



Disclaimer: unless otherwise agreed by the Council of UPOV, only documents that have been adopted by the Council of UPOV and that have not been superseded can represent UPOV policies or guidance.

This document has been scanned from a paper copy and may have some discrepancies from the original document.

---

Avertissement: sauf si le Conseil de l'UPOV en décide autrement, seuls les documents adoptés par le Conseil de l'UPOV n'ayant pas été remplacés peuvent représenter les principes ou les orientations de l'UPOV.

Ce document a été numérisé à partir d'une copie papier et peut contenir des différences avec le document original.

---

Allgemeiner Haftungsausschluß: Sofern nicht anders vom Rat der UPOV vereinbart, geben nur Dokumente, die vom Rat der UPOV angenommen und nicht ersetzt wurden, Grundsätze oder eine Anleitung der UPOV wieder.

Dieses Dokument wurde von einer Papierkopie gescannt und könnte Abweichungen vom Originaldokument aufweisen.

---

Descargo de responsabilidad: salvo que el Consejo de la UPOV decida de otro modo, solo se considerarán documentos de políticas u orientaciones de la UPOV los que hayan sido aprobados por el Consejo de la UPOV y no hayan sido reemplazados.

Este documento ha sido escaneado a partir de una copia en papel y puede que existan divergencias en relación con el documento original.



**BMT/3/6**

**ORIGINAL : English**

**DATE : August 25, 1995**

**INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS**

**GENEVA**

**WORKING GROUP ON BIOCHEMICAL AND MOLECULAR TECHNIQUES  
AND DNA-PROFILING IN PARTICULAR**

**Third Session**

**Wageningen, Netherlands, September 19 to 21, 1995**

**THE ESTIMATION OF MOLECULAR GENETIC DISTANCES  
IN MAIZE FOR DUS AND ED PROTOCOLS: OPTIMIZATION  
OF THE INFORMATION AND NEW APPROACHES OF KINSHIP**

*Document prepared by experts from France*

**The estimation of molecular genetic distances in Maize for DUS and ED  
protocols :  
optimisation of the information and new approaches of kinship**

**C. Dillmann**<sup>1</sup>, A. Charcosset<sup>2</sup>, A. Bar-Hen<sup>3</sup>, B. Goffinet<sup>4</sup>, J. S. Smith<sup>5</sup>, Y. Dattée<sup>1</sup>, J. Guiard<sup>1</sup>

<sup>1</sup>GEVES, La Minière, F-78285 Guyancourt Cedex;

<sup>2</sup>INRA, Station de Génétique Végétale, Ferme du Moulon, F-91190 GIF/Yvette;

<sup>3</sup> The Institute of Statistical Mathematics, TOKYO;

<sup>4</sup>INRA, Biométrie, BP27, 31326 CASTANET-TOLOSAN Cedex;

<sup>5</sup> Pioneer Hi-Bred International, Inc. Plant Breeding Division, Department of Biotechnology Research, 750 NW 62nd Avenue, JOHNSTON, IA 50131 USA.

**Corresponding author :**

C. Dillmann, GEVES, La Minière, F-78285 Guyancourt Cedex

tel : 19-1-30-83-36-73

fax : 19-1-30-83-36-29

e-mail : [dillmann@diamant.jouy.inra.fr](mailto:dillmann@diamant.jouy.inra.fr)

The estimation of molecular genetic distances in Maize for DUS and ED  
protocols :  
optimisation of the information and new approaches of kinship

C. Dillmann, A. Charcosset, A. Bar-Hen, B. Goffinet, J. S. Smith, Y. Dattée, J. Guiard

GEVES, La Minière, F-78285 Guyancourt Cedex

## 1. Introduction

Maize (*Zea Mays* L.) can be considered as a reference species for the study of molecular genetic distances. The genetic map of the species is well known (Helentjaris et al, 1986, Hoisington, 1986, Burr et al, 1988) and large data sets pooling molecular, morphological and agronomical data are available. Moreover, the breeding work is essentially based on the obtention of parental inbred lines for hybrid varieties, which considerably simplify the studies on genetic distances. Therefore, the considerations developed in this paper mainly concern maize inbred lines and will be extended later to other species.

While answering two different questions, both Distinction (DUS) and future Essential derivation (ED) protocols imply the comparison of two maize lines. In the DUS approach, the question to answer is whether a new released variety is distinctly different from previously released varieties. The ED approach goes one step further in the comparison and considers the scope of breeder's right in more details. The question is whether a given inbred line B, distinct from another inbred line A, is predominately derivated from A (the initial inbred), while retaining the expression of the essential characteristics that result from the genotype or combination of genotypes of the initial line (UPOV convention, 1991). Both approaches require the computation of a distance index and the drawing up of rules of decision.

### 1.1 State of the art in DUS

Today, DUS is based on morphological and in some cases on biochemical traits (electrophoresis). In most countries, two inbred lines are distinctly different if they show at least one sufficient difference for one trait. In France, a phenotypic index named LCLM (Gruau, 1989) is computed from about 40 morphological and biochemical traits by affecting each trait with a different weight, depending on its genetic determinism and on the influence of the environment. Depending on the traits, one to three sufficient differences are necessary for two inbred lines to be declared distinct. The non distinct inbred lines are then submitted to the judgement of maize experts, who take the final decision.

### 1.2 State of the art in ED

Essential Derivation has been clearly defined by the UPOV convention in 1991. As soon as it will be ratified in the different countries, a plant breeder shall judge if a released inbred line B is essentially derivated from his own inbred A and may act before the court to inforces his rights on line B recognised. In this context, it is important to define precise tools allowing an objective evaluation of the distance between A and B and of the original breeding work accomplished on line B. From the UPOV convention, the notion of essential derivation is genetic. The point is therefore to estimate a posteriori the relatedness between the two lines.

The two approaches are illustrated on figure 1, where inbred lines B, C and D are distinct from inbred line A, but line B is within the sphere of essential derivation of line A. The difference between DUS and ED clearly appears here as a difference in the minimum distance between two lines. The purpose of this paper is (i) to discuss the elaboration of a decision protocol for DUS and ED and (ii) to apply it to the case of molecular distances.

## 2. Elaboration of a decision protocol

Any decision protocol involves three successive steps

- (1) Choice of a distance index
- (2) Determination of the minimum distance for distinction/essential derivation
- (3) Estimation of the distance from the data and test

which are to be detailed here.

### 2.1 Choice of a distance index

The choice of a distance index mostly releases on the choice of a discriminant trait.

Morphological traits describe the phenotype. They are the result of the expression of the genotype in a given environment. Based on morphological traits, several distance indexes have been proposed like Euclidean distance (Mahalanobis, 1936), empirical distances (LCLM), F1 yield or heterosis. Their relationship with pedigree or molecular data have been studied (Bar-Hen, 1993, Smith et al, 1990, Smith and Smith, 1989). Their obtention may be time-consuming and they poorly correlate with molecular data, showing that the same phenotype may be obtained from different genetic backgrounds. Therefore, morphological data are unable to discriminate closely related inbred lines. However, as the definition of distinction only involves the expressed part of the genome, they perfectly suit DUS studies.

Pedigree data may be difficult to obtain from plant breeders. Moreover, the kinship coefficient between two inbred lines (Malecot, 1948) doesn't take into account the selection process, nor the random genetic drift. When the parental inbred lines are themselves related, the kinship coefficient underestimates the real kinship.

Molecular data reflects both expressed and non expressed parts of the genome. The molecular distance between two inbred lines is the percentage of loci which differ between the two lines (Rogers, 1972, Nei, 1972). It is directly related to the kinship coefficient for unselected inbred lines (Cox et al, 1985, Lynch, 1988) and reflects the evolutionary pressures like selection, mutation and random genetic drift that acted on the lines. However, it doesn't take into account the existence of apparently identical allele which are not identical by descent. The relationship between the molecular distance and the coancestry coefficient has been studied in maize by Bernardo (1993). Finally, the molecular distance is calculated by sampling molecular markers throughout the genome. It is therefore only an estimation of the real genetic distance, whose precision depends on the sampling of markers..

### 2.2 Determination of the minimum distance

In the DUS approach, the question to answer is whether two inbred lines A and B are distinct or not. The distance between A and B has to be compared to the minimum distance from which two inbred lines may be declared distinct. Such a minimum distance may be assessed empirically. For example, the LCLM distance index used in France is based on the discriminatory power of morphological and biochemical traits and ranges from 0 to infinity. A minimum distance of 6 was set after a comparison between the LCLM value and the opinion of maize experts in a reference population of maize inbred lines.

