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agreed by the Council of UPOV, only documents that have been adopted by the Council of UPOV and that have not been superseded can represent UPOV policies or guidance.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

E

UPOV

BMT/5/4
ORIGINAL: English
DATE: September 9, 1998

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS
GENEVA

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

Fifth Session
Beltsville, United States of America, September 28 to 30, 1998

THE POTENTIAL OF AFLP MARKERS FOR DISTINGUISHING BETWEEN RYEGRASS VARIETIES

Document prepared by experts from Belgium, France, the Netherlands and the United Kingdom

The potential of AFLP markers for distinguishing between ryegrass varieties

P. Dubreuil, F. Van Eeuwijk, C. Baril, Ch. Dillmann, M. De Loose, J. Law, I. Roldan-Ruiz

Abstract : The potential of AFLP markers for distinction studies in perennial diploid ryegrass (*Lolium perenne* L.) was investigated from a set of 11 cultivars assayed for DNA polymorphism using two primer combinations. The discriminatory power provided by AFLPs was analysed and various statistical approaches for testing distinctness were compared. Special attention was paid to (i) redundancy among markers, and (ii) optimal sample sizes of both individuals and markers that are required to minimize the variance of the molecular distances between cultivars. To this end, bootstrap sampling strategies were carried out. Statistical procedures aimed for testing distinctness included (i) estimation of population predictors, (ii) Analysis of Molecular Variance (AMOVA), (iii) stepwise regression procedures, and (iv) partial least squares regression procedures. The results prove that AFLP markers are discriminant enough to distinguish between the closest cultivars although a large redundancy was observed. They also suggest that a relatively small sample of individuals per cultivar (~20-30) may suffice for testing distinctness. The evolution of the sampling variance of the distances between cultivars (when varying number of individuals and markers surveyed) showed that it is better to examine a large number of markers rather than a large number of individuals to improve the accuracy of the distance estimate.

A. Introduction

This document reports on the work of a group of scientists (statisticians, geneticists, and molecular biologists) from Belgium, France, The Netherlands and United Kingdom, who have engaged a discussion on the use of molecular markers for assessing distinctness between crop cultivars. The most important questions that this group proposed to deal with included :

- How do molecular markers perform for assessing distinctness as compared to traditional morphological traits ?

- Are the results obtained with different marker systems distinct, and if so, which marker system should be used ?

- How to deal with the information from molecular markers statistically for DUS testing ?

- Is molecular marker information reliable enough and which are the most significant sources of errors among (i) genetic heterogeneity, (ii) differences between DNA extractions from the same plant, (iii) differences between fingerprints from the same DNA extraction.

In this context, the use of AFLP markers in ryegrass is likely to represent one of the most complex examples of using molecular markers for distinction. It combines several problems arising from (i) the use of markers without known genetic determinism (dominant and multilocus), (ii) the high amount of somewhat redundant information provided per primer combination, and (iii) the variability among samples of individuals within heterogeneous populations.

Ryegrass cultivars are commercialized as synthetic populations obtained from the polycross between a variable number of parents (usually between 5 and 15). DUS testing is currently carried out on the second (Syn2) or the third (Syn3) generation of multiplication. Up to now, it involves the measurement of morphological traits and the evaluation of enzymatic systems. One significant difference for any of the traits considered is sufficient for distinction. Uniformity is evaluated for the same traits by testing for heterogeneity among variances of different samples from the same variety. Stability is assumed to be highly correlated to uniformity and is not further investigated.

It is noteworthy that 95% of distinction decisions are based on earlyness (*ie.* date of ear emergence) which underlines the small genetic basis of commercial cultivars. Molecular markers can therefore be very useful to (i) help in DUS establisment, and (ii) distinguish between cultivars that morphological traits failed to distinguish, supposing that plant breeders would be interested in protecting varieties showing the same morphological characteristics from different genetic backgrounds.

B. Plant material

To date only registered European cultivars of perennial ryegrass (*Lolium perenne* L.) have been included in this study (Table 1). They were obtained from four European breeding companies or institutes. Each company was asked to select a representative set of the material from their own breeding programs. By the time the selection of the cultivars was made, the

objective of the study was only to test for the discriminatory power of AFLP markers among released cultivars, and not among cultivars subjected to DUS testing. Each cultivar was represented by a sample of 43 to 54 individual plants.

C. AFLP assays

Individual plants were separately assayed for AFLPs by using the primer combinations E*co*RI-ACG / M*se*I-CTT (primer combination M), and E*co*RI-AGG / M*se*I-CTT (primer combination N). Details on the AFLP protocol used (Perkin Elmer) are reported by Roldán-Ruiz *et al.* (1997). The E*co*RI primer was labelled with a fluoresceine group « JOE ». The samples were loaded on a 5% polyacrylamide gel and analysed with an ABI Prism 377 DNA sequencer. The computer program GeneScan 2.0.2. was used to analyse the data and for the generation of the sample files.
The sample files were further analysed using Genotyper 2.0. Only polymorphic bands which showed a relatively good amplification in at least one of the plants analysed and that were easy to identify, if present, in the rest of the plants were selected. Each marker was coded by 1 or 0 whether present or absent in an individual plant to form a binary matrix of size NxP where N and P are the total number of individual plants and the number of markers, respectively. Because ryegrass cultivars are genetically heterogeneous, typically not all plants in a cultivar will share the same bands or lack other ones.

D. Morphological evaluation

Morphological data were obtained for all cultivars from PTS as averages over 3 to 10 years of evaluation depending on the cultivar considered. Morphological traits included angle in year of sowing (AYS), spring height (SH), date of ear emergence (DEE), height of plant at ear emergence (HEE), width of plant at ear emergence (WEE), length of flag leaf (LFL), width of flag leaf (WFL), length of the longest stem at DEE + 30 days (LLS), length of ear (LE), and spikelet number (SN).

E. Results

A total of 133 polymorphic bands were selected over the entire set of individuals (532) from the two primer combinations assayed. For the first primer combination (*ie.* E*co*RI-ACG/M*se*I-CTT) 59 polymorphic bands were scored ranging from 77 to 418 bp, whereas for the second one (*ie.* E*co*RI-ACG/M*se*I-CTT) 74 polymorphic bands were scored between 87 and 486 bp.

