Concordance between IBD probabilities and linkage disequilibrium.

Florence Ytournel, Hélène Gilbert, Didier Boichard

UR337, INRA, Station de Génétique Quantitative et Appliquée, F-78352 Jouy en Josas cedex

Corresponding author. Station de Génétique Quantitative et Appliquée, INRA,
78352 Jouy en Josas cedex, France. Tel: 33 1 34 65 28 19. Fax: 33 1 34 65 22 10.
e-mail: florence.ytournel@jouy.inra.fr.

Keywords: Identity By Descent/ Linkage Disequilibrium/ Selection

Summary

Studies (Grapes *et al.,* 2006; Zhao *et al*, 2007) have been conducted to estimate the optimal haplotype length to compute IBD probabilities as defined by Meuwissen and Goddard (2001). We simulated different genetic maps with various lengths, marker densities and markers composing them (SNP, microsatellites or a mixture of both). The simulated populations included 100 individuals. IBD probabilities were computed either at the QTL position or at the middle point of the marker bracket preceding the QTL.

We evaluated:

- the distribution of the IBD probabilities for haplotypes of 4, 6 or 10 markers depending on the real IBD status at the QTL, to evaluate the ability of the probabilities to discriminate IBD from non-IBD QTL,
- the evolution of the correlations between the real QTL IBD status and IBD probabilities depending on a) the location of marker in highest linkage disequilibrium with the QTL (LD, evaluated with χ^{2}' (Yamazaki, 1977)) and b) the values of LD between the QTL and its closest marker.

It appeared that non-IBD QTL were better identified than IBD ones with all designs. The discrimination ability improved with the presence of microsatellite markers in the haplotype, with the increase of the haplotype length or the increase of the number of markers defining the haplotype, and when IBD probabilities were computed at the QTL position. The correlation between IBD probabilities and real IBD grew with the intensity of LD and with shortest distances between the QTL and the marker in maximum LD with it.

I. INTRODUCTION

Linkage Disequilibrium (LD) has become of common use for fine-mapping purposes. Methods have been developed to take LD into account, for example through the use of the excess of association of a marker allele with a gene allele (Terwilliger, 1995; Abdallah *et al*., 2004; Farnir, 2002) or probabilities to be Identical By Descent (IBD) (Meuwissen and Goddard, 2000, 2001). Meuwissen and Goddard (2001) proposed to compute these IBD probabilities analytically according to the Identity By State (IBS) of markers for a given haplotype centred on the supposed QTL location.

The pertinence of these probabilities for fine-mapping purposes depends on their capacity of discriminating between IBD and non-IBD QTL. Previous studies (Grapes *et al*., 2006; Zhao *et al*., 2007) focused on the resolution or power of fine-mapping achieved when using IBD probabilities. Up to our knowledge, no further look was given at their ability to actually discriminate between loci IBD or not. On one hand the discrimination ability may be influenced by the molecular information (number of

markers used to calculate the IBD probabilities, length of the haplotype, di- or multi-allelic markers). On the other hand, it is also well-known that selection influences LD (Cannon, 1963); one can thus expect the distribution of the IBD probabilities to be also affected. We evaluated the impact of these factors on the distribution of the IBD probabilities according to the real IBD status of the QTL. We also looked at the correlation between the IBD probabilities and IBD statuses depending on the LD intensity between the QTL and its closest marker.

II. METHODS

1. Simulated designs

The populations were composed of 100 individuals and resulted from 100 separated generations. Matings were at random, possibly with selfing.
Phenotypes were simulated as the sum of a polygenic effect, an additive QTL effect and a residual effect, where the polygenic and residual effects were normally distributed with mean 0 and constant variances over time. The QTL effect was fixed: in the founder population, the QTL explained 20% of the genetic variance and the heritability of the trait was 5%. Truncation selection could be applied over the last 20 generations: animals with highest phenotypes in proportion q=0.8 (80% selected) were retained as potential parents for the next generation, else q=1, which means no selection. Designs were selection was applied are thereafter noted s.