The problem is slightly different in the ED approach, where the question is not only to assess whether the two inbred lines are essentially derivated or not, but also how much they are. The probability of two inbred lines A and B being essentially derivated may be computed by making hypotheses about the mode of derivation of B. Smith et al (1991) calculate the probability of recovering one parent by selfing from an hybrid ( $P_{F1}$ ) or from a single backcross ( $P_{BC1}$ ) to the recurrent parent. They consider the case of 40 unlinked polymorphic loci and show that the chances of coming up with a line more than 75% similar by selfing from a hybrid without selection are extremely low. The 75% level is

equivalent to that which could be achieved with a single backcross to either parent. With linkage, the probabilities  $P_{F1}$  and  $P_{BC1}$  would depend on the recombination rate between the loci *i.e.* on the genetic map. Figure 2 shows the probability of recovering one parent for different genetic maps and table 1 shows the probability of the distance to the recurrent parent to be lower than the minimum distance, for different values of the minimum distance. The effect of linkage is to increase the frequency of non-recombinant types, which correspond to the extreme distances to the recurrent parent. It therefore increases the variance of the genetic distance in the population, as illustrated in figure 2 by comparing the cases of 40 unlinked loci and 40 loci evenly spread on 10 linkage groups. Then, linkage enhances the chances of coming up with a line more than 75% similar by selfing from a hybrid. The latter is 20 times higher with 100 loci evenly spread on 10 linkage groups than with 40 unlinked loci. With repeated backcrosses, the frequency of the non-recombinant types similar to the recurrent parent increases faster than the frequency of the non-recombinant types similar the donor, thus decreasing the chances of coming up with a line less than 75% similar by selfing from a single backcross (table 1).

### 2.3 Estimation of the distance

#### *The Rogers distance*

The genetic distance between inbred lines A and B is defined as the percentage of loci which differ between the two lines. Let  $L$  be the total number of loci over the whole genome and  $d_{AB}^k$  be the distance between A and B at locus  $k$ . If lines A and B have the same allele at locus  $k$ ,  $d_{AB}^k = 0$ , else  $d_{AB}^k = 1$ . Then

$$d_{AB} = \frac{1}{L} \sum_{k=1}^L d_{AB}^k \quad (1)$$

Practically, the data only represent the sampling of  $M$  marker loci ( $M < L$ ). In most cases (see Smith, 1995), the sampling can be considered as random over the genome. The  $d_{AB}^k$ 's are then random variables, with the chance for  $d_{AB}^k$  to be equal to 1 being  $d_{AB}$ , and the chance for  $d_{AB}^k$  to be equal to 0 being  $(1 - d_{AB})$ . Hence, the Rogers distance

$$\hat{d}_{AB} = \frac{1}{M} \sum_{m=1}^M d_{AB}^m \quad (2)$$

is a random variable with a Binomial distribution :

$$P(d_{AB} = d) = \binom{M}{d} (d_{AB})^d (1 - d_{AB})^{(M-d)} \quad (3)$$

Figure 3 show the distribution of  $\hat{d}_{AB}$  for  $M=100$  loci and  $d_{AB} = 0.2$ . The expectation of  $\hat{d}_{AB}$  is

$$E(\hat{d}_{AB}) = d_{AB} \quad (4a)$$

and it's variance is

$$Var(\hat{d}_{AB}) = \frac{d_{AB}(1 - d_{AB})}{M} \quad (4b)$$

It is therefore possible to construct a confidence interval around the Rogers distance estimated from the data, which would depend on  $d_{AB}$  and on the sample size.

If  $M$  is large enough, the distribution of  $\hat{d}_{AB}$  may be approximated by a Normal distribution with parameters  $\mu = d_{AB}$  and  $\sigma = \sqrt{\frac{d_{AB}(1-d_{AB})}{M}}$ .

#### General case

More generally, a linear estimator of  $d_{AB}$  can be any linear combination of the  $d_{AB}^k$ 's :

$$d'_{AB} = \sum_{m=1}^M a_m d_{AB}^m \quad (5)$$

with expectation  $E(d'_{AB}) = \left(\sum_{m=1}^M a_m\right) d_{AB}$  and with variance

$$Var(d'_{AB}) = \left[ \sum_{m=1}^M a_m^2 Var(d_{AB}^m) + \sum_{m=1}^M \sum_{k \neq m} a_m a_k Cov(d_{AB}^m, d_{AB}^k) \right] \quad (6)$$

If the markers are sampled at random over the genome, there is no correlation between distances at marker loci and the variance of the new estimator becomes  $Var(d'_{AB}) = \left(\sum_{m=1}^M a_m^2\right) d_{AB}(1-d_{AB})$ .

In case of non-zero correlation's between distances at marker loci, the covariance terms in (6) have to be taken into account in the computation of the variance of the estimator.

#### Properties

The best estimator has a minimum variance, so that the confidence interval around the estimation is as small as possible, given the data. Another desirable property for an estimator is unbiasedness, *i.e.* its expectation is equal to the parameter to estimate. Note that with random sampling of the markers, the Rogers distance is the best linear unbiased estimator of the real genetic distance between two inbred lines.

### 2.4 Hypothesis testing

Suppose that a distance index has been chosen and that a minimum distance  $d_{MIN}$  has been set. The problem is now to test the hypothesis  $H_0: d_{AB} \leq d_{MIN}$  against the alternative  $H_0: d_{AB} > d_{MIN}$ , given the data. The decision rule will have the form :

(1) Calculate  $\hat{d}_{AB}$

(2) If  $\hat{d}_{AB} \leq d_T$  then accept  $H_0$ , else reject  $H_0$ .

where  $d_T$  depends on  $d_{MIN}$  and on the level of significance  $\alpha$ .

Table 2 gives the exact value of  $d_T$  for the test using the Rogers distance for different values of  $d_{MIN}$  and for two sample sizes. Note that  $d_T$  is also the upper value of the confidence interval for  $\hat{d}_{AB}$  at the level of significance  $2\alpha$  (bilateral test). It can be seen that the discrepancy between  $d_{MIN}$  and  $d_T$  is relatively high and ranges from one half to one third of the  $d_{MIN}$  value. Figure 4 gives the power of the test for two different values of  $d_{MIN}$  and show that increasing the sample size to more than  $M=120$  marker loci doesn't significantly increases the power of the test. Those exact results are in agreement with previous results obtained with Jackknife (Melchinger et al, 1991, Bernardo, 1993) or bootstrapping (Tivang et al, 1994).

### 3. Application to molecular distances : taking the genetic map into account

#### 3.1 The Rogers distance

We have previously seen that when the marker loci are a random sample of the genome, the Rogers distance is the best linear unbiased estimator of the genetic distance between two inbred lines. In some cases however, it is not possible to consider the markers as a random sample. For example, the AFLP method may produce several markers with only one probe, but the latter seem to be very close on the genetic map and cannot be considered as independent. Moreover, as the number of available markers in species like maize is becoming very large, it is frequent to select a shorter set of markers for routine studies. If the markers are selected on their position in the genetic map, the distances at marker loci may become dependent from each other when the inbred lines under study are related.

As a matter of fact, relatedness induces multilocus identity by descent and the apparition of genomic blocks within which all the loci are identical by descent. Random sampling of markers destroys those genomic blocks, which therefore only appear when the marker loci are non independent. In this case, the Rogers distance is no more the best linear estimator because it doesn't take into account the covariance's between distances at marker loci.

Two methods can be proposed to overcome the problem of non random sampling of the marker loci. The first one is to practise stratified sampling and the second one is to find a new estimator with minimum variance by using (6).

#### *Stratified sampling*

Consider the following situation : (i) a genetic map has been established, which includes the extremities of the chromosomes, (ii) based on this map, the genome is divided into M segments of cM each, (iii) a very large number of probes is available, so that it is possible to select sets of probes that fulfil the condition that each segment contains one, and only one probe. In this case, for a given segment s, two inbred lines are identical for a proportion  $p_s$  of the segment, different for a proportion  $1 - p_s$  of the segment. The variance of distance estimation, over the set of probes that answer condition (iii) is :

$$\sigma_{cond}^2 = \frac{1}{M^2} \sum_{s=1}^M p_s (1 - p_s) \quad (7)$$

The difference between this variance and the variance of the estimations under the hypothesis of random sampling of the loci (4b) is

$$\sigma_{cond}^2 - Var(\hat{d}_{AB}) = -\frac{1}{M^2} \sum_{s=1}^M (p_s - d_{AB})^2 < 0 \quad (8)$$

and depends on the variation of  $p_s$  over the segments. This illustrates that condition (iii) leads to a decrease in the variance of the estimate. Thus, as could be expected, choosing loci to optimise genome coverage increases the precision of the estimates. It is a classical result of stratified versus random sampling (Cochran, 1977). However, since  $p_s$  cannot be estimated, it is not possible to derive an estimator of the variance of  $\hat{d}_{AB}$  from formula (8).

#### *Covariance's between distances at marker loci conditionally to the genetic map*

In section 2.3, we have shown that when the markers are not a random sample of the genome, relatedness between inbred lines induces covariance's between distances at marker loci. In this section, two methods will be proposed to estimate those covariance's conditionally to the genetic map. Consider the following points :

(1) In the DUS and ED approaches, the test of  $H_0: d_{AB} \leq d_{MIN}$  implicitly consists in testing the hypothesis that the two inbred lines A and B have a common genetic history and that they can be



related to a given population. For example, in the ED approach, one can wish to test the hypothesis that lines A and B are both derived by successive backcrosses from the same hybrid.