1. Preliminary screening of the data

Association among individuals as revealed by principal components analysis (PCA) and examination of individual scores for the uniqueness measure (Messmer *et al.*, 1991) pointed out 15 outliers from the cultivars Mongita (13), Merbo (1), and Barpolo (1). Further control of the AFLP patterns confirmed problems during AFLP reactions for these individuals which were subsequently removed from the original dataset. Therefore, most of the statistical analyses were performed using a dataset reduced to 517 individuals and 128 markers.

Associations among cultivars as revealed by AFLP markers and morphological traits were graphically depicted through PCA. The location of the cultivars was defined by the two first principal components, which explained together 71.3 % of the total variation at the phenotypic level (Figure 1.1), and 38.2 % of the total variation at the molecular level (Figure 1.2). Comparison between both PCAs did not show concordant groupings of cultivars. PCA based on marker data did not reveal close associations among cultivars while PCA based on phenotypic data exhibited a clear separation between the cultivar Barylou and others. This cultivar was characterized by low values for the date of ear emergence (DEE), the height of the plant at ear emergence (HEE), the length of the ear (LE), the number of spikelet (SN), the length of the longest stem (LLS), and the angle in year of sowing (AYS), which all were highly positively correlated with the first axis.

A discriminant analysis was performed on the AFLP data for all the cultivars to see whether differences between cultivars could be easily found. It turns out that after four axes explaining together 60.2 % of the total variation, allocation is very acceptable already (*ie.* between 80 and 100% of the plants from a cultivar were assigned to the right cultivar) and cannot be improved much by including more axes (data not shown). This result already indicates that it will be quite easy to find distinctness and also that it must be possible to find subsets of markers that will do just as well as the full set of markers.

2.Identification of discriminative markers and evaluation of redundancy

The polymorphism information content (PIC) is usually seen as a convenient parameter to identify the markers with high discriminatory power. Assuming that each marker corresponded to a single biallelic locus, the PIC was computed as $PIC = 2f(1-f)$, where f is the frequency of the band (*ie.* the amplified allele), and $(1-f)$ is the frequency of the null allele. PIC values ranged from 0.004 to 0.499, thus spaning almost the whole possible range for this parameter when markers are biallelic. As it is shown figure 2, the distribution of the PIC among the entire set of markers was almost uniform, and no clear discrepancy between both primer combination was shown.

To estimate the amount of redundant information among AFLP markers, we developed a strategy that consists in ordering the markers according to their effect on the distance between two cultivars as estimated by the ϕ_{st} through AMOVA (Excoffier *et al.*, 1992; Dillmann, 1996; further in this text). The procedure starts by removing each marker in turn to identify the marker that minimizes the square deviation (SD) between the distance computed from the full n-set of markers and the distance computed from the remaining (n-1)-set of markers. When the first least informative marker is identified, the procedure is rerun with all the markers except the one removed in the previous step. Then, the second least informative marker is identified and the process is repeated until all the markers are ordered from the least to the most informative ones. As an example, the figure 3.1 shows the relationship between the SD and the number of markers removed for the comparison between the closest cultivars (*ie.* Herbie and Merganda). The shape of the curve suggests a large redundancy among AFLP markers since number of them were removed before the distance between the cultivars considered became significantly different from the original distance. As expected, all the markers with low PIC values were removed at first as the least informative, whereas those which were removed at last as the most informative had always high PIC values (Figure 3.2).

Nonetheless, the subset of markers with high PIC values also included low informative markers. This confirms that the PIC is not solely sufficient to identify the subset of the most informative markers. It only estimates the potential discriminatory power of an individual marker and not the actual discriminatory power that also depends on the correlations with the other markers in the dataset.

Another way to identify the most discriminative set of markers is by using a subset selection procedure in a regression with the cultivar membership indicator (1 for the cultivar *i* and 0 for the cultivar *j*) as the dependent variable, and the markers as the independent variables. This procedure was carried out for all pairwise comparisons between cultivars (*ie.* (11*10)/2=55 in total), and the number of times a marker was selected by stepwise regression as a part of the most discriminative set was counted. This accounts for the ability of a marker to distinguish between two cultivars among all possible pairwise comparisons. This appoach has the advantage of selecting the subset of the most discriminant markers among all pairwise comparisons. The results obtained agree with those obtained by the SD-AMOVA procedure (above). The relationship between the PIC and the number of times a marker was selected showed a triangular shape (Figure 4). Markers often selected always have high PIC values while those which are rarely (or never) selected can have either low or high PIC values. Selecting *a priori* the markers with the highest PIC values therefore ensure us that the most informative ones for purposes of distinction will be included.

3.Statistical approaches to the problem of distinction from AFLP markers

We present a number of statistical approaches that all succeed in distinguishing between the cultivars for all pairwise comparisons. Some of them (the population predictor method, and the AMOVA) work with the full set of markers, whereas others work with the subset of the most discriminative markers (multiple regression). We have also tested the partial least squared regression method (PLS regression) which can be considered as intermediate between the previous methods since it works with all markers but gives more weight to the most informative ones. Not all proposed methods have been tested with the same detail. Some methods are just proposed to indicate the direction of thought that followed from discussions. This is specially true for the first method to be described here.

3.1. The population predictor method

Ryegrass cultivars are best compared on the basis of marker frequencies. A usefull concept is the so-called population predictor. This predictor contains 1 when the frequency of a marker in a cultivar exceeds 0.5 and 0 otherwise. When the cultivars are highly differentiated, most of the positions in the fingerprints of individual plants from a unique cultivars will be correctly predicted.
As an example (Table 2), the population predictor for the cultivar 1 (*ie.* predictor 1) predicts correctly 5 marker positions out of 5 for the plant 4 in this cultivar. Summed over all marker positions for cultivar 1, the predictor 1 predicts 18 out of 25 marker positions correctly for presence or absence. In the same way, the predictor for cultivar 2 (*ie.* predictor 2) predicts 21 out of 30 marker positions for this cultivar. Because there is not a complete separation between both cultivars, the predictor 1 also predicts 12 out of 30 marker positions correctly for the cultivar 2, whereas the predictor 2 predicts 9 out of 25 marker positions correctly for the cultivar 1.

