Structure of linkage disequilibrium and length of IBD segments in simulated populations.

Florence F. Ytournel, Hélène Gilbert, Tom T. Druet, Didier Boichard

STRUCTURE OF LINKAGE DISEQUILIBRIUM AND LENGTH OF IBD SEGMENTS IN SIMULATED POPULATIONS

F. Ytournel, H. Gilbert, T. Druet, D. Boichard

INRA Station de Génétique Quantitative et Appliquée 78352 Jouy en Josas Cedex, France

INTRODUCTION

Methods for Quantitative Trait Loci (QTL) fine mapping using Linkage Disequilibrium (LD) have been developed (e.g. Meuwissen and Goddard 2000, 2001; Abdallah *et al* 2004), generally assuming that LD originates from random drift in populations in Hardy-Weinberg equilibrium. But LD, which quantifies the non-random association between alleles at different loci, can also be generated by selection, migration or mutation (Lynch and Walsh 1998), especially in active livestock population. A simulation program was developed in order to test the robustness of fine mapping methods to those different sources of LD. In this paper, various scenarios were simulated, and resulting LD was described by considering localisation of the maximum D' value (Hedrick 1987) and length of segments derived from founders.

METHODS

Simulations. Haplotypes were composed of 20 evenly spaced markers and one QTL located in the middle of the haplotype (figure 1). Distance between adjacent markers was 1 cM or 0.25 cM, for a total haplotype length of 19 cM and 4.75 cM, respectively. Loci were numbered according to their rank in the haplotype, from -10 to -1 for the left markers, from 1 to 10 for the right markers, and 0 for the QTL. Each locus had 5 alleles in the founder population. It was also given a copy number corresponding to the number of the founder haplotype it belonged to, making it possible to follow true IBD segments. The individual phenotypes were simulated as the sum of a polygenic effect, a residual effect (both of them being normally distributed with mean 0) and the QTL effect. The QTL explained 20% of the genetic variance and the heritability of the trait was 0.17.

The simulated populations were composed of 100, 150 or 200 individuals per generation. They were simulated under random drift (q=1) or they were submitted to selection by truncation: the best 80% (q=0.8) or 90% (q=0.9) of the individuals were selected on phenotype as potential parents for the next generation. Matings were random. Generations were separate.

Different criteria were computed after 10, 25, 50, 75, or 100 generations of evolution of these populations. For each situation, 2,000 replicates were performed and those where the QTL was fixed were discarded.

Location of the maximum D'. The *D'* (Hedrick 1987) is used to measure LD between two loci. It is defined as $D' = \sum_{i=1}^{I}\sum_{j=1}^{J} p_i q_j \left|D'_{ij}\right|$ where *I* and *J* are the number of alleles at loci A and B, respectively, p_i (q_j) is the frequency of the allele *i* at locus A (B) and $|D'_{ij}|$ is the

absolute value of Lewontin's (1964) normalized LD measure. D'_{ij} is computed as $D'_{ij} = \frac{x_{ij} - p_i q_j}{D_{max}}$ with x_{ij} the observed frequency of the haplotypes A_iB_j and

$$D_{max} = \begin{cases} \min(p_i q_j, (1-p_i)*(1-q_j)) \text{ if } (x_{ij} - p_i q_j) < 0 \\ \min(p_i*(1-q_j), (1-p_i)*q_j) \text{ if } (x_{ij} - p_i q_j) > 0 \end{cases}.$$

For each simulation, the marker with highest D' with the QTL was identified (mD'). Similar analyses were performed with haplotypes of 2 (H2) or 4 (H4) marker haplotypes, applying the former formula where each haplotype was assimilated to a different allele. The haplotype with the highest LD with the QTL was identified and called mhD'. Haplotypes were numbered according to figure 1, from -9 to 9 for 2-loci haplotypes, and from -8 to 8 for 4-loci haplotypes.