(2) The determination of an estimator of the genetic distance requires some knowledge on the sampling distribution of the estimator, *i.e.* on the variation of the information carried by each locus involved in the comparison of a given couple of inbred lines distant of  $d_{AB}$ .

(3) If the loci are independent, it is equivalent to consider the variation of the information carried by each locus for the couple AB, and the variation of the information carried by one given locus over the subset of couples of inbred lines of the population from which A and B originated that are distant of  $d_{AB}$ .

(4) Similarly, it is equivalent to consider the variation of the information carried by all the couples of loci distant of  $c$  centimorgans in the genetic map and involved in the comparison of A and B, and the variation of the information carried by a given couple of loci distant from  $c$  centimorgans over the subset of couples of inbred lines of the population which are distant of  $d_{AB}$ .

(5) Reasoning conditionally to a genetic map consists in considering the sampling of  $M$  markers not over the whole genome, but over the subset of loci which preserve the genetic map. In other words, the subset of loci which preserve the distances between loci. For a given distance matrix, there are a large number of different but equivalent genetic map, as illustrated by figure 5. In this example, loci numbered 1,5,6 and 7 are independent and may be located anywhere in different arms of different chromosomes. Loci numbered 2,3 and 4 are linked but the distance matrix is the same whether locus 3 is closer to locus 2 or to locus 4.

(6) Therefore, reasoning conditionally to the genetic map consists, for a given couple of inbred lines, in measuring the variation of the information carried by the set of couples of loci distant from  $c$  centimorgans in the genetic map. From (3) and (4), this is equivalent to measuring the information carried by a given couple of loci distant from  $c$  centimorgans over the subset of couples of inbred lines of the population which are distant of  $d_{AB}$ .

Therefore, if  $R_{pop}^{mn}$  is the correlation between the genetic distance at locus  $m$  and the genetic distance at locus  $n$  in the population from which A and B originated, the covariance of the distance between marker loci, conditionally to the genetic map is equal to

$$Cov_{map}(d_{AB}^m, d_{AB}^n) = d_{AB}(1 - d_{AB})R_{pop}(d_{AB}^m, d_{AB}^n) \quad (9a)$$

where

$$R_{pop} = \frac{Cov_{pop}(d_{AB}^m, d_{AB}^m)}{Var_{pop}(d_{AB})} \quad (9b)$$

(9a) is the covariance for the genetic distance at marker loci  $m$  and  $n$  in the subset of lines of the population that are distant of  $d_{AB}$ .

Hence, using (9a) is a quite simple way to compute the theoretical value of the covariance between distances at marker loci, given some hypothesis about the population from which the lines to be compared originated. It is then possible, by using (6) to compute the sampling variance of any estimator of the genetic distance conditionally to the genetic map.

If there is no idea about the population from which A and B originated, it is also possible to estimate  $R_{pop}^{mn}$  from the data.

Theoretical value of  $R_{pop}^{mn}$  : Table 3 gives theoretical values of  $E_{pop}$ ,  $Var_{pop}$  and  $Cov_{pop}$  for different breeding schemes and without selection. Note that in this case,  $R_{pop}$  only depend on the recombination rate between the loci. For example, if A and B are haplodiploidized from an hybrid,  $R_{pop}^{mn} = (1 - 2r_{mn})^2$  and the conditional variance of the Rogers distance becomes

$$Var_{map}(\hat{d}_{AB}) = \frac{d_{AB}(1-d_{AB})}{M^2} \left[ M + \sum_m \sum_{n \neq m} (1 - 2r_{mn})^2 \right] \quad (10)$$

where  $r_{mn}$  is the recombination rate between marker loci  $m$  and  $n$  obtained from  $c_{mn}$  by using the Haldane's mapping function (Haldane, 1919). In a general case,  $R_{pop}$  depends on allelic frequencies and on the linkage disequilibrium between the two loci (Charcosset and Essioux, 1994). It is related to the two locus kinship coefficient, describing the situations of identity by descent at two locus, and being equal to zero when the alleles at locus  $m$  and  $n$  are non identical by descent *i.e.* when A and B are unrelated.

Estimation of  $R_{pop}^{mn}$  from the data : If there is no information on the population from which A and B originated, it is possible to use (9a) to estimate the conditional covariance's between distances at marker loci from the data, knowing that if A and B are related, then  $R_{pop}$  is a function of the recombination rate between the markers :

$$R_{pop}^{mn} = a + br_{mn} + cr_{mn}^2 + \dots + \varepsilon \quad (11)$$

where  $\varepsilon$  is a random error. For example, if there are  $M$  marker loci spread every  $c$  centimorgans on one chromosome, there are  $M(M-1)/2$  couple of loci. Therefore, for a given couple of inbred lines,  $(M-1)$  data are available to estimate the covariance between distances at marker loci separated from  $c$  cM,  $(M-2)$  data to estimate the covariance between distances at marker loci separated from  $2c$  cM, and so on. By rearranging the data, it is possible to create a new variable  $Y$  with  $M(M-1)/2$  entries, which takes the value  $\hat{d}_{AB}^2$  if the lines have the same allele at both loci,  $-\hat{d}_{AB}(1-\hat{d}_{AB})$  if the lines have the same allele at one locus and a different allele at the other locus and  $(1-\hat{d}_{AB})^2$  if the lines have different alleles at both loci. (11) can then be estimated from  $Y$  by non linear regression methods.

To illustrate this approach, two data sets were considered. The first one (data set 1) was supplied by A Charcosset and consists of 50 recombinant inbred lines derived from the cross F2 x Io (Charcosset et al, 1994). Molecular data were obtained from 133 public RFLP probes. The second data set (data set 2) was supplied by JS Smith and consists of 37 highly selected, elite inbred lines of maize representing a broad range of diversity in coefficient of parentage (0 to 95%) from the central US Corn Belt (Smith et al, 1990). Molecular data were obtained from 110 RFLP probes supplied by Ben Burr (Brookhaven National Laboratory) or Dave Hoisington (University of Missouri-Columbia). Non linear regression was performed by maximum likelihood with the function 'nls' of Splus (Splus, Statistical sciences Inc.) using a Gauss-Newton algorithm.

Figure 6A shows the relationship between the value of  $R_{pop}$  estimated from the data and the recombination rate for 4 couples of inbred lines from data set 1. It is slightly lower than expected but the results for couples 6, 7 and 8 are significant. Figure 6B is similar for 4 couples of inbred lines from data set 2. This illustrates the fact that it is possible to estimate  $R_{pop}$  from the data. However, the precision of the estimation is difficult to assess, except for asymptotic results and is probably quite bad.

Figure 7 shows the relationship between the slope of  $R_{pop}$  and the Rogers distance in the two data sets. As would be expected, there is no correlation in data set 1, which consist of non selected

recombinant inbred lines for which  $R_{pop}$  does only depend on the recombination rate between the loci (10). On the contrary, there is a triangular relationship between  $R_{pop}$  and the Rogers distance in data set 2, which consist of more or less related, highly selected inbred lines. High Rogers distances are associated with zero or positive slopes, while low Rogers distances are associated with highly negative or zero slopes. The effect of selection on  $R_{pop}$  is still to explore but it can be noticed that the lowest slope is obtained with the couple (PA632, C4) which are two BSSS lines with a kinship coefficient of 0.82.

The lack of precision in the estimation of  $R_{pop}$  from the data precludes its direct utilisation to compute the conditional covariance's. However, it should be possible to test several hypothesis about the population from which the lines to be compared originated. For each of the hypothesis, the theoretical value of  $R_{pop}$  may be assessed. The data can then be used to compute the likelihood of each hypothesis and choose the one with the highest probability (the highest LOD score).

### 3.2 The ED case

Once a model has been set for  $R_{pop}$ , it is possible to use (6) to find the linear combination of the  $d_{AB}^k$ 's which would lead to the minimum variance estimator of  $d_{AB}$ , conditionally to the genetic map (Searle, 1971). The problem is to find an appropriate weight for each marker locus in order to take into account the redundancy of the information when distances at marker loci are correlated. Let call this new distance the ED distance.