One way to quantify the success of the markers in distinguishing between the two cultivars is by counting the sum of the number of correct predictions by the population predictor for the plants in the proper class *ie.* W=18+31, versus the average number of correct predictions by the predictor from the other cultivar *ie.* B=9+12. A criterion for distinctness between the two cultivars is then (W–B). This criterion is that used in the maximal predictive classification (Gower, 1975), where it is optimized in search for an optimal number of groups.
In our case, we have only two groups and we want to test for distinctness between these groups. To answer the question, the statistic (W-B) is computed from a number of datasets obtained from the original data set by permuting the cultivar membership vector (*ie.* the cultivar indicator variable). This procedure comes down to allocate randomly haplotypes to both cultivars and calculate the statistic from the data sets so obtained. We order the (W-B) scores obtained from permutations with the (W-B) score from the original data set. Assuming that 99 permutations were carried out, this makes a series of 100 in total. If the original (W-B) score is among the 5% of the largest (W-B) scores then the cultivars can be considered as distinct at the 0.05 level of significance.

3.2. The analysis of molecular variance (AMOVA)

The AMOVA (Excoffier *et al.*, 1992) for DUS testing (Dillmann, 1996) consists in partioning the variation among distances between haplotypes from two cultivars into variation among haplotypes within cultivars and variation among cultivars. The variance components are estimated using a standard analysis of variance from which the differentiation between cultivars can be assessed as $\phi_{st} = \sigma_B^2 /(\sigma_B^2 + \sigma_W^2)$, where σ_W^2 and σ_B^2 are the variances among haplotypes within cultivars and among cultivars, respectively. When the distance between two haplotypes is defined by the sum of squared differences between band frequencies (1 if the band is present, 0 otherwise) over the markers, then it is equivalent in performing an analysis of variance per marker from the two-way table of plants by markers with the two cultivars as grouping factors. In this case, the ϕ_{st} statistic corresponds to the best linear prediction of the F-value from a set of independent markers. The main interests of the AMOVA are that (i) either distance matrix can be used, and (ii) a permutation procedure can be applied to obtain the null distribution and to test for the hypothesis of no significant difference between cultivars.

All pairwise comparisons of cultivars were performed using AMOVA from the euclidean distance matrix computed on the full set of markers. The ϕ_{st} values ranged from 0.088 to 0.263, and were all highly significant based on the permutation test (500 permutations) (Table 3). Using only markers from one primer combination did not change the conclusion that all pairs of cultivars were significantly differentiated (data not shown). This result showed that markers from only one primer combination may be discriminant enough for proving distinctness between commercialized cultivars.

3.3. The multiple regression

The method is based on the multiple regression of the cultivar indicator variable (as those used in the population predictor method) on the set of markers. The statistic is formed by counting the number of well predicted population memberships following this rule : When the predicted value on the regression is above 0.5 for the cultivar 1 or below 0.5 for the cultivar 2, then it is counted as a right prediction, which is reasonable when the cultivars have roughly

equal sampling size. Subsequently, a permutation test can be done, completely analogous to that used for the previous methods.
An important complication in using this method for testing distinctness is that markers may provide redundant information. In standard regression, this is called the multi-collinearity problem. To stay away of this problem, the regression can be performed on a subset of markers selected by either available methods (*ie.* forward selection, backward selection, stepwise selection). A recommended procedure consists in (i) using a subset selection procedure in a regression with the cultivar membership indicator as the dependent variable and the markers as the independent variables, (ii) predicting the cultivar membership with the selected set of markers and calculating the number of right classified haplotypes, (iii) permuting the original cultivar membership indicator variable, and performing again a subset selection procedure followed by prediction and calculation of the number of right classified haplotypes, (iv) repeating (iii) a number of times, and (v) assessing the position of the original dataset statistic and estimating at which level of significance the cultivars may be deemed as distinct.

Before starting the regression procedure, a pre-selection of the markers was made by removing all markers that had a frequency above 75% or below 25% over the two cultivars that were to be compared. Afterwards, all pairwise comparisons between cultivars were subjected to multiple regression analysis using the stepwise selection method. The number of haplotypes misallocated is given for each comparison in Table 4.1. As we can see, only rarely were haplotypes not classified into the cultivar from which they were sampled. Reducing the cultivars to half and one third of the plants sampled did not change greatly the results (Table 4.2). Hence, a relatively small sample of plants per ryegrass cultivar may suffice for finding distinctness.

3.4. The PLS regression

The purpose of this approach is to use all markers without suffering from the problem of multi-collinearity. Basically, a predictor for cultivar membership is formed from all markers, where each marker is weighted with its correlation with the indicator variable for cultivar membership. The latent variable in PLS is constructed to describe as good as possible the information in the independent variables (*ie.* the markers). The first PLS latent variable is thus the linear combination of markers that correlates as high as possible with the indicator variable for the two cultivars while describing the maximum amount of variation in the markers as well. The coefficients for the markers are (more or less) the correlations between each marker and the indicator variable for cultivar membership.
A series of such compound predictors (latent variables) can be computed by repeating the procedure a number of times on the residuals of both the indicator variable for cultivar membership and the marker information. The second latent variable is constructed by that part of the marker variability that was not caught in the first latent variable. This is done by regressing all markers on the first latent variable and what is left forms the building blocks for the second latent variable. The indicator variable for the cultivar membership is similarly regressed on the first latent variable. The second latent variable is then constructed by finding weights for the (corrected) markers that make the latent variable correlate maximally with the corrected indicator variable for cultivar membership. Again, the weights for the (corrected) markers are just correlations with the corrected indicator variable.

Practical work on PLS regression used only two predictors for all pairwise comparisons of cultivars. Just like for the stepwise regression, the procedure was performed (i) using all plants sampled within cultivars, (ii) using half of the plants per cultivars, and (iii) using one third of the plants per cultivars. Results were similar to those obtained with the stepwise regression. Very few misclassification occurred when using all the plants (data not shown). However, when only one third of the plants per cultivar were used, the procedure started to break down and many misclassifications were noted.