Two statistics were derived:
- A 95% confidence interval of the length of the segment (in number of markers) containing both the QTL and the marker (or haplotype) with the largest D'-value, and the generation when this number was minimum. This confidence interval was obtained by discarding the 5% largest values over the simulations.
- The proportion of maximum LD observed in the neighbourhood of the QTL, and the generation when it was reached. For individual markers, the neighbourhood of the QTL was defined by markers -2 to 2 whereas neighbour haplotypes were required to contain at least marker 1 or marker -1.

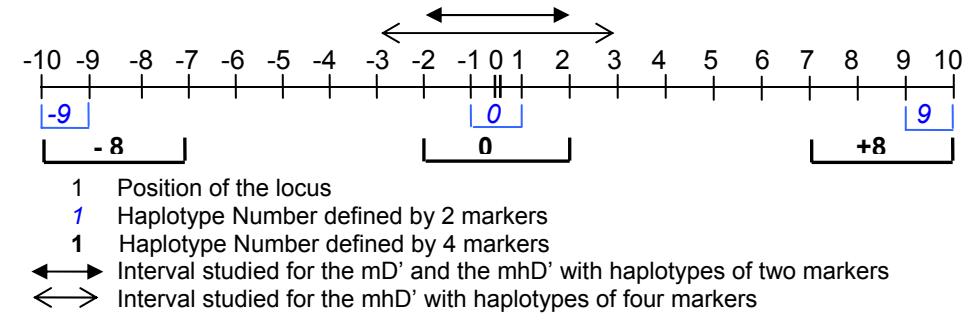

Figure 1. Definition of haplotypes and intervals studied.

Length of the segments derived from the founders. A recombination event was assumed to occur in the middle of an interval between two adjacent loci when the copy numbers differed from each other. The length of the segments derived from a unique ancestor with no apparent recombination was calculated for any segment or only for the segment including the QTL.

RESULTS AND DISCUSSION

Location of the maximum D'. The results are summarized in Table 1. After discarding the simulations where the QTL was fixed (D'=0 for all loci), the distributions were obtained with 522 (100 individuals, q=0.9 and 100 generations) to 2000 replicates.

Table 1. Maximal proportion of m(h)D' in the QTL interval and 95% confidence interval of the length (cM) of the interval containing m(h)D' and the QTL, according to population size and map density .

Map	Population size	Criterion	Maximal LD observed in the QTL neighbourhood		Segment containing both the QTL and the marker or haplotype with highest D'	
			% [(a)]	Gene-ration[(a)]	95% CI of length[(a)]	Generation[(a)]
Sparse	100	mD'	91/*83*/73	50/*25*/25	7/*8*/12	25&50/*25*/10
		mhD'–H2	83/*62*/61	50/*50*/25	5/*8*/12	50/*25*/10&25
		mhD'–H4	49/*49*/45	75/*25*/25	6/*8*/10	50&75/*25*/25
	150	mD'	95/*85*/73	50/*25*/25	5/*10*/12	50/*25*/25
		MhD'–H2	89/*73*/62	75/*25*/25	4/*8*/12	50/*25*/25
		MhD'–H4	52/*50*/46	50/*50*/25	5/*8*/9	50-100/*25*/25
	200	mD'	96/*85*/72	50/*25*/25	4/*7*/12	75&100/*25*/25
		MhD'–H2	92/*73*/64	75/*25*/25	4/*7*/11	50&100/*25*/25
		MhD'–H4	53/*48*/46	75/*25*/25	4/*7*/9	50-100/*25*/25
Dense	100	mD'	67/*51*/43	50/*50*/50	3.5/*4*/4.5	50/*50*/10&25
		MhD'–H2	63/*43*/35	75/*50*/25	3.25/*4*/4	50/*10&25*/10
		MhD'–H4	45/*38*/33	75/*50*/25	2.5/*3.5*/3.5	50&75/*25*/25
	150	mD'	78/*53*/41	75/*50*/50	2/*3.75*/4.5	75/*50*/25
		MhD'–H2	70/*46*/36	100/*50*/25	2.5/*3.5*/4	50&75/*25*/25
		MhD'–H4	46/*39*/33	100/*50*/25	2/*3.25*/3.5	75&100/*25*/25
	200	mD'	84/*54*/44	100/*50*/50	2.25/*4*/4.25	75/*50*/25
		MhD'–H2	73/*45*/35	100/*50*/25	2.25/*3.75*/3.75	75&100/*25*/25
		MhD'–H4	47/*38*/32	100/*50*/25	2/*3.25*/3.5	75&100/*25&50*/25

