E

BMT/5/3
ORIGINAL: English
DATE: August 12, 1998

UPOV

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

PREDICTION OF VARIETY RELATEDNESS: MOST SIMILAR VARIETY COMPARISONS AS A PRE-SCREENING TOOL

Document prepared by experts from the United Kingdom and the United States of America

PREDICTION OF VARIETY RELATEDNESS :-
Most Similar Variety Comparisons as Pre-screening Tool.

John Law[1], Robert Cook and Stephen Smith[2]
[1]NIAB, Huntingdon Road, Cambridge, CB3 0LE, UK.
[2] Pioneer Hi-Bred International, Johnson, Iowa, 50131, USA.

1. Introduction.

A number of authors have considered the potential application of some molecular marker methods to variety identification and discrimination. In maize, Smith *et al* (1991), Bar-Hen *et al* (1995) and Dillmann *et al* (1997) have focused on comparisons of molecular methods with methods based on the crop morphology and pedigree. Similar studies in barley have been reported by Graner *et al* (1994) and Russell *et al* (1997).

At the request of the 4th (1997) UPOV BMT meeting, the TWC undertook a study on the comparison of the most similar variety based on both morphological data and molecular methods.

Data sets from maize, wheat and barley have been studied in detail together with pedigree information where available. This paper reports findings from the analysis of the maize data.

2. The Data

Data from 35 maize inbred lines have been analysed. Listed below are the types of data used, the number of 'characters' (ie for morphological data the number of characters, for molecular data the number of polymorphic bands) and the scoring method utilised:-

Type	No. Characters	Similarity Scoring System
Morphology	50	Euclidean
AFLP	347	Jaccard
APPCR	258	Jaccard
RFLP	951	Jaccard
SSR	63	City Block
Pedigree	~	As supplied

AFLP Amplified fragment length polymorphism
RFLP Restriction fragment length polymorphism.
SSR Simple sequence repeat
APPCR Arbitrarily primed polymerase chain reaction

3. Methods.

The data presented were first screened to remove a few monomorphic bands from the raw scored data sets. Data provided may be part of a larger set of varieties which could have shown polymorphisms. Such data cause computational difficulties if retained. Morphological characters that contained 'missing' values were also excluded at this stage.

Each type of data requires individual scoring algorithms to create the pair-wise variety similarity matrices. Data from AFLP, RFLP and APPCR analysis generate a binary data set which relates to the presence/absence of clearly observed bands. Several authors (e.g. Law *et al* 1997) have shown that the Jaccard method of constructing a similarity matrix performs satisfactorily for AFLP data and this has been used in this case. Unique banding patterns occur as a result of SSR analysis and these are scored using the City Block approach as there is an element of increased mobility in the bands but not sufficient to warrant a fully ordered treatment. Morphological data have been treated as 'Euclidean' for the purpose of this study. Pedigree similarity coefficients were used as supplied.

For each method the most similar variety has been identified for each individual variety, with the observed level of similarity recorded. Such an approach seeks to quantify the agreement between molecular methods and then to compare with both the pedigree and morphological methods. A secondary approach, less prone to influences due to 'near-misses', identifies the most similar variety by morphological data and then determines the ranked position of that variety in the similarity set as expressed through molecular methods. This method can become cumbersome if the molecular method has multiple tied first rank for the "most similar variety" or if the pedigree matrix contains many identical low similarities.

Correlations and scatter plots also aid interpretation of agreement between methods, based on individual variety assessments.

4. Results

4.1 Plots.
An assessment of the overall level of agreement between the similarity matrices for DNA methods compared with similarities derived from morphological data can be seen in the pair-wise scatter plots (Fig. 1). The scatter plot of the morphology similarities (range 0.6 - 1.0) versus SSRs (range 0.5 to 0.95) shows a symmetrical cloud of data points with little discernible grouping or clustering of points within the overall cloud. However when similar plots versus RFLP and AFLP data are studied it is very noticeable that points are clumped with an over-population of low RFLP/AFLP values and fairly uniform scattering of points in the remainder of the two-dimensional space. The APPCR plot appears to have a form in between those of the SSRs and the RFLP/AFLP. The plot of similarities from morphology against the supplied pedigree similarities shows a more extreme form of the RFLP/AFLP-type situation. This is caused by many pair-wise similarities based on pedigree information being effectively zero.