Let  $V$  be the matrix of variances and covariance's of the marker loci, conditionally to the genetic map. For example, with three loci,  $V$  is equal to

$$V = d_{AB}(1 - d_{AB}) \begin{bmatrix} 1 & R_{pop}^{12} & R_{pop}^{13} \\ R_{pop}^{12} & 1 & R_{pop}^{23} \\ R_{pop}^{13} & R_{pop}^{23} & 1 \end{bmatrix} \quad (12)$$

Let  $e_{mn}$  be the element of  $V$  inverse corresponding to the marker loci  $m$  and  $n$ . Then, the appropriate weight for each marker locus  $m$  is

$$a_m = \frac{\sum_{n=1}^M e_{mn}}{\sum_{n=1}^M \sum_{l=1}^M e_{nl}} \quad (13a)$$

and the best linear unbiased estimator of  $d_{AB}$ , conditionally to the genetic map is

$$\tilde{d}_{AB} = \sum_{m=1}^M a_m d_{AB}^m \quad (13b)$$

which is called the ED distance. The variance of  $\tilde{d}_{AB}$  is equal to

$$Var(\tilde{d}_{AB}) = \frac{1}{\sum_{n=1}^M \sum_{l=1}^M e_{nl}} \quad (14)$$

Note that when  $R_{pop}$  is equal to 0, the ED distance is equal to the Rogers distance. Otherwise, each marker locus is weighted by a coefficient which depend on its position on the genetic map.

Those coefficients are compared to the coefficients of the Rogers distance on figure 8 in a trivial case with 10 loci randomly spread in one chromosome of 100 cM. Loci closed together have a small coefficient like loci 1 and 2 or loci 7,8 and 9 of figure 8, while isolated loci like locus 5 and 6 share a

higher coefficient. The high coefficients of the loci situated at the extremity of the chromosome are a side effect of the model, resulting from the reasoning conditionally to the genetic map.

Figure 9 shows the coefficients of the marker loci for the maize genetic map provided by M. Causse (Causse, 1995). As previously, the closer the loci are, the lower the coefficient of the ED genetic distance. Table 4 compares the ED distance and the Rogers distance for the eight couples of inbred lines from data sets 1 and 2 previously studied. It can be seen that the variance of the ED distance is always about 20 percent lower than the variance of the Rogers distance, thus increasing the precision of the estimation.

The ED variance seems perfectly suited for ED problems when markers are not randomly sampled. At first, it lowers the conditional variance significantly. At second, it is dependent on the hypothesis about the relatedness between the two lines to be compared via the model chosen for  $R_{pop}$ . It is therefore specific of the couple of lines to compare.

Once the data are collected, the following ED protocol may be proposed

1. Chose an appropriate set of markers with known position on the genetic map.
2. Calculate the distance between the two lines at each marker locus.
3. Use the data to find the best suited value of  $R_{pop}$
4. Use  $R_{pop}$  to compute the coefficients of each marker locus.
5. Calculate the ED genetic distance and its variance
6. If the number of markers is large enough, a confidence interval may be computed by using a normal distribution with mean  $\tilde{d}_{AB}$  and variance  $Var(\tilde{d}_{AB})$ .

### 3.4 The DUS case

The problem is slightly different in DUS studies where a new inbred line has to be compared to the whole set of previously released inbred lines. It may be desirable for the estimator of the genetic distance to have the same variance for all the comparisons so that each distance may be assessed with the same precision. Instead of reasoning conditionally, it is possible to consider the set of markers as fixed, the variability coming from the information at the marker locus, i.e. from the variability of the inbred lines of the population. The distance to measure here is the distance  $d_{AK}$  between a given new line  $K$  and an inbred line  $A$  considered as a random sample of the population. Then, the optimum estimator would be the one which minimise the mean squared error (MSE) between the real genetic distance and the estimation in the population,

$$MSE = E_{pop}(d_{AK} - d'_{AK}) \quad (15)$$

thus leading to the best linear unbiased predictor of  $d_{AK}$  in the population. This approach is quite similar to the BLUP approach in plant or animal breeding (Henderson, 1975). Let call this estimator the DUS genetic distance.

With only one locus, the new estimator would be

$$d_{AK}^* = E_{pop}(d_{AK}) + \frac{Cov_{pop}(d_{AK}, d_{AK}^m)}{Var_{pop}(d_{AK}^m)} (d_{AK}^m - E_{pop}(d_{AK}^m)) \quad (16)$$

Using the same notations as in (13) and defining

$$g_k = Cov_{pop}(d_{AK}, d_{AK}^m) \quad (17)$$

the new weight of each marker locus over a total of  $M$  markers is,

$$a_m^* = a_m \left( 1 - \sum_{n=1}^M \left( \sum_{l=1}^M g_l e_{nl} \right) \right) + \left( \sum_{l=1}^M g_l e_{ml} \right) \quad (18)$$

the  $e_{mn}$  's being this time the elements of the inverse of the variance-covariance matrix  $V^*$  in the whole population. With three markers, we have, for example,

$$V^* = \text{Var}_{pop}(d_{AB}) \begin{bmatrix} 1 & R_{pop}^{12} & R_{pop}^{13} \\ R_{pop}^{12} & 1 & R_{pop}^{23} \\ R_{pop}^{13} & R_{pop}^{23} & 1 \end{bmatrix}$$

As previously, the covariance's between distances at marker loci may be either estimated from the data which consist of the whole population of released inbred lines, or computed theoretically, making hypothesis on the relatedness between the individuals of the population. This would be the case if the DUS genetic distance is used for ED approaches. If the total number of loci in the genome is large enough,  $g_k$  can then be computed as

$$g_k = \int_0^{c_{mi}} \text{Cov}_{pop}(d_{AK}^i, d_{AK}^m) + \int_0^{c_{mf}} \text{Cov}_{pop}(d_{AK}^f, d_{AK}^m) \quad (19)$$

where  $i$  and  $f$  are the beginning and the end positions of the chromosome of  $m$ . Using table 3, it can be shown that  $g_k$  is maximum in the middle of a chromosome and minimum at the two ends of the chromosome. This was to be expected from the fact that a marker in the middle of a chromosome has more adjacent loci than a marker at the end of the same chromosome. If the  $g_k$  's are equal to zero, the DUS distance reduces in the ED distance.

A comparison of the Rogers distance, the DUS distance and the ED distance is presented on figure 8, considering a population of inbred lines derived from an hybrid and line  $K$  being one of them. It can be seen that the two distances are very close together, except for the side effect in the ED distance, which is moderated in the DUS by the  $g_k$  coefficient. However, the expected variance seems to be larger with the DUS distance than with the ED distance.

It is therefore possible to propose the following DUS protocol based on molecular markers :

1. Choose an appropriate set of markers with known position on the genetic map.
2. Choose a value for  $d_{MIN}$ .
2. Enter the molecular data for each newly released inbred line into a database.
3. Calculate the distances between each pair of lines at each marker locus.
3. Use those data to find the best suited value of  $R_{pop}$
4. Use  $R_{pop}$  to compute the coefficients of each marker locus and the sampling variance of the genetic distance.
5. For each new entry, compute the DUS genetic distance to each line of the database and realise the test  $H_0: d_{AK} \leq d_{MIN}$ .
6.  $R_{pop}$  can be re-evaluated each year.

The above protocol clearly introduces a new concept in DUS by considering the previously released inbred lines as a population. Such populations do exist *de facto* for the inbreds belonging to the same combining ability group or having the same origin. It is quite possible to realise the DUS protocol for each of these *de facto* populations. This would be a natural extension of the French approach of distinction, which consider the danger of founding distinctness on only one trait.

#### 4. Conclusion

The aim of the present paper was to present clearly some of the statistical and genetical implications of the DUS and ED approaches. It emphasises the problem of the correct estimation of the genetic distance and its sampling variance which may be particularly accurate when the former are used to construct phylogenetic trees or dendrograms (Felsenstein, 1985).

Two new estimators were proposed, taking into account the genetic map. DUS and ED distances were distinguished in the text because they represent different philosophies. However, they were shown to produce very similar results and it is quite possible to use the ED distance in DUS approaches and conversely. They both seem to have interesting properties which have to be confirmed by (i) checking by simulation that the real sampling variance is equal to its theoretical value, and (ii) finding a method to properly choose between the different models for  $R_{pop}$ , the correlation between distances at marker loci.

The precision of the estimation of  $R_{pop}$  appears to be the crucial point of the study. However, it is also a new approach of kinship, combining the information carried by single loci to the 2-loci information, the latter being more persistent through time because of linkage.

Finally, those approaches can be extended to other species where the genetic map is not so complete as in maize by , for example, providing a method to weight the information carried out by different chromosomes.

## References

Bernardo R (1993) Estimation of the coefficient of coancestry using molecular markers in maize. *Theor Appl Genet* 85:1055-1062

Bernardo R (1993) Estimation of the coefficient of coancestry using molecular markers in maize. *Theor Appl Genet* 85:1055-1062

Burr B Burr FA Thompson K Albertsen M Stuber CW (1988) Gene mapping with recombinant inbreds in maize. *Genetics* 118:519-526

Charcosset A Cause M Moreau L Gallais A (1994) Investigation into the effect of genetic background on QTL expression using 3 connected maize Recombinant Inbred Line populations. Proc 9th meeting of the Eucarpia section : Biometrics in plant breeding : applications of molecular markers. Eds JW Van Ooijen and J Jansen

Charcosset A Essioux L (1994) The effect of heterosis on the relationship between heterosis and heterozygosity at marker loci. *Theor Appl Genet* (ref ?)