2. Genetic maps

Three different genetic maps were considered, whose general features are presented on Fig. 1. The "sparse" and "centered" maps were composed of 8 markers and one QTL. They differed for the distance between the QTL and its two closest markers: this distance was 0.1 cM on the "sparse" map *vs.* 0.05 cM on the "centered" map, therefore implying a regular marker spacing in the 0.3 cM surrounding the QTL on the centered map . The dense map has the same features as the centered map except that it has been densified in the 0.9 cM central region with a regular marker spacing of 0.1 cM (it was then composed of 12 markers instead of 8).

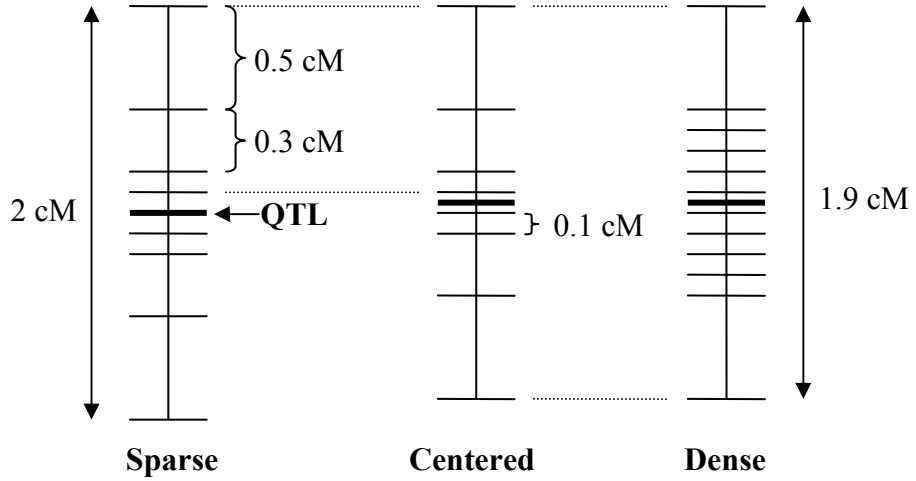

Figure 1: Different maps considered in the study. Thick lines: QTL locus, thin lines: marker loci

We considered different sets of markers:

- Uniform maps: a unique marker type, either with 5-alleles markers (microsatellite–like, noted thereafter "-M") or biallelic markers (SNP-like), that were tested on the three genetic maps,
- Composite maps (noted "–V"): mixture of two marker types, with SNP-like loci at all locations but two positions with microsatellite-like loci, the third and the sixth on the centered map and the fifth and the eighth on the dense map. This was tested only these two maps.

3. Computation of IBD probabilities

The IBD probabilities were computed according to the analytic formula of Meuwissen and Goddard (2001). For each pair of haplotypes, the IBD probabilities were computed at the QTL position.

The IBD probabilities were computed for haplotypes composed of 4 or 6 markers on the two sparser maps (A and B on Fig. 1) and 6 or 10 markers on the denser map (C), the QTL being located in the middle of the haplotype. The corresponding haplotype lengths and the abbreviations used in the following figures are reported in Table 1.

Table 1: Length of the haplotypes used for IBD computations depending on the genetic map and corresponding notations.

Map Haplotypes		Sparse	Centered	Dense
4 markers	Length (cM)	0.4	0.3	-
	Notation	*4*	*4c*	
6 markers	Length (cM)	1.0	0.9	0.5
	Notation	*6*	*6c*	*6d*
10 markers	Length (cM)	-	-	0.9
	Notation			10d

4. Computation of linkage disequilibrium

Among the different measures available to quantify LD, Zhao *et al.* [26] showed that χ^2' (Yamazaki, 1977) better handled the LD decline with respect to the distance between the markers. We used it to evaluate the LD between the QTL Q and its previous marker M, chosen as the best indicator of QTL allele segregation in the population.

If x_{ij} are the observed frequencies of the haplotypes M_iQ_j and p_i (q_j) the frequency of the allele i (j) at marker locus M (QTL), the association disequilibrium between the alleles is given by $D_{ij} = x_{ij} - p_i q_j$.