4.2 Similarity Relationships.
Comparison of the similarity matrices was initially investigated by application of a simple pair-wise correlation. Such correlations when compared to morphology are moderately low at 0.17, 0.16, 0.13, 0.19 and 0.21 for AFLP, APPCR, SSR, RFLP and pedigree respectively. Similar levels of association have been reported by other authors, eg Russell *et al* (1997).
To confirm these findings a MANTEL permutation test was performed based on the R-Package software running on a Mac computer. The method for comparing two similarity matrices works by keeping one matrix fixed and permuting the observations in the second matrix. Test statistics compare the MANTEL statistic with all the other possible statistics generated from the permutations. When numbers of varieties are large an approximation is usually effective based on a sample of over 200 permutations. The normalised MANTEL

statistics confirm that the association between the morphology similarity matrix and the other matrices is weak, although with the large numbers of degrees of freedom both the correlation coefficients and MANTEL statistics achieve a degree of significance.
Of the 50 morphological characters the vast majority are measured and fully justify the use of the Euclidean metric. However it was suggested that the use of the Euclidean metric could be influenced by the presence of a few characters which have numerically large values. Similarity matrices were generated from the original use of the Euclidean metric on all 50 characters; 47 characters were used omitting the 3 large valued characters and the application of the Gower's amalgamation criteria which allows truly measured and scored variables to be combined into a single similarity measure. Comparisons showed differences to be minimal with any conclusions robust in terms of the specific character set tested and approach used.

4.3 Most Similar by Morphology.
The most similar variety to each of the lines of maize taken in turn can be seen in Table 1. For example, in the first four columns of that table, for say target variety number 24, we see that the most similar variety is number 17 with the largest similarity value of 0.922. There are no tied equal maximal similarity coefficients in this analysis but this needs to be allowed for in the more general application of this procedure. It can be noted that the maximal similarity coefficients exceed 0.9 (except for variety 4) in all cases with the majority also greater than 0.95.
Before establishing similar statistics for the DNA methods, it is of interest to note the ranked position of the morphologically most similar variety in the set of similarities from the molecular methods. With perfect agreement across all methods they would each rank the same variety as 1st. This approach allows for slight numerical variation which may affect the similarity ranking, making the agreement appear much worse than it really is. Continuing with the same target variety as before (24) , it can be seen that both AFLP and APPCR rank the selected variety (number 17) second, while SSR and RFLP rank it in first place. Based on pedigree information variety 17 is ranked most similar to the target variety. From Table 1 it can be seen that for varieties 2, 5, 9, 11,12, 16, 19, 24, 26, 29, 31 and 32 each of the DNA methods and pedigree are in good agreement with the most similar variety based on morphological measures. However, for a number of varieties whilst there is good agreement between the DNA methods, collectively the ranking is well away from that established using morphological data (see for example variety numbers 1, 3, 8, 13, 21, 22, 30, 33, 35).

4.4 Most Similar by Molecular Methods.
Results from the application of each DNA method in establishing the most similar variety can be seen in Table 2. Agreements across the methods can be seen (e.g. target variety 4, 6, 9, 10, 16, 17, 20, 25, 27, 31, 33) although the results are totally different from those based on morphology. The absolute levels of maximal similarity for the DNA methods are also of interest. Target variety number 35 has a consistently low level of maximal similarity, showing this to be very different from all other material under test. Excluding this case, the SSRs maximal similarities are the highest and most consistent (range 0.74 - 0.95) while the RFLP values are the most variable with a range between 0.37 and 0.86. Note that for SSRs the maximal similarity values generally exceed 0.8.
Thus far comparisons of the similarities generated by morphological data and DNA methods have been made firstly on the criterion of assessing the rates of exact match for the single most similar variety to each target line and secondly as close matches in a ranked sense. Overall correlations based on pair-wise similarities have already been quoted above and are at best only moderate. However, for each method in turn it is possible to consider the

relationship between rows of the similarity matrices and to form correlations for each target variety.

Consider in Table 3 firstly the summary statistics for comparisons with the morphological results. The maximal relationship is remarkably consistent across all DNA methods and pedigree, at circa 0.66, with minimum correlations all below -0.3, giving a very large range often exceeding 1.0. The median of the varietal correlations for SSRs is noticeably lower than for the other methods. For comparative purposes, results for the pedigree and SSRs are given, with each median relationship, in terms of the correlation coefficients, exceeding 0.77.

4.5 Minimal Set of Most Similar Varieties.

The practical use of 'most similar variety' in pre-screening situations is to select a small set of varieties in which it is highly likely that the 'true' (but unknown) most similar variety occurs. For this exercise target sets of varieties have been sought corresponding to the top 10%, top 20% and the top 30%. The critical cut-off similarity coefficients for morphology sets, corresponding to the most similar coefficients in Table 1, are shown graphically in Fig 2. The graph is ranked in terms of the most similar variety , ie the top variety. It can be seen that, apart from the extremes, the critical similarity values corresponding to the 10%, 20% and 30% criteria are well behaved and follow closely the most similar critical values. This shows that generally there is an even gradient in similarities for a variety. At high values, and to a lesser extent the lower ends, the pattern of distribution of similarities becomes more stretched with the most similar often dominated by a single exceptionally high value.