Cochran WG (1977) *Sampling techniques*. Wiley, 3d Ed.

Cox TS Kiang YT Gorman MB Rodgers DM (1985) Relationship between coefficient of parentage and genetic similarity indices in the soybean. *Crop Sci* 25:529-532

Felsenstein J (1985) Confidence limits on phylogenies : an approach using the bootstrap. *Evolution* 39:783-791

Haldane JBS (1919) The combination of linkage values, and the calculation of distance between the loci of linked factors. *J Genet* 8:299-309

Helentjaris T Slocum M Whright A Scharfer A Nienhuis J (1986) Construction of genetic linkage maps in maize and tomato using restriction fragment length polymorphism's. *Theor Appl Genet* 72:761-769

Henderson CR (1975) Best linear unbiased estimation and prediction. *Biometrics* 31:423-449

Hoisington D (1986) Maize working maps. *Maize Genet Coop Newsl* 60:142

Lynch M (1988) Estimation of relatedness by DNA fingerprinting. *Mol Biol Evol* 5:584-599

Malecot G (1948) *Les mathématiques de l'hérédité*. Masson & Cie, Paris.

Melchinger AE Messmer MM Lee M Woodman WL Lamkey KR (1991) Diversity and relationships among US maize inbreds revealed by restriction fragment length polymorphism. *Crop Sci* 31:669-678

Nei M (1972) Genetic distances between populations. *Am Nat* 106:283-292

Rogers JS (1972) Measures of similarity and genetic distance. *Studies in Genetic VII*. Univ Tex Publ 7213:145-153

Smith JSC Smith OS (1989) The description and assessment of distance between inbred lines of maize : I The use of morphological traits as descriptors. *Maydica* 34:141-150

Smith JSC Smith OS Bowen SL Tenborg RA Wall JS (1991) The description and assessment of distance between inbred lines of maize : A revised scheme for testing of distinctiveness between inbred lines utilizing DNA RFLP's. *Maydica* 36:213-226

Smith OS Smith JSC Bowen SL Tenborg RA Wall JS (1990) Similarities among a group of elite maize inbreds as measured by pedigree, F1 grain yield, grain yield and RFLP's. *Theor Appl Genet* 80:833-840

Tivang JG Nienhuis J Smith OS (1994) Estimation of sampling variance of molecular marker data using bootstrap procedure. *Theor Appl Genet* 89:259-264

Bar Hen A (1994) Généralisation de la distance de Mahalanobis au cas du mélange de variables continues et discrètes. Application aux études de distinction variétale. Thèse de l'Université de Paris-Sud, Orsay, 179pp

Bar Hen A Charcosset A (1995) Precision of the estimation of genetic distance between inbred lines using molecular markers. Submitted

Gruau (1989) Logiciel de Comparaison de Lignées de Maïs.

Mahalanobis PC (1936) On the generalized distances in statistics. *Proc Nat Inst Sci India* 2:49-55

Searle SR (1971) *Linear Models*. Wiley

| genetic map                                | breeding scheme <sup>(3)</sup> | $d_{MIN}$ |        |        |
|--|--------------------------------|-----------|--------|--------|
|  |                                | 0.7500    | 0.8000 | 0.8500 |
| 40 loci, unlinked                          | F1                             | 0.0003    | 0.0002 | 0.0000 |
|  | BC1                            | 0.4395    | 0.1819 | 0.0436 |
|  | BC2                            | 0.9773    | 0.8809 | 0.6160 |
| 40 loci, 10 linkage groups <sup>(1)</sup>  | F1                             | 0.0048    | 0.0009 | 0.0001 |
|  | BC1                            | 0.4609    | 0.2405 | 0.0870 |
|  | BC2                            | 0.9528    | 0.8435 | 0.6116 |
| 100 loci, 10 linkage groups <sup>(2)</sup> | F1                             | 0.0063    | 0.0012 | 0.0001 |
|  | BC1                            | 0.4970    | 0.2694 | 0.1014 |
|  | BC2                            | 0.9640    | 0.8730 | 0.6595 |

**Table 1** : Probability for the distance to the recurrent parent to be lower than  $d_{MIN}$  for different breeding schemes and different genetic maps.

<sup>(1)</sup> loci evenly spaced on the map,  $d=50$  cM. <sup>(2)</sup> loci evenly spaced on the map,  $d=16.7$  cM.

<sup>(3)</sup> F1=selving from an hybrid, BC1= selving from a single backcross, BC2= selving from a double backcross.



| $d_{MIN}$ | $\alpha$ | $M$ | $d_T$ |
|-----------|----------|-----|-------|
| 0.05      | 0.02     | 80  | 0.120 |
|           |          | 160 | 0.083 |
| 0.10      | 0.02     | 80  | 0.180 |
|           |          | 160 | 0.150 |
| 0.15      | 0.02     | 80  | 0.250 |
|           |          | 160 | 0.210 |
| 0.20      | 0.02     | 80  | 0.320 |
|           |          | 160 | 0.260 |
| 0.25      | 0.02     | 80  | 0.370 |
|           |          | 160 | 0.320 |

**Table 2** : Value of the effective minimum distance for the test of  $H_0: d_{AB} \leq d_{MIN}$  with the Rogers distance.

| population | $E_{pop}$      | $Var_{pop}$        | $Cov_{pop}(d_{AB}^m, d_{AB}^n)^{(1)}$                                  |
|------------|----------------|--------------------|--|
| F1         | $\frac{1}{2}$  | $\frac{1}{4}$      | $\frac{e^{-4*c_{mn}}}{4}$  |
| BC1        | $\frac{3}{8}$  | $\frac{15}{64}$    | $\frac{(1 - e^{-2*c_{mn}})^4 - 1}{64}$                                 |
| BC2        | $\frac{7}{32}$ | $\frac{175}{1024}$ | $\frac{((1 - e^{-2*c_{mn}})^3 - 1)((1 - e^{-2*c_{mn}})^3 + 17)}{1024}$ |

**Table 3** : Mean, variance and covariance between distances at marker loci in a population of inbred lines derived by selfing and without selection from F1 : an hybrid, BC1 : a single backcross and BC2 : a double backcross.

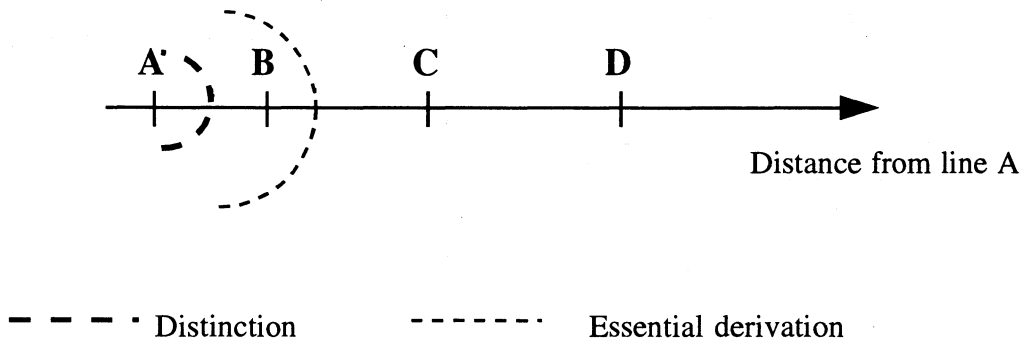
<sup>(1)</sup>  $c_{mn}$  is the distance between marker loci  $m$  and  $n$  expressed in morgans.

| dataset          | inbred lines | Rogers distance |                         | ED distance |          |
|------------------|--------------|-----------------|-------------------------|-------------|----------|
|                  |              | value           | variance <sup>(1)</sup> | value       | variance |
| 2                | B7-C5        | 0.318           | 0.0086                  | 0.298       | 0.0069   |
|                  | PA632-C4     | 0.145           | 0.0049                  | 0.141       | 0.0040   |
|                  | C5-B5        | 0.518           | 0.0099                  | 0.484       | 0.0082   |
|                  | PB73- PA4    | 0.482           | 0.0099                  | 0.576       | 0.0080   |
| 1 <sup>(2)</sup> | L3-L74       | 0.233           | 0.0014                  | 0.216       | 0.0011   |
|                  | L54-L139     | 0.173           | 0.0011                  | 0.170       | 0.0009   |
|                  | L42-L67      | 0.398           | 0.0018                  | 0.434       | 0.0016   |
|                  | L82-L111     | 0.331           | 0.0017                  | 0.318       | 0.0014   |