Given all pairs of alleles (i=1,I ; j=1,J) between M and Q, $\chi^2 = 2N\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{D_{ij}^2}{p_i q_j}$ is computed, and finally χ^2' is obtained as $\chi^2' = \frac{\chi^2}{2N(l-1)}$, where l =min(I,J).

5. Description of IBD probability accuracy

A thousand replicates was obtained for each simulated design. To evaluate the IBD probability relationships with LD, we focused on two criteria:

- the distribution of the IBD probabilities with respect to the real IBD status at the QTL. We first dispatched all IBD probabilities into ten classes, each class covering a range of 0.1: class 1, corresponding to IBD probabilities between 0 and 0.1 up to class 10 containing all concurrencies showing an IBD probability exceeding 0.9. Then, among each class, we distinguished the IBD and non IBD occurrences.;
- the correlations between the real IBD status at the QTL (0 or 1) and the estimated IBD probabilities, depending on the linkage disequilibrium level between the QTL and its closest marker. The correlations were first computed for each simulated population and then averaged over the 1,000 replicates.

Considering each IBD class as a threshold to infer an IBD status from IBD probabilities, we derived from the IBD distributions the type I and type II error rates corresponding to each class of probability. The type I error was defined as the probability for a haplotype pair of being considered IBD while it is not IBD. It was thus computed, for each IBD class, as the ratio of the sum of the non-IBD occurrences with IBD probability equal or higher than the current IBD class and the total number of occurrences in these IBD classes. On the other hand, a type II error was defined as the probability for a haplotype pair of being considered non-IBD while it is. It was computed for each IBD class as ratio of the sum of the IBD occurrences with IBD probability equal or lower than the current IBD class by the total number of IBD occurrences. Power was calculated as one minus the type II error rate. Thresholds for type I (resp. type II) error rates were defined as the lowest (resp. highest) value of IBD probability from the IBD class being the one having a type I (resp. type II) error rate closest to 0.05 but lower than that value.

III. RESULTS

1. Distribution of the IBD probabilities

i. On "uniform" maps in unselected populations

The repartition of the IBD and non-IBD QTL in the classes of IBD probabilities (Fig. 2) showed that the great majority of the highest probabilities corresponded to IBD QTL whereas the lowest probabilities were mostly attributed to non-IBD QTL: for the series 4 and 6d-M, respectively 97.5% and 98.3% of the QTL pairs with probabilities under 0.1 (*i.e.* class 1) were non-IBD while respectively 92.4% and 99.8% of the pairs with probabilities belonging to class 10 were IBD for the QTL.

Non-IBD QTL were much better discriminated than IBD QTL (Fig 3): in more than 75% of the situations where the QTL were not IBD, the IBD probability was lower or equal to 0.1. When the two QTL were actually IBD, the proportion of IBD probabilities over 0.9 was comprised between 27% (with a haplotype composed of 4 SNP on map A) and 60% (haplotypes composed of 10 microsatellite markers on map C).

The segregation ability was improved with a denser map in the QTL neighbourhood, with haplotypes containing microsatellites and with a longer physical length covered by the haplotype for a given number of markers. It seemed that the map density around the QTL had more influence on the IBD probabilities than the haplotype length (comparing for instance the curves 4c and 4).