Similar plots show that the critical cut-off values for 10%, 20% and 30% for SSR are equidistant and parallel to the most similar value. For AFLP's there is a less well behaved system which appears to diverge from the most similar value with, for example, the 30% critical cut-off remaining at about 0.4 while the ranked 'top' variety similarity increases from 0.4 to 0.9.

4.6 Minimal Sets in Practice.

The key remaining question is ... How well do molecular markers perform in terms of producing a limited set of varieties that can be compared with the most similar variety(ies) as defined by using the morphological data only? Variety sets were constructed that, for each molecular method and pedigree, contained the ranked most similar varieties from which the top 10%, 20% and 30% were retained. These sets were then compared, target variety by variety, against the most similar variety based on morphology. This was repeated for the 2nd and 3rd most similar varieties by morphology. As a check on the approach the most similar three varieties were compared for both AFLP and SSR molecular methods.

The results are shown in Table 4, in which it can be seen that, for example with AFLP, the top 10% set (rounded to 4 in this case) contains the most similar variety defined by morphology 46% of the time. This increases to 63% as the set increases to 30% (taken as 10 in this case). There is good agreement between the percentage 'hit rates' with AFLP and SSR and pedigree, although the values are only modest. Of interest is the additional information that at least 20% of the most similar varieties by morphology do not appear in the top 30% of **any** other method (not AFLP nor SSR nor pedigree). For the 3rd most similar variety by morphology the 'total miss rate' is just under 50%. This indicates that morphology is a poor assessor of most similar varieties.

The 'hit rate' is impressive using the most similar variety defined by similarities derived from the AFLP and SSR data. The 'total miss rate' is nil for AFLP and very low for SSR.

5. Conclusions

While the number of varieties/lines available in this project was relatively low, the amount of morphological, molecular and pedigree information utilised was very large. The DNA analysis methods show a measure of internal agreement when compared to the variety selected as the most similar by morphology. However, it should also be noted that for certain target varieties very consistent *but different* conclusions are drawn. The scatter plots show that the range of morphological similarities is relatively low (circa 0.2) compared to those for AFLP and RFLP data at 0.6 and pedigree data at over 0.9.

Overall, the DNA methods appear to give better correlations between each other when identifying a most similar variety, and also correlate better with pedigree data, than does morphology. Morphology thus appears to be a poor assessor of the relationships between varieties and hence of truly similar varieties. The much better molecular methods, used singly or in combination, are able to identify a minimum set of close varieties that are highly likely to contain the truly 'most similar' variety. This is clearly important if molecular methods are to be used in the pre-screening context.

As well as being of interest to the practice of DUS testing *per se*, these results are particularly significant for situations which require knowledge of the relationships between varieties, eg assessments of minimum distance and establishing rational criteria for the definition of essential derivation.

6. Acknowledgements.

This work was supported by the Ministry of Agriculture, Fisheries and Food of the United Kingdom.

References.

Bar-Hen, A., Charcosset, A., Bourgoin, M. and Guiard, J. (1995). Relationship between genetic markers and morphological traits in a maize inbred line collection. Euphytica 84, 145 - 154.

Dillmann, C., Bar-Hen, A., Guerin, D., Charcosset, A. and Murigneux, A. (1997). Comparison of RFLP and morphological distances between maize *Zea mays* L. inbred lines. Consequences for germplasm protection purposes. Theor. Appl. Genet 95, 92 - 102.

Graner, A., Ludwig, W. F. and Melchinger, A. E. (1994). Relationships among European barley germplasm: II. Comparisons of RFLP and pedigree data. Crop Science 34, 1199 - 1205.

Law, J. R., Donini, P., Koebner, R. M. D., Reeves, J. C. and Cooke, R. J. (1997). Advances in Biometrical Genetics. Proceedings 10th Meeting of the EUCARPIA Section, Biometrics in Plant Breeding, Poznan.

Russell, J. R., Fuller, J. D., Macaulay, M., Hatz, B. G., Jahoor, A., Powell, W. and Waugh, R. (1997). Direct comparison of levels of genetic variation among barley accessions detected by RFLPs, AFLPs, SSRs and RAPDs. Theor. Appl. Genet 95, 714 - 722.

Smith, J. S. C., Smith, O. S., Bowen, S. L., Tenborg, R. A. and Wall, S. J. (1991). The description and assessment of distances between lines of maize. III. A revised scheme for testing of distinctness between inbred lines utilizing DNA RFLPs. Maydica 36, 213 - 226.