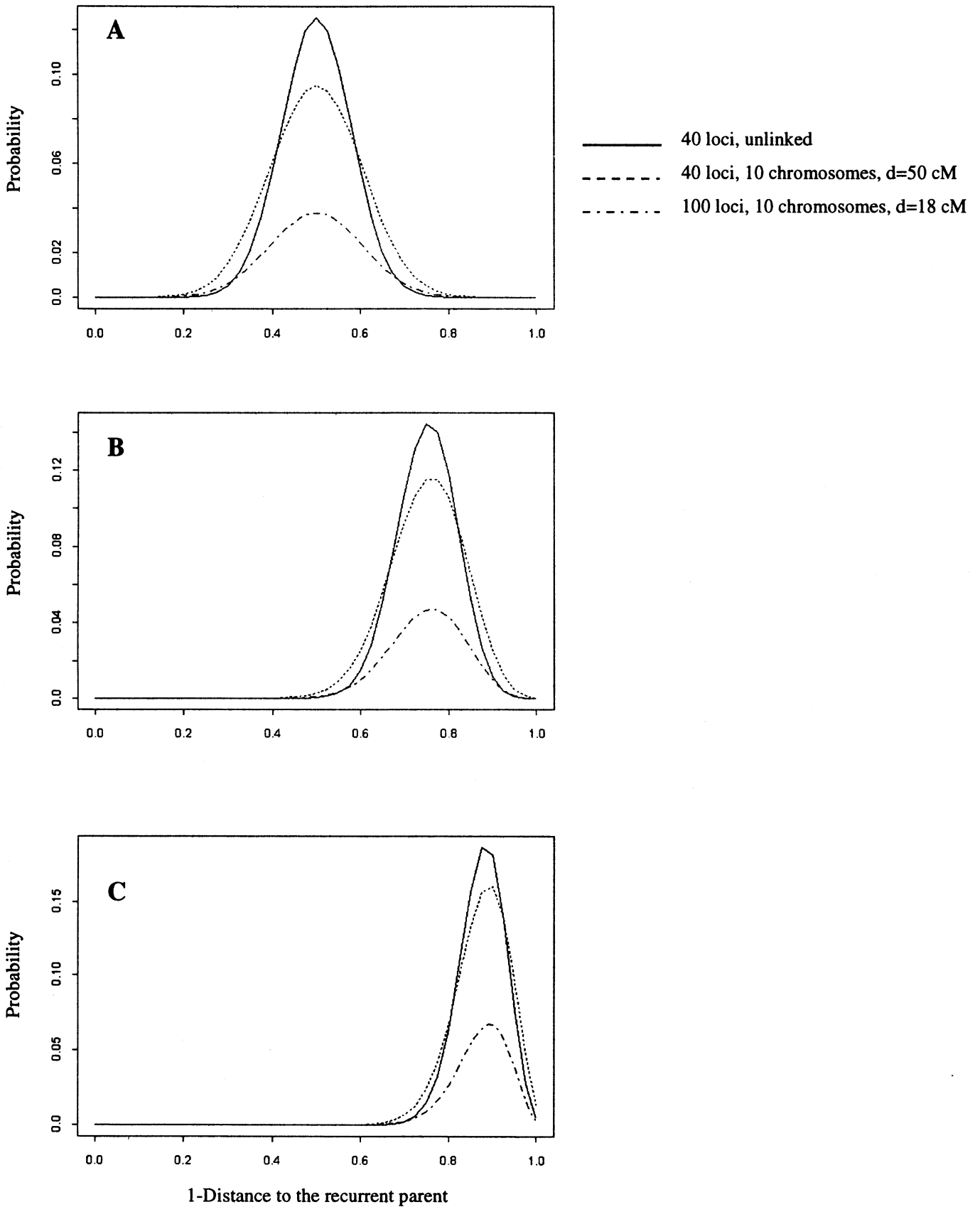
**Table 4** : Comparison of the Rogers distance and the ED distance for the eight couples of inbred lines coming from data sets 1 and 2.

<sup>(1)</sup> Conditionally to the genetic map.

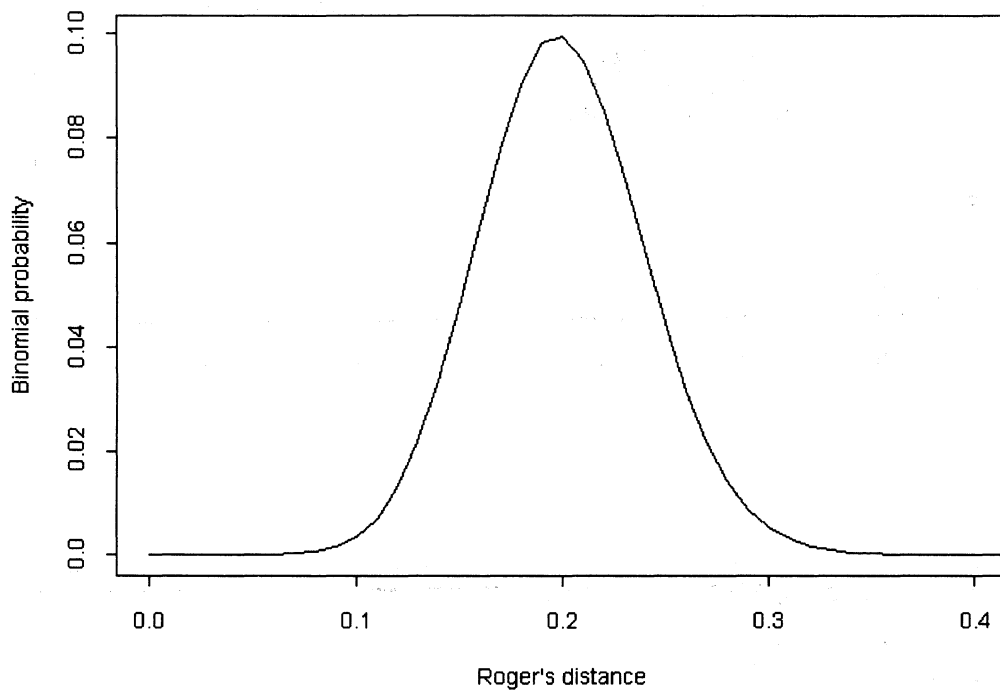
<sup>(2)</sup> L3-L74=couple 5 in figure 6, L54-L139=couple 6, L42-L67=couple 7, L82,L111=couple 8.



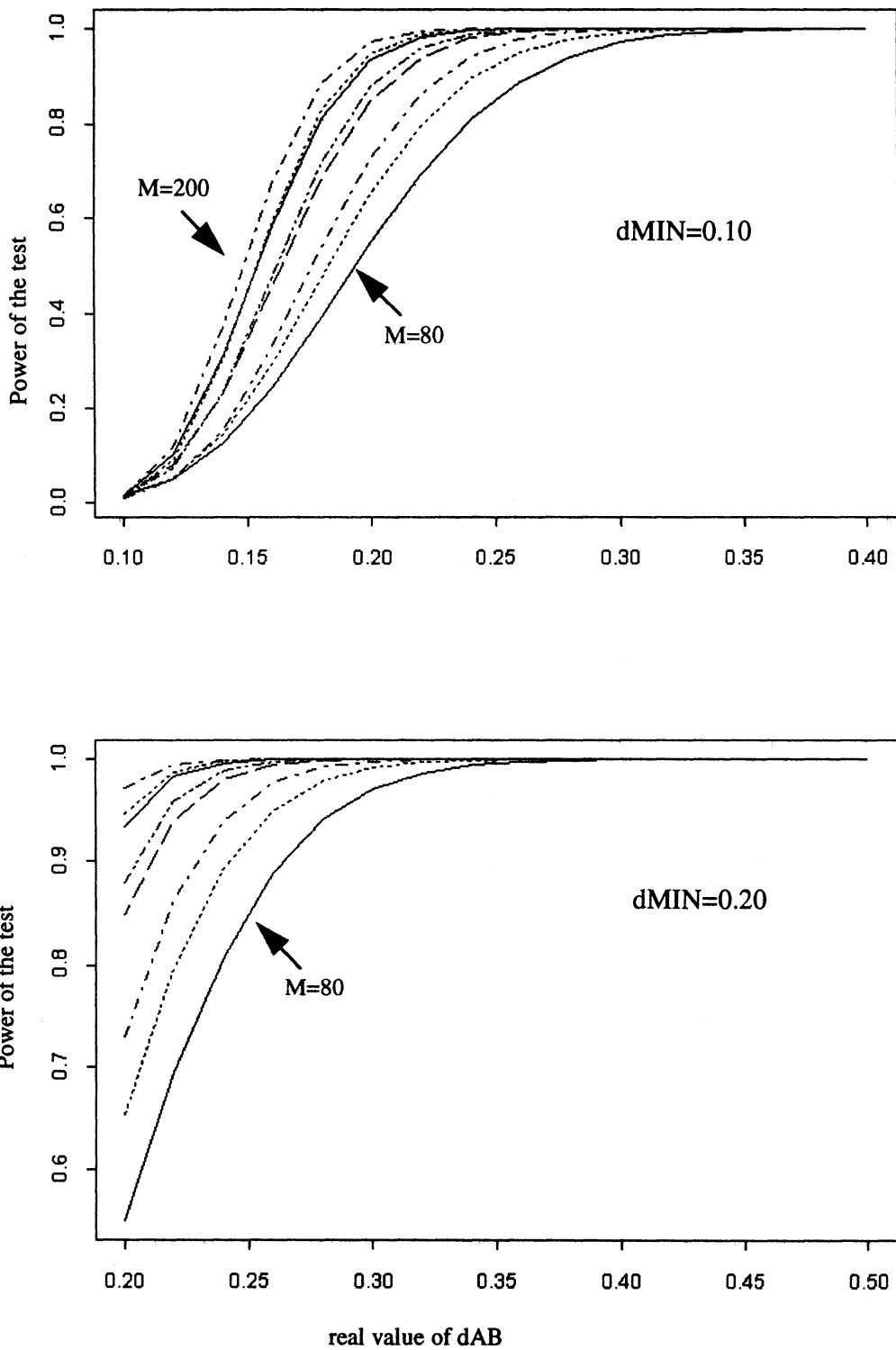
**Figure 1** : Schematic representation of the relationships between four inbred lines in Distinction and in Essential Derivation.