4.Evaluation of the optimal sample sizes of individuals and markers

The sampling variance of distances between cultivars depends on the number of markers used for computing the distance. Because ryegrass cultivars are genetically heterogeneous, it also depends on the number of individuals plants surveyed per cultivar. In order to evaluate whether it is better to survey more markers rather than more individuals to improve the accuracy of distance estimates, we analysed the relationship between the sampling variance of the Nei's distance (1972) between cultivars and the number of markers sampled. An estimate of the mean sampling variance for Dnei over all pairwise distances between cultivars was derived from 500 bootstrap samples (random sampling with replacement from the initial dataset) of both markers and individuals for each pairwise comparison. This was done for sample sizes of markers varying between 10 and 125 and for sample sizes of individuals between 10 and 35 per cultivar.
The evolution of the mean sampling variance for Dnei (MSVD) when the number of markers increases is shown figure 5 for different number of individuals sampled. The sampling variance rapidly decreased up to 50-60 markers, then plateaued when including more markers. As expected, the MSVD decreased as the number of individuals surveyed increased, but the gain in accuracy was low between 20 and 35 individuals per cultivar. Considering these results which only hold for the cultivars used in this study, it seems that it is better to analyse more markers rather than more individuals to minimize the sampling variance for the distance estimation. Moreover, the optimal sample sizes may be around 60 for the markers (that is approximately the number obtained from a single primer combination) and 20 for the individuals.

F.Discussion - Conclusion

This study was devoted to (i) investigate the discriminatory power of AFLP markers among genetically heterogenous ryegrass cultivars and (ii) evaluate appropriate statistical techniques for DUS testing.
Although a large redundancy was revealed among markers, cultivars were found significantly differentiated among all pairwise comparisons. Morever, it also appeared that markers from only one primer combination and less than 50 individuals per cultivars would be sufficient to prove distinctness. This result can be explained by the fact that cultivars were chosen to represent a broad genetic basis and are probably unrelated. Nevertheless, it also questions the sensitivity of the permutation test that has been used. By permuting only complete haplotypes between the two cultivars to be compared, the test assumed that markers were completely linked which is obviously not true. The sensitivity of the test would be decreased by permuting both individuals and markers, but this would come to assume that markers are

completely independent which is also not true. The test procedure has therefore to be refined to take into account that markers are neither completely linked nor independent.
In the continuation of this project attention will also be given to the choice of related material and different generations of the same cultivar. Such a material may allow us to investigate the power of the permutation procedure used for testing distinctness and to gain an insight into the minimum distance for DUS establisment in ryegrass.

Glossary

AFLP (Amplification fragment length polymorphism) : It is a technique for DNA fingerprinting based on the selective PCR (polymerase chain reaction) amplification of restriction fragments from genomic DNA. It involves (i) restriction of the DNA and ligation of oligonucleotide adapters (about 20 bp long), (ii) selective amplification of sets of restriction fragments, and (iii) gel analysis of the amplified fragments.

Bootstrap procedure : It is a resampling technique used for inferring the variability of a given statistic when the actual distribution of this statistic is unknown. The bootstrap procedure consists in resampling many times (say 1000) with replacement the initial n-set of data, each time producing a fictional set of n data from which an estimate of the statistic is computed. The essential idea is that the set of estimates so obtained has a distribution that approximates the distribution of the actual estimate. The variance of the unknown distribution can be inferred by computing the variance of the set of estimates and the confidence limits of the statistic can be approximated from the observed distribution of this set.

Permutation test : This test posits the null hypothesis that there is no genetic differentiation between the cultivars that are compared. Under this hypothesis, samples of individuals from the two cultivars are considered as drawn from a same cultivar, with variation between samples due to random sampling alone. The null hypothesis of no genetic differentiation is obtained by allocating at random each individual to either cultivar. For AMOVA, this comes to random permutation of the rows and corresponding columns of the distance matrix between individuals. The differentiation between cultivars is then estimated for a number (500 in this study) of permuted matrices to obtain the null distribution. If only 5% of the estimates of differentiation computed from the permuted matrices are greater than the original estimate, then the null hypothesis can be rejected at the 95% significance level.

References

C. Dillmann (1997) The use of the Analysis of molecular Variance for distinction studies. Document BMT/4/9, Fourth session of the working group on biochemical and molecular techniques from UPOV, Cambridge, March 1997

L. Excoffier , P.E. Smouse, and J.M. Quattro (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes : application to human mitochondrial DNA restriction data. Genetics 131 : 479-491.

M.M. Messmer, A.E. Melchinger, M. Lee, W.L. Woodman, E.A. Lee, and K.R. Lamkey (1991) Genetic diversity among progenitors and elite lines from the Iowa Stiff Stalk Synthetic (BSSS) maize population : comparison of allozyme and RFLP data. Theor. Appl. Genet. 83 :97-107.

M. Nei (1972) Genetic distance between populations. The American Naturalist 106 (949) : 283-292.

I. Roldán-Ruiz, K. Van Laecke, J. De Riek, J. Dendauw, A. Depicker, E. Van Bockstaele, and M. De Loose (1997) The use of AFLP markers for the identification of ryegrass (*Lolium ssp.*) cultivars. Advances in Biometrical Genetics. Proceedings of the tenth meeting of the EUCARPIA section biometrics in plant breeding. Poznán, 14-16 May 1997, P. Krajewski and Z. Kaczmarek (eds.), pp 231-237.