Table 1. Maize Morphology - Identification of Most Similar Variety

Ranked Position of the Most Similar Variety by Morphology in the set of Similarities Derived by Molecular Methods and Pedigree

Target Variety	Variety most Similar to Target	similarity Coefficient	Number of tied first rank (1)	AFLP	APPCR	RFLP	SSR	Pedigree $
1	21	0.964	1	34	28	15	18	13(=2)
2	11	0.960	1	1	1	2	2	3(=2)
3	2	0.942	1	28	24	27	22	27(=9)
4	7	0.899	1	7	6	6	8	10(=2)
5	26	0.970	1	1	1	1	2	1(=2)
6	1	0.954	1	16	16	15	5	15(=21)
7	1	0.957	1	4	8	7	4	3
8	35	0.950	1	33	23	31	27	31(=5)
9	12	0.973	1	1	1	1	1	1
10	27	0.966	1	20	22	11	13	7(=2)
11	31	0.971	1	3	3	1	1	1
12	9	0.973	1	1	1	1	2	1
13	27	0.905	1	33	34	33	34	8(=28)
14	20	0.961	1	3	2	1	11	5
15	21	0.963	1	10	4	6	6	7(=2)
16	29	0.950	1	1	1	1	1	1
17	22	0.948	1	8	10	7	10	8(=2)
18	20	0.944	1	12	5	20	4	13
19	32	0.936	1	1	1	1	2	1
20	14	0.961	1	3	2	2	27	2
21	1	0.964	1	29	21	18	15	11
22	27	0.948	1	18	25	14	14	10(=4)
23	27	0.925	1	9	16	8	14	6(=3)
24	17	0.922	1	2	2	1	1	1
25	28	0.913	1	10	7	5	20	13
26	5	0.970	1	1	1	1	2	2
27	10	0.966	1	11	6	8	25	7
28	21	0.933	1	15	14	8	5	9
29	10	0.955	1	3	2	1	4	2
30	5	0.911	1	31	21	13	31	15(=3)
31	11	0.971	1	1	1	1	1	3
32	1	0.960	1	2	2	1	1	1
33	23	0.916	1	20	21	34	34	19(=17)
34	14	0.951	1	3	3	3	23	1
35	8	0.950	1	23	20	19	23	16(=20)

$ Values in brackets are the number of tied ranks with equal similarity in pedigree.

Table 2. Maize - Identification of Most Similar Variety by Molecular Methods and Pedigree

Target Variety	Most Similar	AFLP	Most Similar	APPCR	Most Similar	RFLP	Most Similar	SSR	Most Similar	Pedigree
	AFLP	Similarity	APPCR	Similarity	RFLP	Similarity	SSR	Similarity	Pedigree	Similarity
1	32	0.826	32	0.819	32	0.725	32	0.942	33	0.909
2	11	0.885	11	0.869	31	0.717	31	0.943	31	0.885
3	15	0.466	15	0.493	15	0.366	15	0.837	15	0.437
4	31	0.713	31	0.679	31	0.541	31	0.887	31	0.500
5	26	0.879	26	0.901	26	0.799	8	0.953	26	0.882
6	13	0.709	13	0.738	13	0.552	13	0.738	13	0.500
7	32	0.600	32	0.600	31	0.470	20	0.813	31	0.502
8	26	0.800	26	0.846	5	0.777	5	0.953	26	0.882
9	12	0.915	12	0.894	12	0.862	12	0.918	12	0.947
10	29	0.625	29	0.620	29	0.561	29	0.809	29	0.509
11	2	0.885	2	0.869	31	0.779	31	0.952	31	0.885
12	9	0.915	9	0.894	9	0.862	8	0.926	9	0.947
13	6	0.709	6	0.738	6	0.552	25	0.798	25	0.500
14	33	0.681	34	0.648	20	0.603	26	0.843	34	0.553
15	26	0.495	26	0.507	5	0.418	3	0.837	3	0.437
16	29	0.673	29	0.701	29	0.536	29	0.884	29	0.500
17	24	0.684	24	0.706	24	0.543	24	0.910	31	0.500
18	12	0.593	12	0.645	9	0.461	9	0.773	12	0.577
19	32	0.839	32	0.862	32	0.717	1	0.901	32	0.821
20	34	0.758	34	0.682	34	0.673	34	0.867	34	0.524
21	18	0.472	18	0.577	18	0.372	3	0.790	18	0.267
22	11	0.686	31	0.709	31	0.663	31	0.870	31	0.546
23	29	0.626	29	0.602	13	0.499	29	0.812	29	0.500
24	31	0.709	31	0.761	17	0.543	17	0.910	31	0.500
25	13	0.667	13	0.671	13	0.447	13	0.798	13	0.500
26	5	0.879	5	0.901	5	0.799	8	0.944	8	0.882
27	34	0.772	34	0.691	34	0.578	34	0.849	29	0.500
28	10	0.578	17	0.644	17	0.537	24	0.770	10	0.509
29	16	0.673	16	0.701	10	0.561	16	0.884	22	0.546
30	11	0.842	2	0.862	11	0.697	11	0.934	31	0.885
31	11	0.837	11	0.857	11	0.779	11	0.952	30	0.885
32	19	0.839	19	0.862	1	0.725	1	0.942	1	0.908
33	1	0.712	1	0.738	1	0.686	1	0.891	1	0.909
34	27	0.772	27	0.691	20	0.673	20	0.867	14	0.553
35	21	0.441	10	0.497	28	0.275	21	0.706	21	0.187

