted in the previous paragraph already justify the interest of this method.

Table 1. Distribution of the deviations (in marker brackets) of the estimated QTL position from the correct bracket obtained with HAPim under the default simulation scenarios described in Meuwissen and Goddard (2000).

Marker interval (cM)	Deviation of estimated position from the correct interval [a]				
	0	1	2	3	4
	Replicates with frequency of the *Q* allele > 0.1				
1.0	16	17	9	5	3
0.5	12	20	10	2	6
0.25	12	18	8	6	6
	Replicates with frequency of the *Q* allele > 0				
1.0	15	14	6	8	7
0.5	10	17	12	7	4
0.25	11	14	11	5	9

[a] Deviation of 0 means the estimated position was in the correct marker bracket, 1 means the estimated position was one bracket away from the correct position, etc.

REFERENCES

Abdallah, J.M., Mangin, B., Goffinet, B., Cierco-Ayrolles, C. and Pérez-Enciso, M. (2004) *Gen. Res*. **83**: 41–47.

Grapes, L., Dekkers, J.C.M., Rothschild, M.F. and Fernando, R.L. (2004) *Genetic*s **166**: 1561-1570.

Grapes, L., Firat, M., Dekkers, J.C.M., Rothschild, M.F. and Fernando, R.L. (2006) *Genetics* in press

Falconer, D.S. and Mackay, T.F.C. (1996) "Introduction to quantitative genetics" 4th edn., Longman, Essex.

Meuwissen, T.H.E. and Goddard, M.E. (2000) *Genetics* **155**: 421–430.

Meuwissen, T.H.E. and Goddard, M.E. (2001) *Genet. Sel. Evol.* **33**: 605-634.

Terwilliger, J.D. (1995) *Am. J. Hum. Genet*. **56**: 777–787.

Xiong, M. and Guo, S-W. (1997) *Am. J. Hum. Genet.* **60**: 1513–1531.