Table 1. Main characteristics of ryegrass cultivars analysed

#	Cultivar	Breeding company	Genetic origin	Ploidy	Type	Earliness
1	Barlet	Barenbrug (Holland)	Hungary	2n=14	Syn. – 4 parents	Medium
2	Barpolo	Barenbrug (Holland)	Denmark	2n=14	Mass selection	Very late
3	Barylou	Barenbrug (Holland)	Traditional ecotypes	2n=14	Syn. – 4 parents	Very early
4	DP8611[a]	DLF (Denmark)	-	2n=14	-	-
5	Herbie	Van der Have[b]	North-Western Europe	2n=14	Syn. – 4 parents	Late
6	Merbo	RvP (Belgium)	Selected belgian ecotypes	2n=14	Syn.	Medium
7	Merganda	RvP (Belgium)	Selected belgian ecotypes	2n=14	Syn.	Medium
8	Mikado	DLF (Denmark)	-	2n=14	Syn.	Medium
9	Mongita	Mommersteeg[b]	-	2n=14	Syn. – 6 parents	Medium
10	Morimba	Mommersteeg[b]	Cultivar Morenne	2n=14	Syn. – 6 parents	Medium
11	Paddok	RvP (Belgium)	-	2n=14	Syn. – 6 parents	Late

[a] cultivar called Hamlet in the market, [b] Van der Have and Mommersteg now belong to the same breeding company

Table 2 Example of population predictor method

Cultivar	**Haplotype**	**Indicator Variable**	*Marker*					Well predicted by	
			1	**2**	**3**	**4**	**5**	**PP1**	PP2
1	Plant 1	1	1	0	0	0	1	3/5	0/5
1	Plant 2	1	1	1	0	1	1	5/5	2/5
1	Plant 3	1	0	1	0	1	1	4/5	3/5
1	Plant 4	1	1	1	0	1	1	5/5	2/5
1	Plant 5	1	1	1	1	0	1	3/5	2/5
Marker frequency in cultivar 1			0.80	0.80	0.20	0.60	1.0		
Population Predictor 1 (PP1)			**1**	**1**	**0**	**1**	**1**	**18/25**	**9/25**
2	Plant 1	0	0	0	1	1	0	1/5	4/5
2	Plant 2	0	0	1	1	1	0	2/5	5/5
2	Plant 3	0	1	1	1	0	1	3/5	2/5
2	Plant 4	0	1	0	1	1	0	2/5	3/5
2	Plant 5	0	0	1	1	0	0	1/5	4/5
2	Plant 6	0	0	1	0	1	1	4/5	3/5
Marker frequency in cultivar 2			0.33	0.67	0.83	0.67	0.33		
Population Predictor 2 (PP2)			**0**	**1**	**1**	**1**	**0**	**12/30**	**21/30**

Table 3 ϕ_{st} estimates between pairs of cultivars (below diagonal) and associated levels of significance as estimated from 500 permutations of individuals (above diagonal).

#	**Cultivars**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1	Barlet	-	***	***	***	***	***	***	***	***	***	***
2	Barpolo	0.196	-	***	***	***	***	***	***	***	***	***
3	Barylou	0.210	0.247	-	***	***	***	***	***	***	***	***
4	DP8611	0.169	0.144	0.147	-	***	***	***	***	***	***	***
5	Herbie	0.151	0.162	0.121	0.094	-	***	***	***	***	***	***
6	Merbo	0.250	0.204	0.250	0.215	0.176	-	***	***	***	***	***
7	Merganda	0.194	0.218	0.107	0.128	**0.088**	0.179	-	***	***	***	***
8	Mikado	0.162	0.125	0.235	0.106	0.119	0.188	0.201	-	***	***	***
9	Mongita	0.207	0.195	0.225	0.141	0.149	0.234	0.224	0.145	-	***	***
10	Morimba	0.201	0.221	0.162	0.109	0.115	0.263	0.130	0.161	0.181	-	***
11	Paddok	0.197	0.219	0.150	0.122	0.097	0.216	0.096	0.155	0.190	0.116	-

***, P<0.001

Table 4.1 Number of plants wrongly allocated (out of ~ 100) after stepwise regression on the full marker dataset

#	Cultivars	1	2	3	4	5	6	7	8	9	10	11
1	Barlet	-										
2	Barpolo	1	-									
3	Barylou	0	2	-								
4	DP8611	0	4	1	-							
5	Herbie	1	2	4	13	-						
6	Merbo	2	0	0	0	1	-					
7	Merganda	0	5	2	11	13	2	-				
8	Mikado	2	3	1	3	8	3	1	-			
9	Mongita	0	2	1	3	3	1	0	2	-		
10	Morimba	0	1	1	3	6	3	7	17	0	-	
11	Paddok	0	1	1	5	7	3	6	7	0	4	-

Table 4.2 Number of plants wrongly allocated after stepwise regression on the full marker dataset when only half of the plants per cultivar (~ 25) are used (range over 3 random samples) (above diagonal), and when only one third of plants per cultivar (~ 15) are used (one random sample) (below diagonal)

#	Cultivars	1	2	3	4	5	6	7	8	9	10	11
1	Barlet	-	0-3	0-2	0-1	0-2	0-1	0-3	1-3	0-0	0-3	0-1
2	Barpolo	0	-	0-2	1-3	1-7	0-0	0-3	2-4	0-3	0-3	0-0
3	Barylou	0	0	-	0-1	2-4	0-0	0-5	0-0	0-0	0-1	0-1
4	DP8611	2	5	0	-	2-3	0-0	1-6	5-7	0-2	0-1	0-4
5	Herbie	0	0	0	5	-	1-1	5-6	0-1	0-0	0-2	1-1
6	Merbo	1	0	0	0	1	-	0-1	0-1	0-1	0-0	0-0
7	Merganda	0	1	0	11	2	0	-	0-1	0-0	0-2	0-7
8	Mikado	0	2	0	2	1	0	0	-	1-4	2-10	1-1
9	Mongita	1	2	0	3	2	3	0	3	-	0-2	0-0
10	Morimba	5	0	1	1	6	0	4	1	2	-	0-3
11	Paddok	0	0	0	0	1	0	3	2	1	0	-

Fig 1.1 PCA on morphological traits (axis1: 48.7%, axis2: 22.6%)

