P. S. Virk · H. J. Newbury · M. T. Jackson
B. V. Ford-Lloyd

The identification of duplicate accessions within a rice germplasm collection using RAPD analysis

Received: 8 September 1994 / Accepted: 8 December 1994

Abstract A set of accessions of *Oryza sativa* from the International Rice Research Institute (Philippines) that included known and suspected duplicates as well as closely related germplasm has been subjected to RAPD analysis. The number of primers, the number of polymorphic bands and the total number of bands were determined that will allow the accurate discrimination of these categories of accessions, including the identification of true and suspected duplicates. Two procedures have been described that could be employed on a more general basis for identifying duplicates in genetic resources collections, and further discussion on the values of such activities is presented.

Key words *Oryza sativa* · Rice · Genetic resources · RAPD · Molecular markers · Cluster analysis

Introduction

Plant germ plasm has been accumulated over many decades and is now stored in genebanks in countries around the world. The Consultative Group on International Agricultural Research (CGIAR) Centres conserve about 500,000 accessions of more than 30 crop species and wild relatives. The gene bank at the International Rice Research Institute (IRRI) in the Philippines has a collection of more than 80,000 accessions of rice, including *Oryza sativa, O. glaberrima,* the 20 wild species in the genus *Oryza* and representative species from all genera in the tribe Oryzeae. Management of a large germplasm collection is a complex and costly task. To ensure long-term preservation, management procedures are aimed at safe multiplication or periodic rejuvenation of the conserved germplasm in order to reduce the chances of genetic erosion during these activities. The value of germplasm is enhanced by obtaining additional information concerning patterns of morphological variation (characterisation) and responses to biotic and abiotic stresses (evaluation). In this way, plant breeders and other researchers can select germplasm for their specific needs. Lastly, particularly in the case of seeds such as those of rice, which can be stored at low moisture content and low temperature (the so-called orthodox seeds), it is essential to place in storage only those materials that have a high germination rate and therefore a predicted longevity in storage. This entails carrying out germination tests prior to storage and periodically monitoring the delay of germination potential over time.

The race against genetic erosion has yielded thousands of accessions safely stored in genebanks. However, genebanks have a finite capacity, and it is apparent that they often conserve more than one sample of the same genotype. In other words, the plant genetic resources collections contain duplicate materials, yet the scale of the problem can not be determined with certainty.

From a purely management point of view, there are distinct advantages in trying to identify duplicate accessions and thereby only conserving in the collection unique genetic materials. Until now the identification of duplicate accessions has had to rely on a comparison of morphological characters, some of which are subject to environmental variation, together with passport data including (amongst others) variety name and origin. The identification of duplicates of vegetatively propagated species, such as potato for example, is more straightforward than for seed-propagated crops such as rice. At the International Potato Center (CIP) duplicate accessions have been routinely identified through the comparison of tuber proteins separated in polyacrylamide gels, and complemented by field observations of morphological characters. However, now it is possible to make routine use of molecular markers based directly upon genomic DNA for the identification of duplicate accessions of seed-propagated species.

Communicated by J. W. Snape

P. S. Virk (✉) H. J. Newbury B. V. Ford-Lloyd
School of Biological Sciences, University of Birmingham,
Edgbaston, Birmingham B15 2TT, UK

M. T. Jackson
International Rice Research Institute, PO Box 933, Los Baños
Manila, Philippines

We have tested a procedure for the designation of duplicates using DNA-based markers, and the results are presented here. For this purpose a set of 44 rice accessions was used at IRRI to compile 47 seed samples that included pairs of true duplicates and suspected duplicates. These were characterised for morphological features in The Philippines, and the random amplification of polymorphic DNA (RAPD) analysis was carried out in Birmingham. The efficiencies of morphological and molecular techniques for recognising duplicates have been compared.

Materials and methods

Material

Forty-four accessions of *O. sativa* comprised material for the present study (Table 1). Seeds of 3 accessions arbitrarily chosen from the above were split into two lots at IRRI (we shall refer to these as true duplicates) and all 47 seed packets were allocated random numbers so as to undertake the random amplified polymorphic DNA (RAPD) analyses in a blind fashion. Among these 47 were three pairs of accessions which had been classified as 'suspected duplicates' in the IRRI germplasm collection on the basis of similarities of name, place, country of origin and seed source etc. Nine accessions [Random numbers (RN) 2, 7, 9, 12, 30, 33, 35, 38 and 39] derived by inbreeding a landrace 'Radin Ebos' represent closely related materials, while the remaining 26 accessions representing ecogeographical diversity were chosen at random.