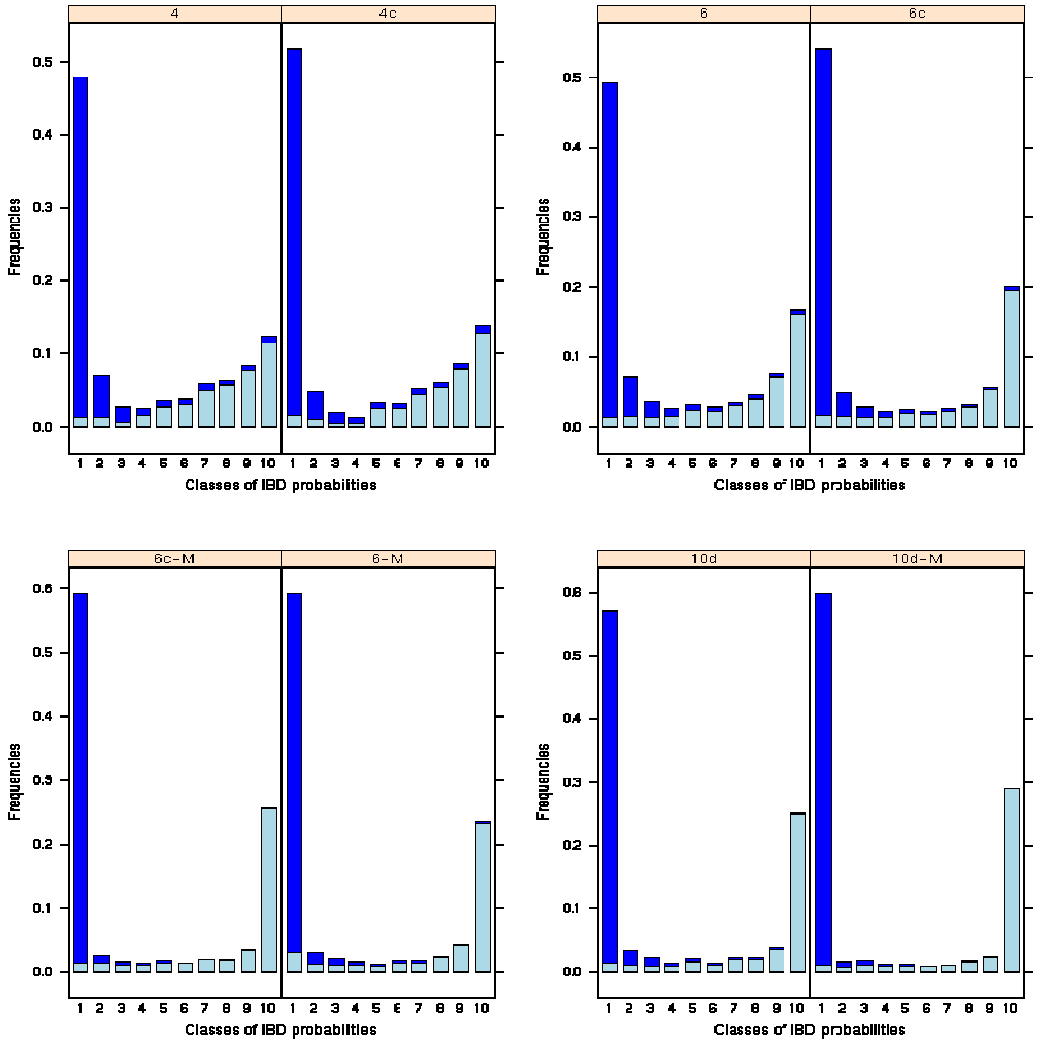

Figure 2: Repartition of the pairs of haplotypes in the classes of IBD probabilities for IBD QTL (light blue) or non-IBD QTL (blue), depending on the haplotype length, the molecular information and the density around the QTL.

ii. On “composite” maps in unselected populations

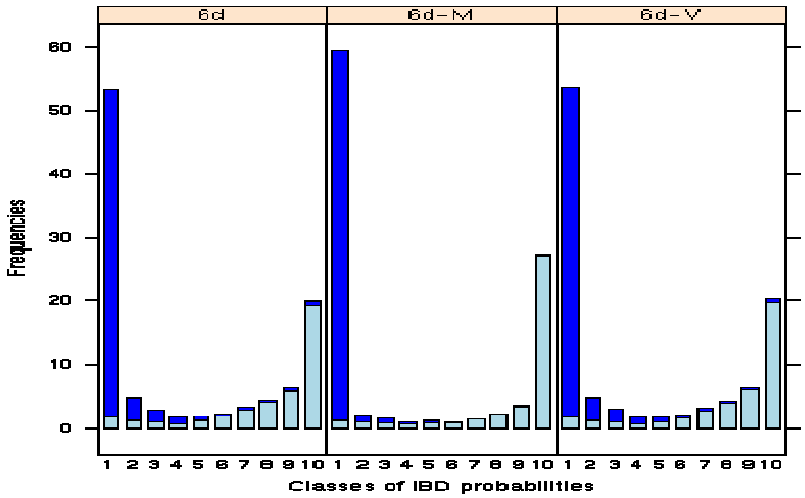

Figure 3: Frequencies of the haplotype classes and repartition of IBD QTL (blue) or non-IBD QTL (red) among them when using only microsatellites (6d-M), only SNP (6d) or two microsatellites and four SNP (6Cd) on map C