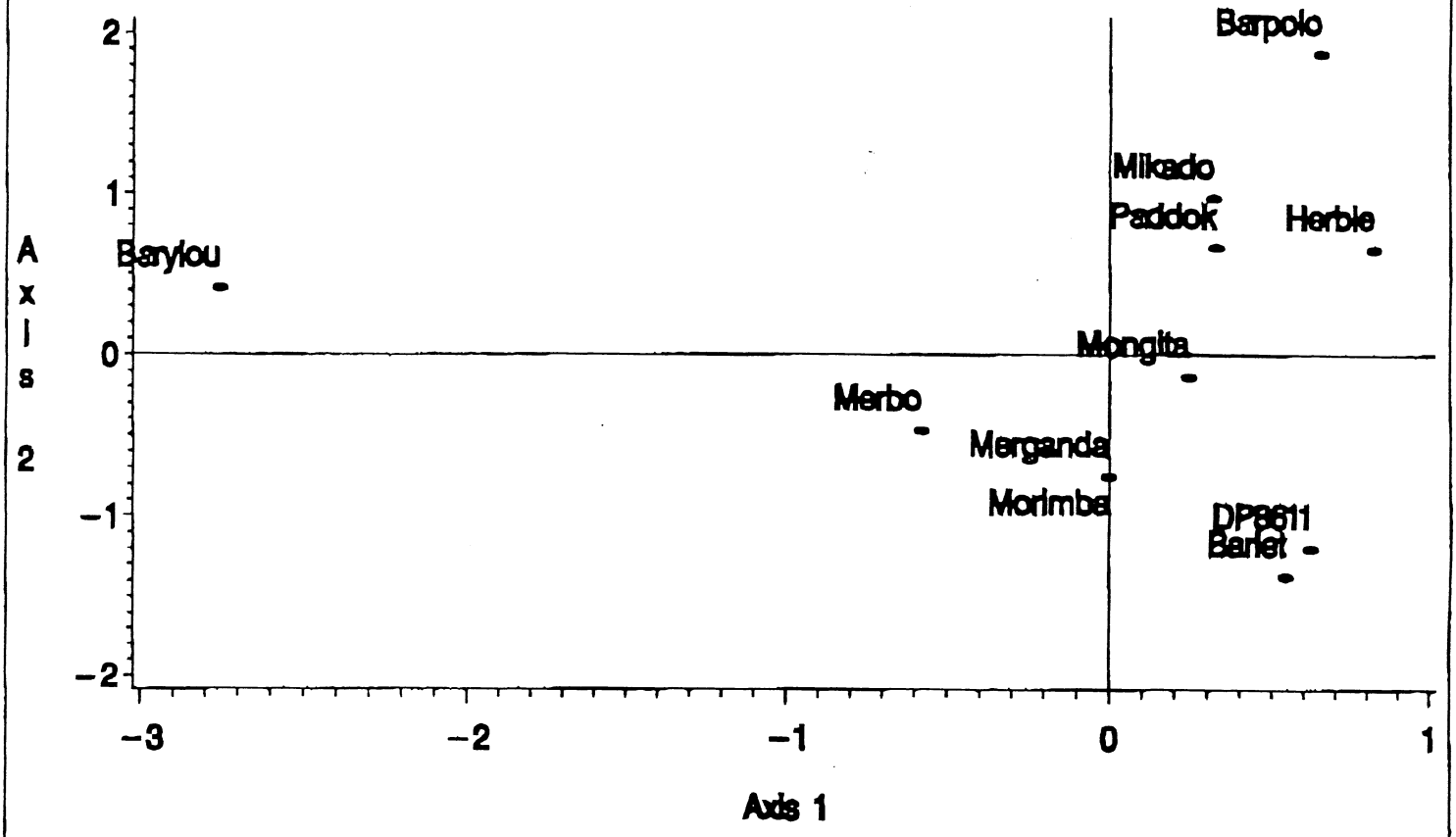


Fig 1.2 PCA on AFLP band frequencies (axis1: 21.5%, axis2: 15.7%)