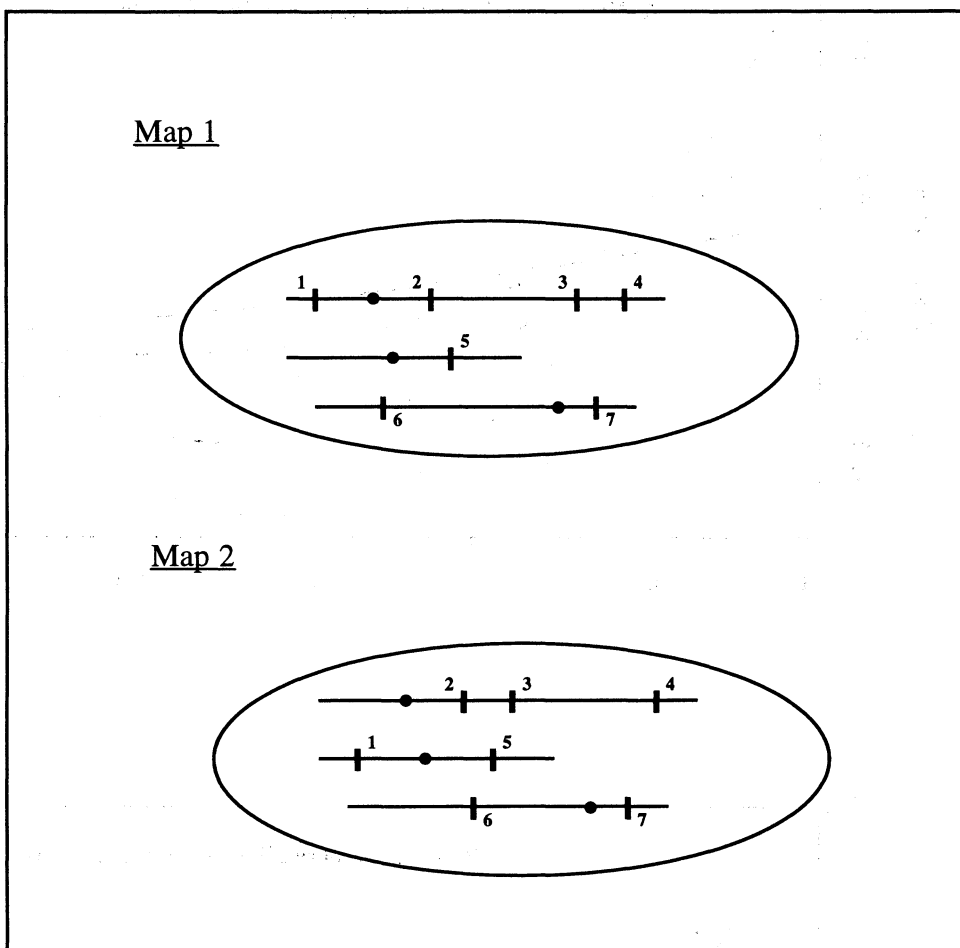
**Figure 2** : Probability of recovering one parent by selfing from (A) an hybrid, (B) a single backcross, (C) a double backcross.



**Figure 3 :** Distribution of the Roger's distance for  $M=100$  loci and  $d_{AB}=0.2$

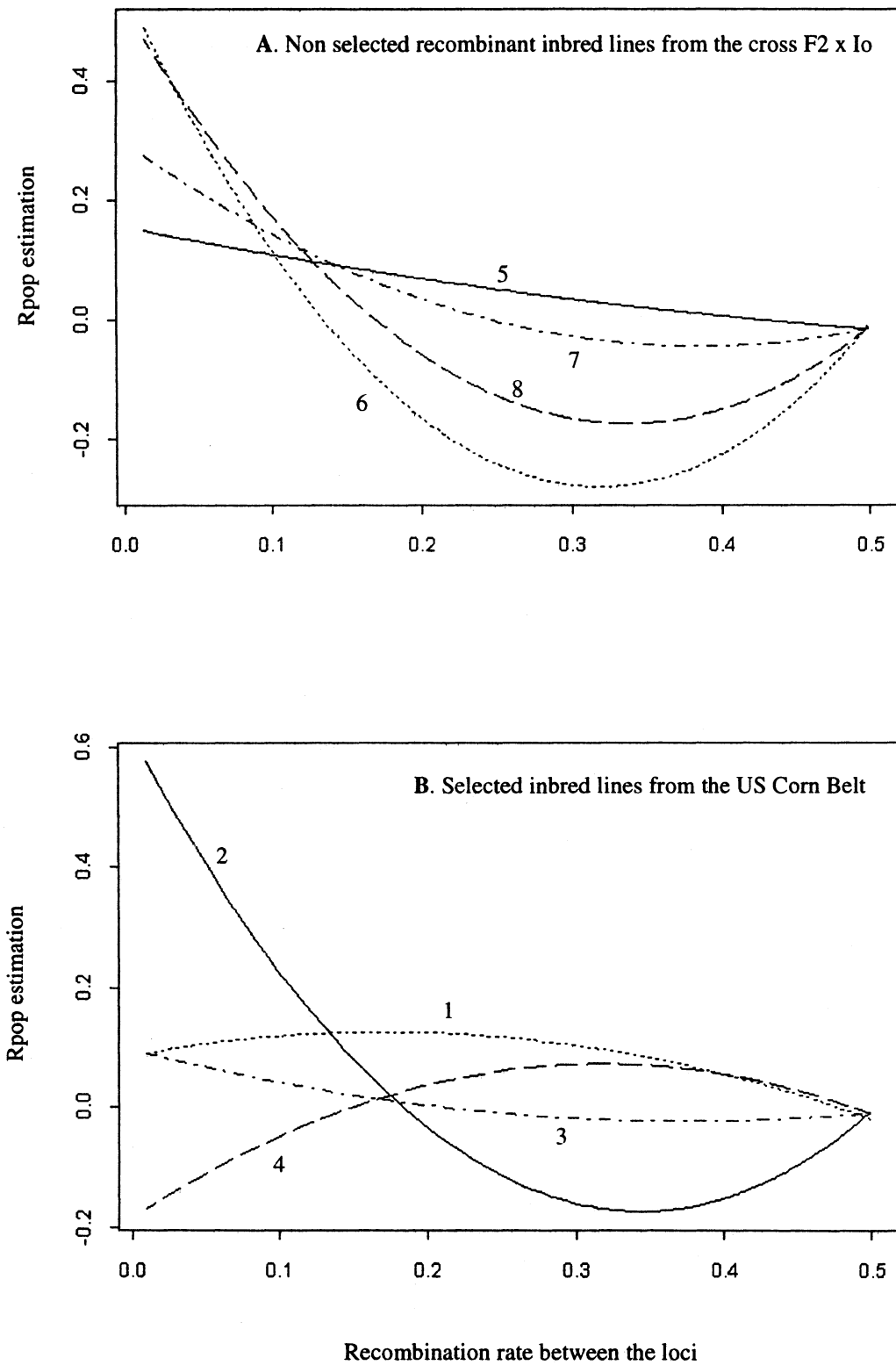


**Figure 4** : Exact power of the test  $d_{AB} < d_{MIN}$  for sample sizes ranging from 80 to 200 and for different values of  $d_{MIN}$  and calculated from the binomial distribution.