Table 3. Summary of Individual Variety Correlations Between Characteristic Similarities

	MORPHOLOGY versus				
	AFLP	APPCR	SSR	RFLP	Pedigree
Max	0.68	0.64	0.68	0.69	0.69
Min	-0.36	-0.45	-0.30	-0.38	-0.34
Mean	0.23	0.24	0.18	0.25	0.27
Median	0.27	0.26	0.17	0.25	0.28
Range	1.03	1.09	0.98	1.07	1.03

	Pedigree versus				
	AFLP	APPCR	SSR	RFLP	Pedigree
Max	0.98	0.98	0.94	0.98	
Min	0.57	0.57	0.08	0.61	
Mean	0.90	0.90	0.74	0.92	
Median	0.92	0.92	0.80	0.93	
Range	0.40	0.40	0.86	0.37	

	SSRs versus				
	AFLP	APPCR	SSR	RFLP	Pedigree
Max	0.93	0.93		0.96	0.94
Min	0.10	0.26		0.13	0.08
Mean	0.72	0.73		0.75	0.74
Median	0.81	0.77		0.80	0.80
Range	0.83	0.66		0.82	0.86

Table 4.

Most Similar (1st, 2nd and 3rd) Variety by Target Method (Morphology, AFLP and SSR) Contained in top x% by other Methods.

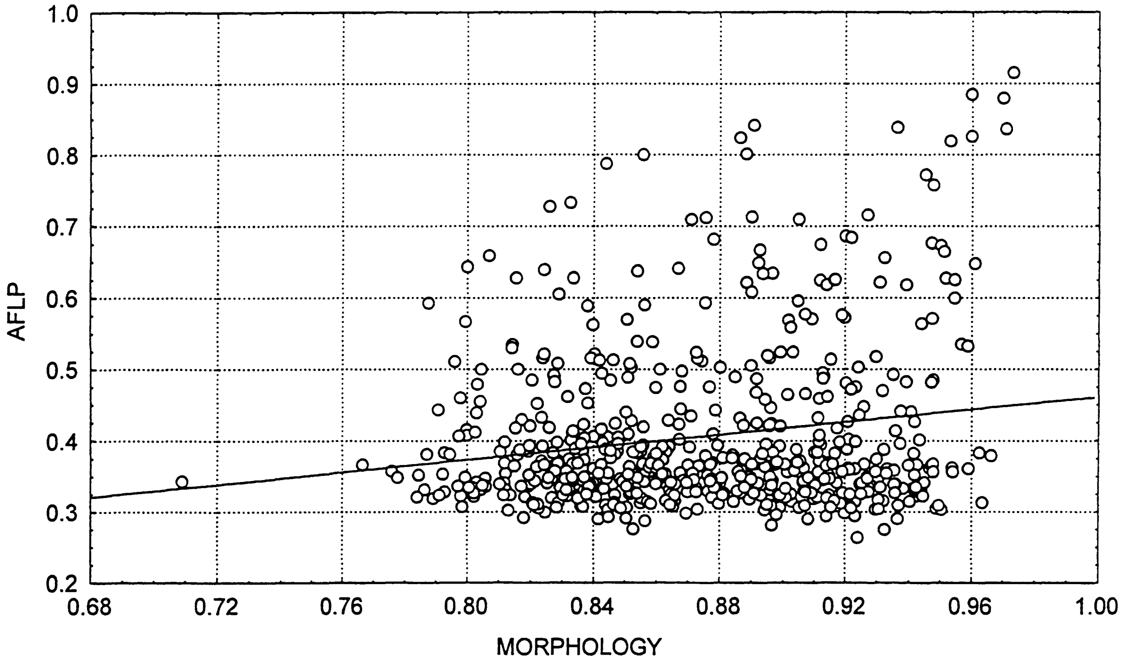
	Target Method Morphology					Target Method AFLP					Target Method SSR		
		1st	2nd	3rd			1st	2nd	3rd			1st	2nd	3rd
Morphology	10%	~	~	~	Morphology	10%	69	23	29	Morphology	10%	57	37	28
	20%	~	~	~		20%	77	43	34		20%	69	43	40
	30%	~	~	~		30%	80	60	46		30%	71	63	51
AFLP	10%	46	49	34	AFLP	10%	~	~	~	AFLP	10%	89	66	69
	20%	49	63	46		20%	~	~	~		20%	89	74	86
	30%	63	63	51		30%	~	~	~		30%	91	83	89
SSR •	10%	40	51	37	SSR	10%	89	63	49	SSR	10%	~	~	~
	20%	49	54	46		20%	97	77	71		20%	~	~	~
	30%	57	63	46		30%	100	91	77		30%	~	~	~
RFLP	10%	46	57	69	RFLP	10%	100	91	71	RFLP	10%	91	71	69
	20%	57	66	66		20%	100	91	97		20%	91	80	74
	30%	49	49	57		30%	100	94	97		30%	97	86	89
APPCR	10%	43	54	60	APPCR	10%	100	83	77	APPCR	10%	89	74	71
	20%	57	60	66		20%	100	91	94		20%	91	84	86
	30%	43	46	49		30%	100	94	97		30%	91	84	94
Pedigree	10%	43	46	40	Pedigree	10%	97	83	87	Pedigree	10%	94	66	51
	20%	54	60	46		20%	100	94	94		20%	97	71	71
	30%	69	63	54		30%	100	97	97		30%	97	83	97