Frequencies of the extreme classes of IBD probabilities increased when the whole haplotype was composed of microsatellites instead of SNP (1.3% more non-IBD QTL in class 1 and 2.0% more IBD QTL in class 10 when comparing the designs 6d and 6D-M). However, replacing two SNP by microsatellites did not really improve those frequencies (designs 6Cd and 6d in Fig.3).

iii. Uniform and composite maps in selected populations

Frequencies of the highest class of IBD probabilities were strongly reduced when selection was applied on the populations (Fig. 4): for the series 6s *vs.* 6 and 6ds *vs.* 6d, there were respectively 23 and 28% less IBD QTL pairs in class 10 for the populations submitted to selection. This coincided with an increase of the intermediate frequencies (IBD probabilities between 0.5 and 0.8) when the QTL were truly IBD, corresponding to an increase of the overall frequency of these classes of IBD probabilities. However, the distribution of the IBD probabilities of the non-IBD QTL was not affected by the selection, and the frequencies in the lowest classes of IBD probabilities changed little, independently from the IBD status of the loci.

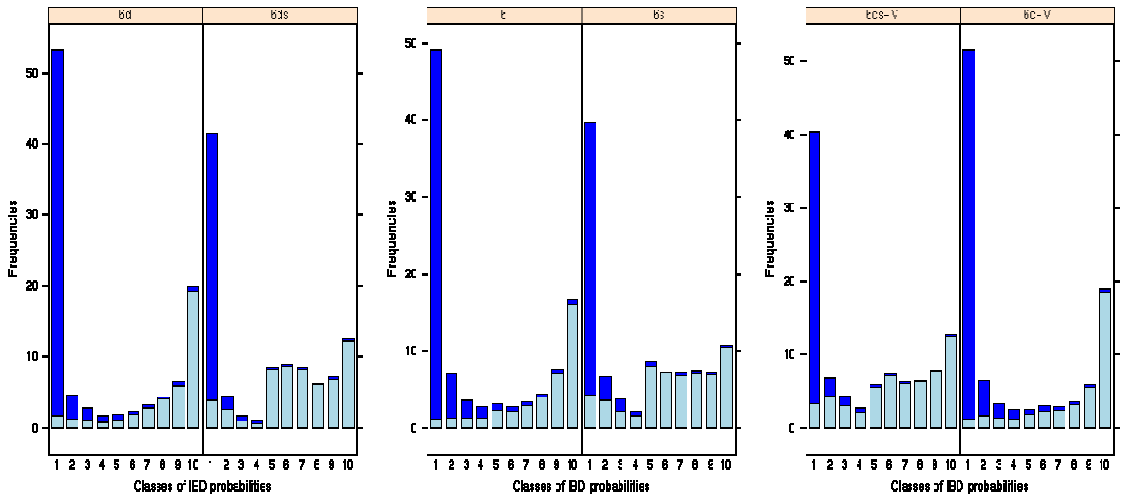

Figure 4: Distribution over 1,000 replicates of the IBD probabilities (class 1 corresponding to IBD probabilities under 0.1 and class 10 to probabilities between 0.9 and 1.0) according to the real IBD status at the QTL location depending on the haplotype length, the molecular information and the density around the QTL and selection (solid lines). C: "composite" haplotypes comprising two microsatellite markers, else haplotypes composed of SNP.

iv. Thresholds for type I and type II error rates

Table 2: IBD probability thresholds to infer IBD status with type I or type II error rates of 5%.