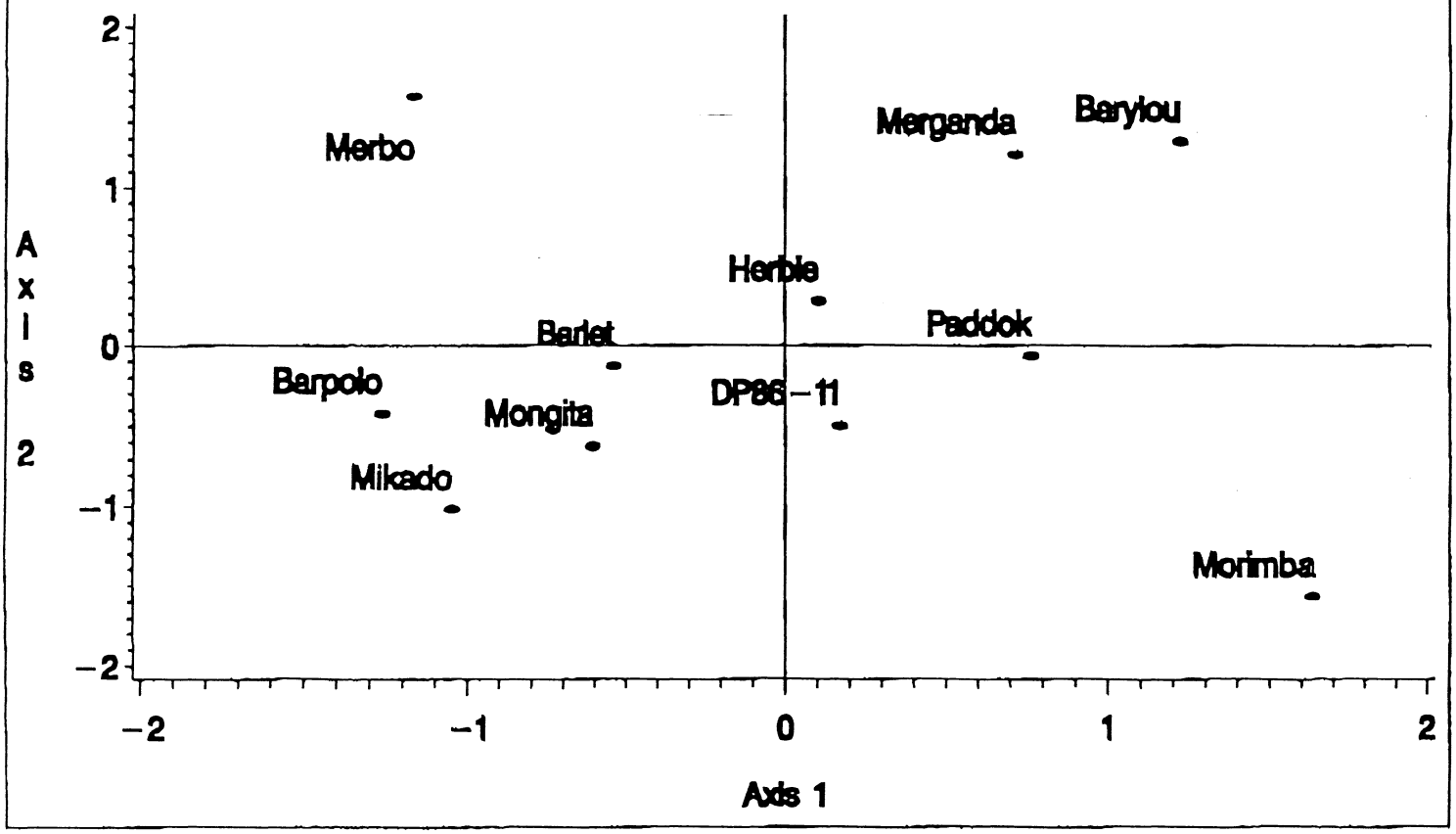


Fig. 2 Distribution of PIC values among AFLP markers

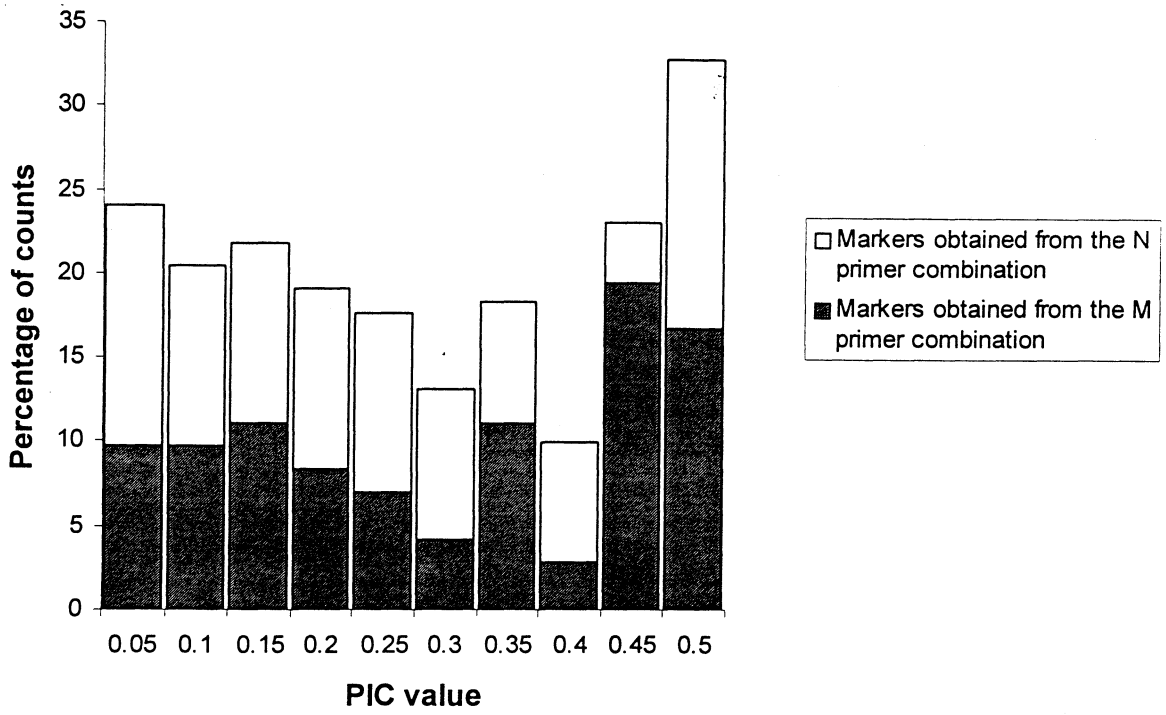


Fig.3.1 Relationship between the distance square deviation (SD) and the number of markers removed for the comparison between the cultivars 5 and 7

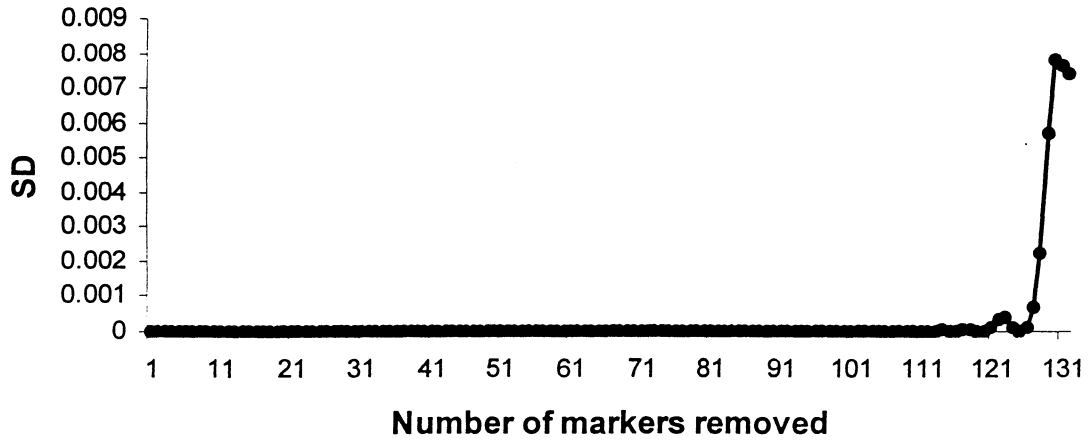


Fig. 3.2 Relationship between the PIC and the order of removal for the comparison between the cultivars 5 and 7