% Total Miss Rate ie the % of 1st, 2nd and 3rd Similar Varieties not contained in the Top 30% of any Other Method.

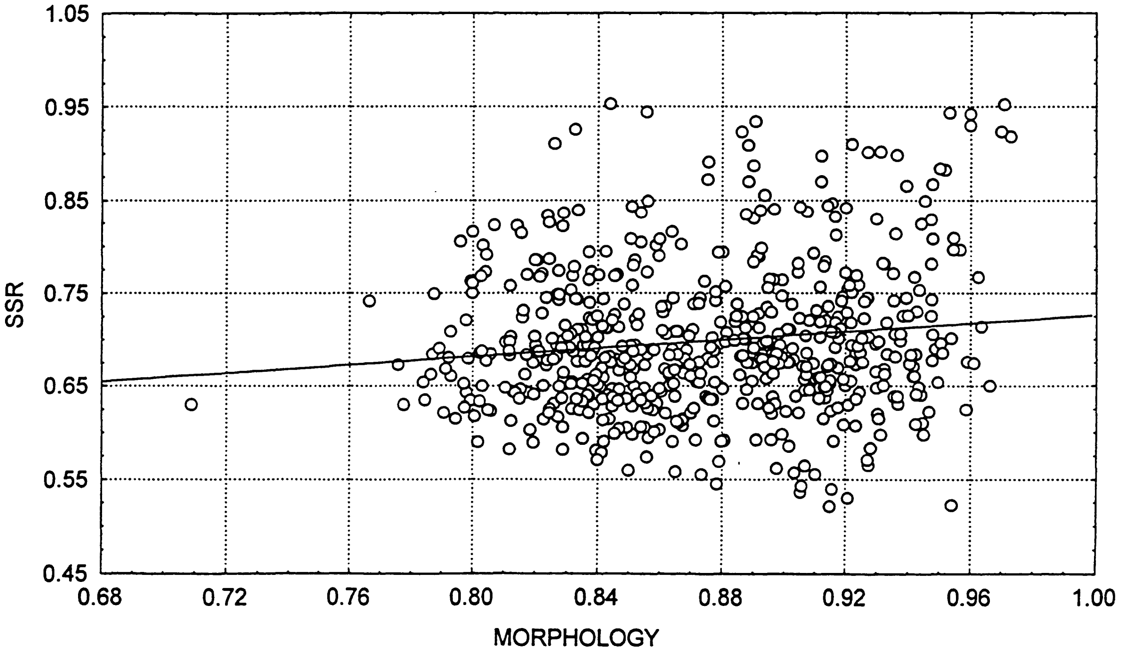
	1st	2nd	3rd
Morphology	20	31	46
AFLP	0	0	0
SSR	0	9	0

Figure 1
Scatter Plot of Pairwise Similarities from AFLP, SSR, APPCR, RFLP and Pedigree v Morphology