Name of the series	Threshold for a type I error of 5% (*respective power %)*	Threshold for a type II error of 5% (*respective type I error %*)
6	0.9 *(41)*	0.1 *(21)*
6-M	0.5 *(83)*	< 0.1 *(61)*
6s	0.4 *(66)*	< 0.1 *(43)*
6d	0.8 *(64)*	0.1 *(20)*
6d-M	0.2 *(92)*	0.1 *(6)*
6ds	0.3 *(86)*	< 0.1 *(42)*
6d-V	0.7 *(76)*	0.1 *(20)*
6c-V	0.7 *(70)*	0.1 *(23)*
6cs-V	0.3 *(78)*	< 0.1 *(43)*
10d	0.5 *(86)*	0.1 *(14)*
10d-M	0.2 *(93)*	0.2 *(5)*
10ds	0.2 *(82)*	< 0.1 *(42)*

From Table 2, it appeared clearly that, as expected, 5% type I errors were achieved with low thresholds for the IBD probabilities, particularly when the map was composed only of microsatellites and/or if the population was submitted to selection. Using denser maps helped in lowering this threshold (6 vs. 6d) while the use of two microsatellites had little influence (6d vs. 6d-V). No general threshold for the probability distributions could be obtained from these results. The associated powers were frequently over 50%, and even larger then 70% when selection was applied.

Concerning the threshold probability for a power of 95%, the maximum probabilities were rather similar across all designs. They were all lower or equal to 0.2 and were even under 0.1 when the population was under selection pressure. However, they were achieved at the expense of high type I errors, except on a dense map with microsatellite markers without any selection (6d-M and 10 d-M).

2. Relationships between Linkage Disequilibrium and IBD probabilities

The average correlations between the real IBD status between the QTL and the IBD probabilities between the haplotypes were over 0.70 for all designs and for all intensities of LD between the QTL and its closest marker (Table 3, only LD<0.1 or LD>0.9 are presented). Clear trends were difficult to point out, and the number of occurrences in the class of LD exceeding 0.9 was rather low (commonly around 50 to 70 replicates among the 1,000), so these results must be taken with caution. The correlations slightly augmented with an increased LD (from + 0.03 to + 0.11). They were higher when the haplotype was totally composed of microsatellites, when it was composed of more markers (comparing 4c and 6c for example) or physically longer (comparing 6c-M and 10d-M for instance). However, it seemed that using 10 microsatellite markers instead of 6 did not increase the concordance between IBD probabilities and the real IBD statuses; it even tended to be deleterious when the population was submitted to selection (-0.05 between 10ds-M and 6ds-M). This effect of selection when increasing the number of markers defining a haplotype (and at the same time, the physical length covered by it) held also when the haplotypes were defined with SNP only. More generally, selection led to reduced correlations, independently from the LD level. Finally, replacing two SNP with microsatellites did not help in increasing the correlation.

Table 3: Average correlations between IBD probabilities and true IBD status over 1,000 replicates, given two classes of LD between the QTL and its previous marker, comprised either between 0.0 and 0.1 or between 0.9 and 1.0.

Series name	Average correlation between IBD probabilities and IBD status	
	$0.0 < LD < 0.1$	$0.9 < LD \leq 1.0$
4c	0.78	0.86
6	0.80	0.87
6s	0.75	0.86
6c	0.82	0.87
6c-V	0.82	0.87
6cs-V	0.78	0.85
6c-M	0.86	0.93
6d	0.85	0.93
6ds	0.80	0.86
6d-V	0.86	0.92
6d-M	0.91	0.94
6ds-M	0.83	0.87
10ds	0.77	0.86
10d-M	0.90	0.94
10ds-M	0.78	0.87

IV. DISCUSSION AND CONCLUSION

This study aimed at establishing the discrimination ability of IBD probabilities, their adequacy with the real IBD status of the QTL locus and their evolution with an increasing LD. A practical implication concerns better definitions of clustering thresholds to reduce the IBD matrix size, make it positive definite and reduce computation time.