**Figure 5 :** Example of 2 equivalent genetic maps preserving the distance matrix.

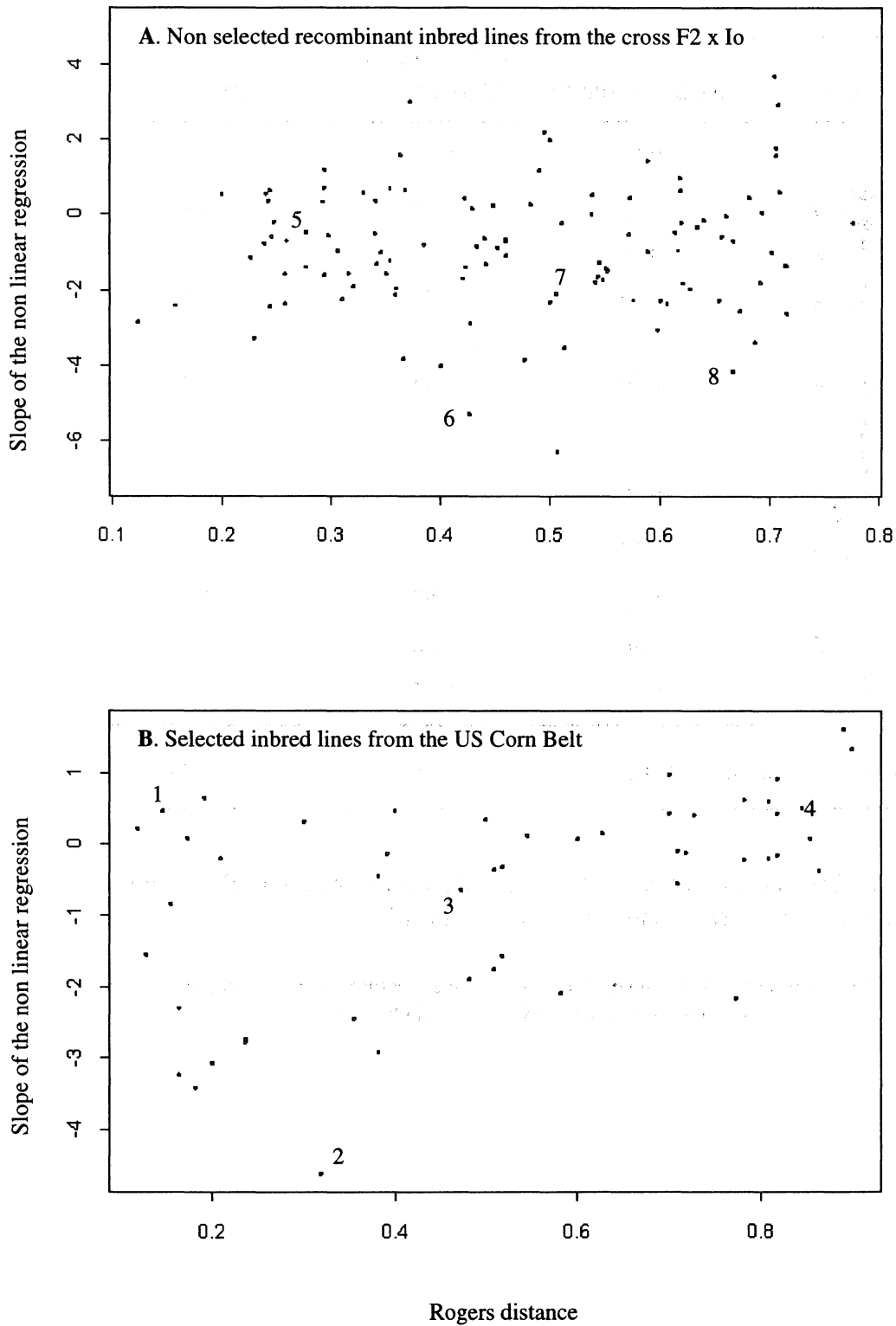




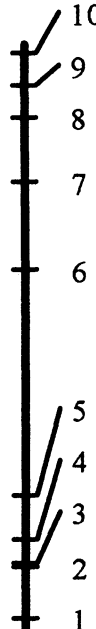
**Figure 6** : Estimation of  $R_{pop}$  from the data for 8 couples of maize inbred lines.

1=(B7,C5) 2=(PA632,C4) 3=(C5,B5) 4=(PB73,PA4)

5-8=four couples of recombinant inbred lines from the cross F2 x Io

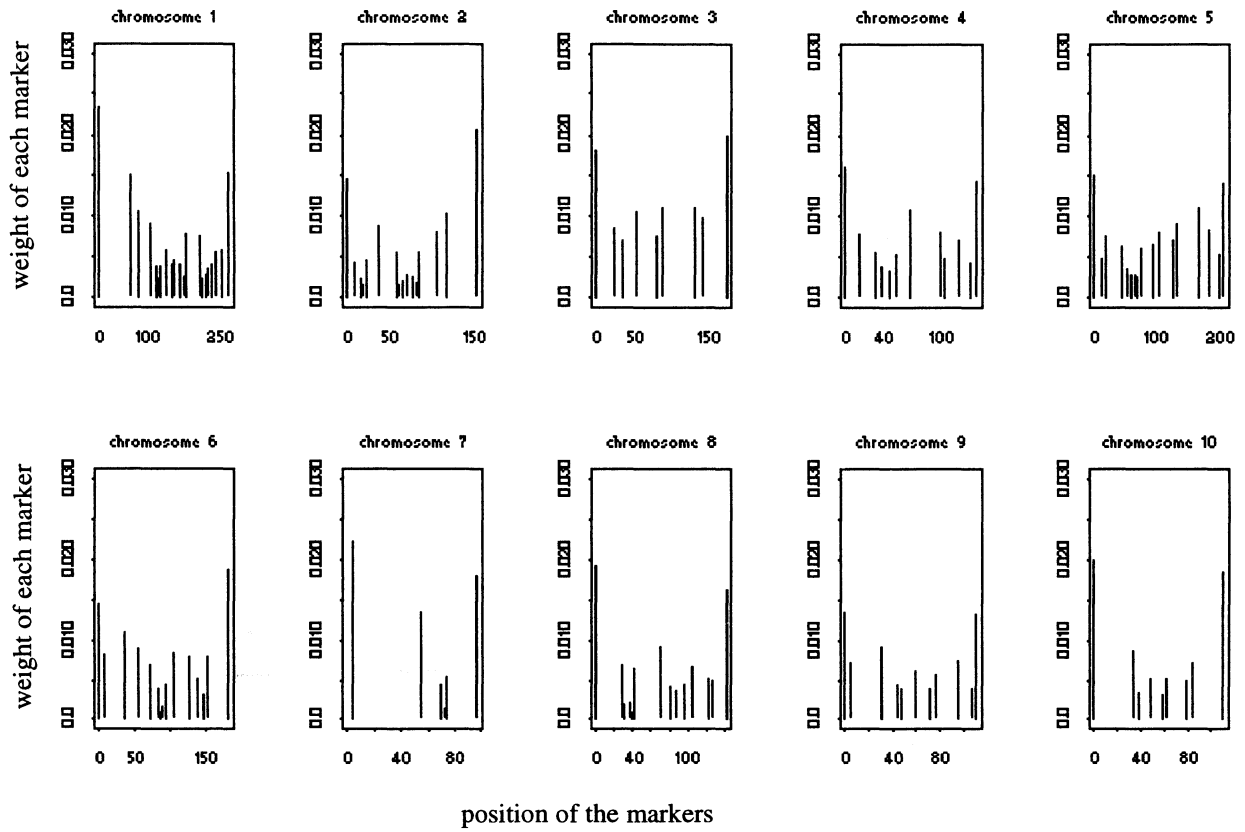


**Figure 7** : Relationship between the slope of  $R_{pop}$  estimated from the data and the Rogers distance for two different data sets. A= recombinant inbred lines and B=highly selected inbred lines from the US Corn Belt.

| Genetic map  | Rogers distance | coefficients of<br>ED distance | DUS distance |
|--|-----------------|--------------------------------|--------------|
|  | 0.10            | <b>0.22</b>                    | 0.09         |
|  | 0.10            | 0.03                           | 0.04         |
|  | 0.10            | 0.01                           | 0.02         |
|  | 0.10            | 0.04                           | 0.06         |
|  | 0.10            | 0.14                           | <b>0.19</b>  |
|  | 0.10            | <b>0.17</b>                    | <b>0.23</b>  |
|  | 0.10            | 0.09                           | 0.12         |
|  | 0.10            | 0.05                           | 0.07         |
|  | 0.10            | 0.04                           | 0.05         |
|  | 0.10            | <b>0.21</b>                    | <b>0.13</b>  |
| Mean sampling variance   | 0.101           | 0.096                          | 0.100        |

**Figure 8** : Comparison of the three methods for the estimation of the genetic distance in a trivial case of one chromosome with 10 loci <sup>(1)</sup>.

<sup>(1)</sup> The covariances between distances at marker loci were computed by supposing the two inbred lines to be compared derived by selfing from a single hybrid.



**Figure 9** : Genetic map of maize for 130 RFLP markers (Causse, 1995) and weight of each marker for the computation of the ED genetic distance, supposing the two inbred lines to be compared derived by haplodiploidization from the same hybrid. Horizontal axis : position of the markers in the chromosome. Vertical axis : value of the weighting coefficient for each marker locus.