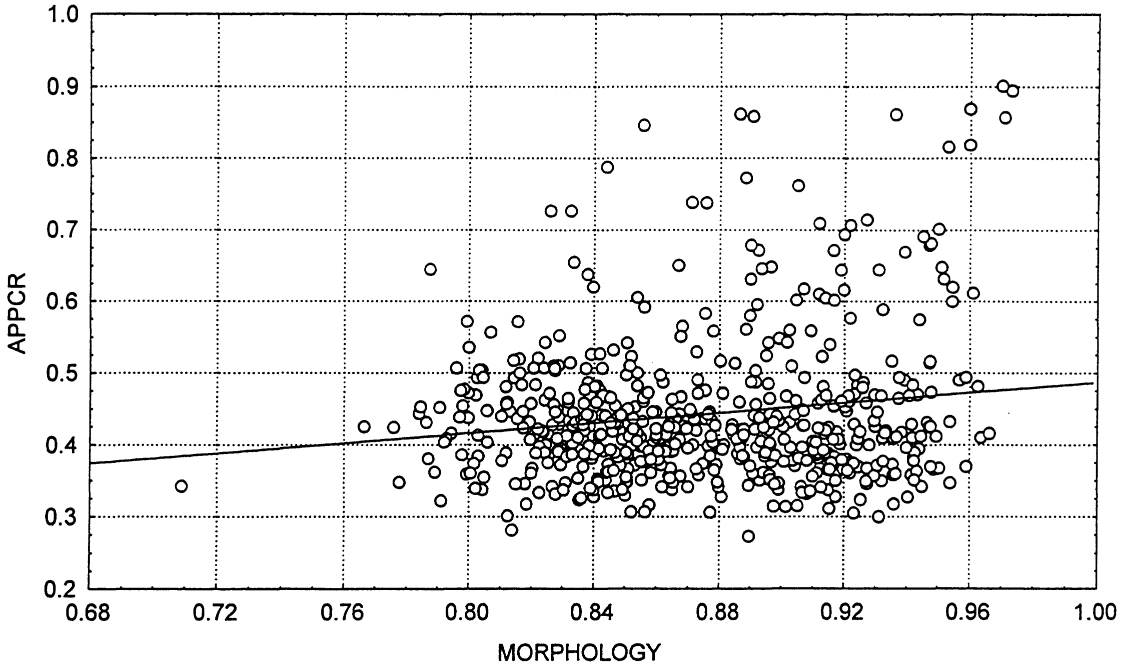
MAIZE RELATIONSHIPS
AFLP v MORPHOLOGY SIMILARITIES



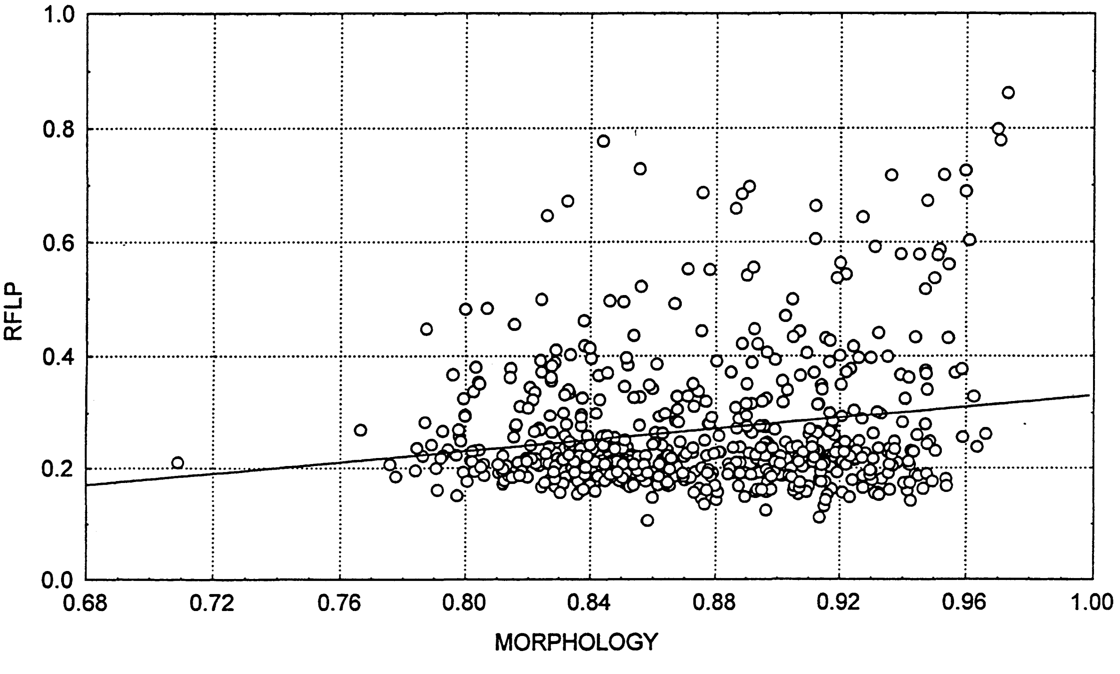
MAIZE RELATIONSHIPS
SSR v MORPHOLOGY



MAIZE RELATIONSHIPS
APPCR v MORPHOLOGY



MAIZE RELATIONSHIPS
RFLP(PI) v MORPHOLOGY