(a) q=1 / *q=0.9* / q=0.8

The generation number before reaching the maximal concentration increased with population size (25 for a size of 100 *vs* 100 for a size of 200), while it reduced with selection (100 where q=1 *vs* 25 where q=0.8 for a population size of 200). Similar observations stood for the maximum LD location around the QTL and the length of the segments containing 95% of the m(h)D'. The maximum concentration seemed lower when the population was submitted to selection, but higher concentrations might be reached for steps not studied here (15 or 20 generations, for example). This concentration was probably due to the reduction of the number of haplotypes in the population, but also to optimum frequency of the favourable QTL allele: the maximal concentration was reached for QTL frequencies around 0.5 in selected populations (data not shown). After the maximum was reached, the m(h)D' was re-distributed over the whole chromosomal region while LD still increased at the population level, probably due to a reduction of haplotype variability. This could have led to localizations of m(h)D' at positions away from the QTL, giving spurious estimated locations of the QTL based only on LD.

The proportion of mhD' obtained with haplotypes of four markers in the studied interval was generally lower than that obtained with haplotypes of two markers or with a single marker. This was in agreement with the results of Abdallah *et al.* (2004) who found that the multipoint composite likelihood of Terwilliger (1995) performed better with haplotypes of two markers.

Length of the segments derived from the founders.

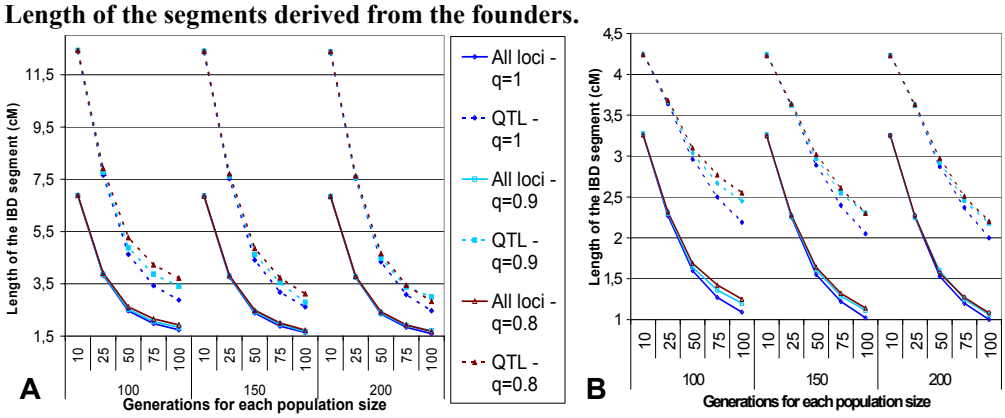

Figure 2. Evolution of the length of the segments derived from the founders according to the generation, the population size and the map density (A: sparse map, B: dense map).

The average length of IBD segments slightly increased with selection. The difference observed on the overall length was due to the difference on the segment including the QTL. The mean length of the segment was greater for small population sizes.
With a four-fold map density, the mean length was only divided by 1.2 to 1.4 after 100 generations. Resolution of QTL fine-mapping methods is limited by the length of IBD segments and, therefore, does not proportionally increase simply with map density.

CONCLUSION

These results show that selection influences the structure of the LD as well as the length of the IBD segments inherited from a founder. One can expect that the observed differences would affect fine-mapping results on selected populations. The robustness of fine mapping methods to those different structures will be further investigated.

REFERENCES

Abdallah J. *et al.* (2004) *Genet. Res.* **83** : 41-47.
Hedrick P.W. (1987) *Genetics* **117** : 331-341.
Lewontin R.C. (1964) *Genetics* **49** : 49-67.
Lynch M. and Walsh B. (1998) « Genetics and Analysis of Quantitative Traits ». Sinauer Associates, Inc. Publishers, Sunderland, Massachusetts, USA.
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** : 421-430.