In our study, the distributions of IBD probabilities are influenced by molecular factors such as the genetic marker informativity in the haplotype, its length and the map density around the tested position. For all studied designs, the distributions of the IBD probabilities provided a good discrimination ability, particularly for the non-IBD QTL, leading to very low chances of observing false-positives. This is partly due to the U-shaped distribution of the probabilities with very few occurrences in the middle-range IBD probabilities. Maps composed only of microsatellites provided the most discriminant distributions of IBD probabilities, *i.e.* those with the greatest proportions of IBD QTL with IBD probabilities superior to 0.9 and of non-IBD QTL with IBD probabilities lower than 0.1. As it is unrealistic to imagine haplotypes composed of microsatellites as dense as those simulated in this study, we checked the properties of haplotypes composed of both microsatellite and SNP markers, to provide additional haplotype variability. However, may be due to the intern disequilibrium of information it created within the haplotype, this could not help much the discrimination of IBD loci using IBD probabilities. It suggested that best results may be obtained with dense, balanced and variable haplotypes, which may only be obtained with high densification of SNP.

The correlations between IBD probabilities and real QTL IBD statuses increased slightly with the LD level when computed between the QTL and its closest marker locus. However, even with low LD, the correlations were high (over 0.7), which indicates that these probabilities do not only take LD into account. They rely by construction on IBS and that is finally what will be caught, wherever it comes from. This implies that high number of IBS loci, due to recent population mixture for example, may lead to spurious associations using IBD probabilities. There seemed to be an optimal number of markers to be used on the densest map: the use of 10-marker haplotypes was equally or less powerful than 6-marker haplotypes. However, this related to the additional distance covered by 10-marker haplotypes, because the markers furthest from the middle of the haplotype did not bring additional information as there may have been more recombinations. This was previously stated in studies on the fine-mapping efficiency of IBD probabilities which concluded that 4 and 6-marker haplotypes were the most efficient (Grapes *et al.*, 2006). In our simulations, we conclude that 6-marker haplotypes are better in terms of adequacy between the IBD probabilities and the real QTL IBD status, while Grapes *et al.* focused on the ability of locating correctly the QTL.

When the populations were submitted to selection, the frequency of IBD QTL having IBD probabilities between 0.5 and 0.8 strongly increased, corresponding to a decrease of the frequency of

IBD QTL with IBD probabilities over 0.9. On the other hand, the distributions of IBD probabilities for non IBD QTL were unaffected. As there are many computational difficulties linked to the use of these IBD probabilities (particularly the matrix being singular), one could group haplotypes with high IBD probabilities. According to the results of this study, if the population is submitted to selection, it seems possible to set a probability threshold for doing so to 0.5, corresponding to a type I error rate lower than 5%. A so low threshold is supported first by few occurrences of non-IBD QTL with IBD probabilities over 0.3, and second by a reduced number of haplotypes segregating in populations submitted to selection. Haplotypes are thus more frequently IBD, even when only the two central markers (*i.e.* the two markers flanking the tested position) are IBS. When looking at the shape of the overall distributions, there was a sudden increase of the frequencies in class 5 in selected populations, which appeared only in class 9 or 10 in unselected populations. It thus seems that looking at the overall distribution of the IBD probabilities may be an indicator of the clustering probability to use while maintaining a reasonable type I error rate.
Selection thus slightly decreased the correlation between the real IBD status and the IBD probabilities. It seemed that using 10-marker haplotypes was deleterious to that correlation in selected populations, comforting the idea that using 6-marker haplotypes was probably the best compromise for all situations.

References:

Cannon G.B., 1963. The effects of natural selection on linkage disequilibrium and relative fitness in experimental populations of Drosophila *melanogaster*, Genetics 48: 1201-1216.

Grapes L., Firat M.Z., Dekkers J.M.C., Rotschild M.F., Fernando R.L., 2006. Optimal haplotype structure for linkage disequilibrium-based fine mapping of quantitative trait loci using identity by descent. Genetics, 172: 1955-1965.

Meuwissen T.H.E., Goddard M.E., 2000. Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics, 155: 421-430.

Meuwissen T.H.E., Goddard M.E., 2001. Prediction of identity by descent probabilities from marker-haplotypes. Genet. Sel. Evol., 33(6): 605-34.

Yamazaki T., 1977. The effects of overdominance on linkage in a multilocus system. Genetics, 86: 227-236.

Zhao H.H., Fernando R.L., Dekkers J.C.M., 2007. Power and precision of alternate methods for linkage disequilibrium mapping of quantitative trait loci. Genetics, 175: 1975-1986