MAIZE RELATIONSHIPS
PEDIGREE v MORPHOLOGY SIMILARITIES

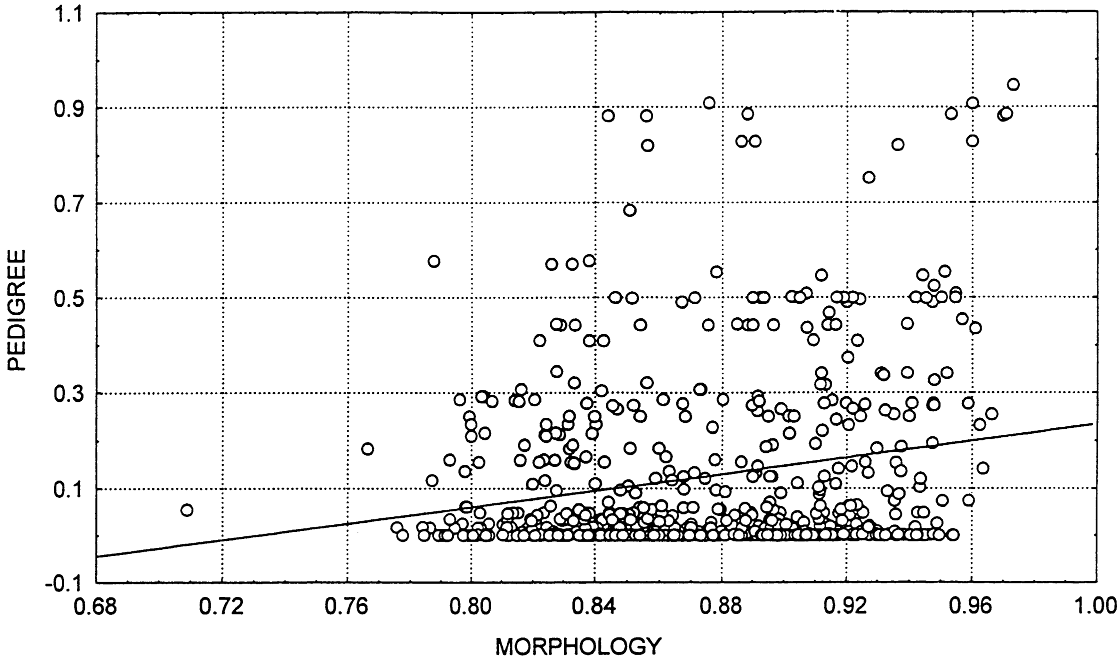
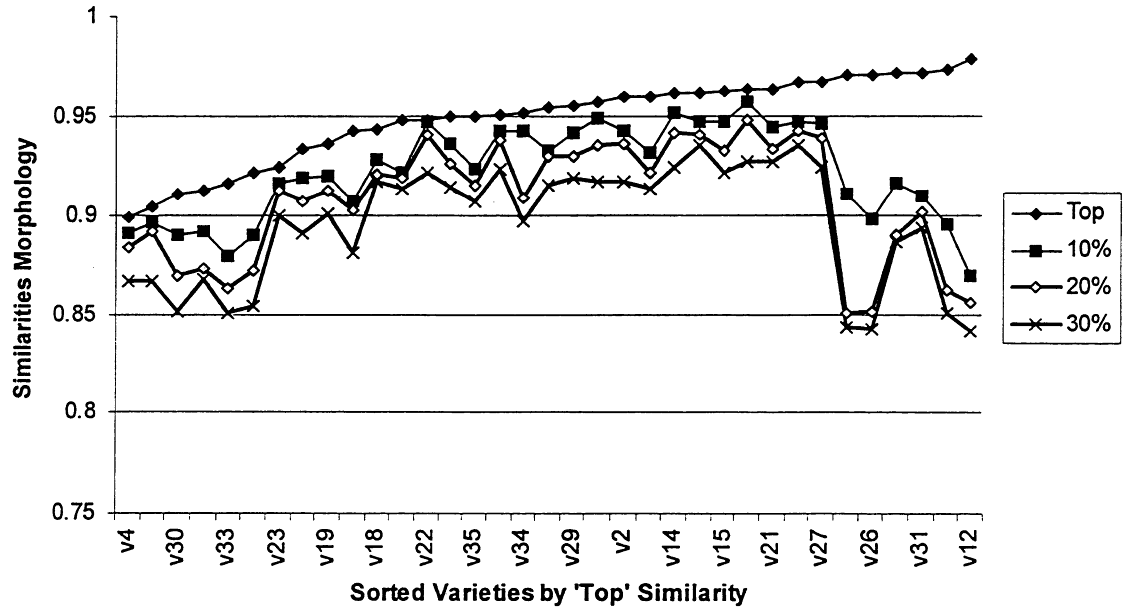


Fig 2. Maize (BMT) Morphology ~ Critical Cut-off Similarity Values
Most Similar (Top), 10%, 20% and 30% Points



[End of document]