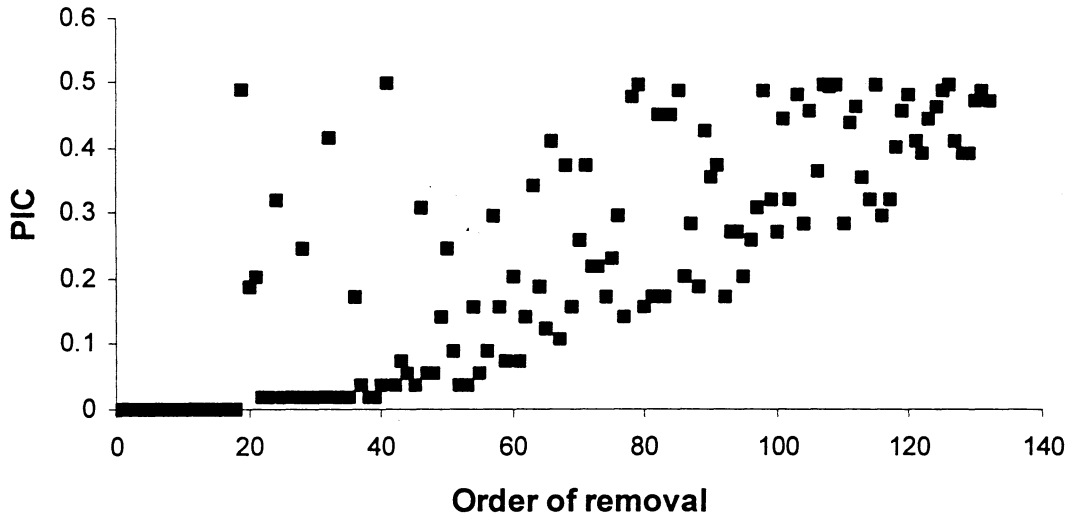


Fig.4 Relationship between the PIC value and the discriminatory power as estimated by stepwise regression

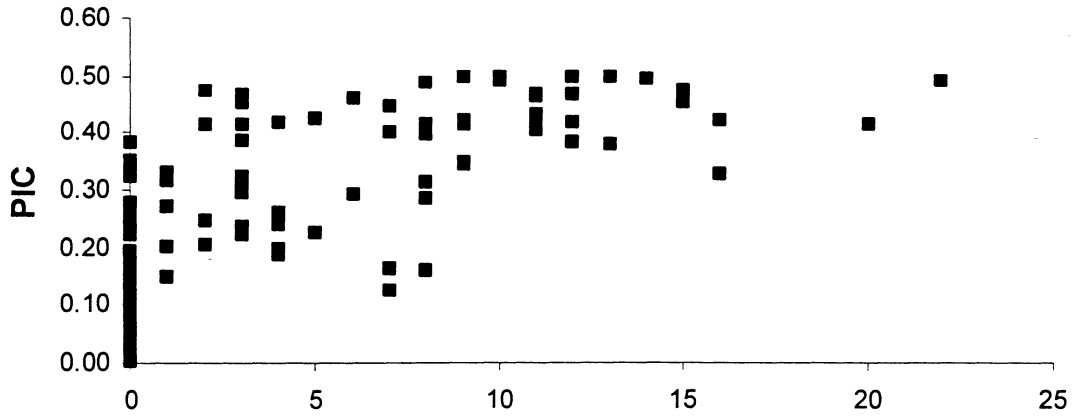
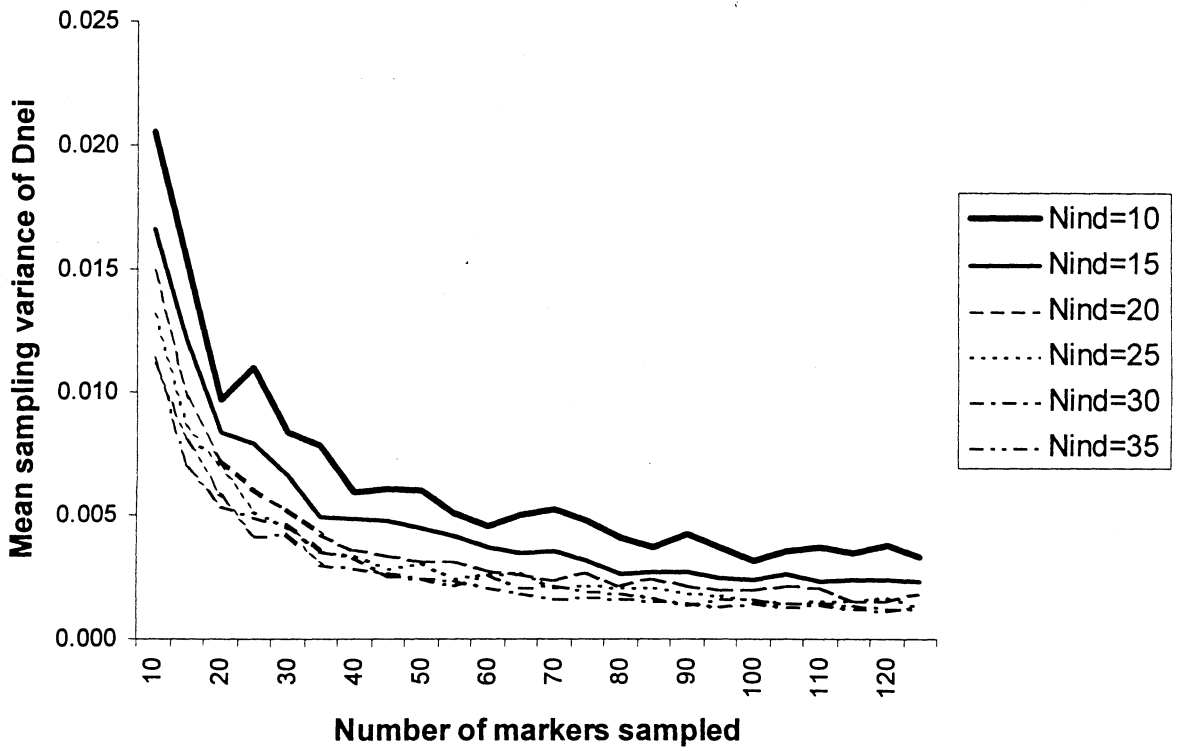


Fig. 5 Relationships between the mean sampling variance of Nei's distance and the number of markers sampled for different number of individuals surveyed



[End of document]