Morphological data

Morphological data were obtained during routine germplasm characterisation at Los Baños, Philippines. Data were available for all accessions except 'IRGC 12048' and '71543' for 36 qualitative and quantitative traits scored based on the IBPGR-IRRI descriptors for rice (1980) on ten representative plants of each accession. However, in the present investigation we analysed data for only 22 qualitative traits: blade pubescence, blade colour, basal leaf sheath colour, flag leaf angle, ligule colour, collar colour, auricle colour, culm angle, internode colour, culm strength, panicle type, panicle exsertion, panicle threshability, awning, awn colour, apiculus colour, stigma colour, lemma and palea colour, spikelet fertility, seed coat colour, endosperm type and leaf senescence.

Table 1 List of accessions used in the present investigation

Accession number (IRGC)	Name	Random number	Origin
1723	Carolina Gold	3	USA
3627	Mas 2401	8 & 24	Indonesia
5418	Sintane Diofor	34	Burkina Faso
5854	DA9	16	Bangladesh
6087	Makalioka 34	21	Madagascar
6464	Tres Meses	10	Brazil
8191	Mansaku	14	Japan
8234	RTS12	41	Vietnam
8237	Peh-Kuh-Tsao-Tu	31	Taiwan
8245	Macan Binundok	29	Philippines
9074	JC 157	43	India
10769	Lambayque 1	17	Peru
12048	Moroberekan	18	Guinea
13375	Jumula 2	32	Nepal
13861	Radin Ebos 63	9	Malaysia
13863	Radin Ebos 67	7	Malaysia
13865	Radin Ebos 72	2	Malaysia
13866	Radin Ebos 74	30	Malaysia
13867	Radin Ebos 75	12	Malaysia
14202	Radin Ebos 32	39	Malaysia
14204	Radin Ebos 35	38	Malaysia
14207	Radin Ebos 39	33	Malaysia
14482	Radin Ebos 33	35	Malaysia
14483	Radin Kuning	1	Malaysia
15292	Sinnanayam	4	Sri Lanka
15310	Sinnakayan	46	Sri Lanka
16073	Pate Blanc Mn 1	44	Ivory Coast
16381	Djawa Gempolan	42 & 45	Indonesia
23729	Hawm Om	19	Thailand
24649	Sikuneng	47	Indonesia
24651	Si Kuning	22	Indonesia
27967	Jhona 26	28	Pakistan
29676	Sa Tang	15	Laos
29726	Chaing Roneal	25	Cambodia
30359	Kao Ipoua	13 & 27	Laos
32362	Tchampa	23	Iran
32399	Phudugey	36	Bhutan
33189	Kaukkyi Ani	20	Myanmar
33270	Kyar Amagyi	5	Myanmar
33272	Kyarmagyi	40	Myanmar
40275	Black Gora (NCS 12)	26	India
43394	Gogo Lempuk	37	Indonesia
51064	Sinna Sithira Kali	11	Sri Lanka
71643	Undus	6	Malaysia

DNA extraction and polymerase chain reaction (PCR)

DNA was extracted from 20 mg bulked leaf material obtained in equal amounts from ten 2-week-old seedlings of each sample following Virk et al. (1995), and its concentration was monitored by running samples on 0.7% (w/v) agarose gels in TBE buffer (Sambrook et al. 1989).

A total of 53 decanucleotides of arbitrary sequence were used in the present study. Three primers namely, BFL-02 (5′GGGAGAGTCA), BFL-12 (5′GTGCGTATGG) and BFL-13 (5′GACAGACAGA), were obtained from Alta Bioscience (University of Birmingham), while the other primers were purchased from Operon Technologies Inc. These were OPA-01 to OPA-20; OPK-01 to OPK-20; OPF-01, F-02, F-06, F-09, F-13, F-14, F-17; OPC-07, C-15; OPD-08.

Unless otherwise indicated, the DNA amplification reactions were performed in a volume of 25 μl containing approximately 1 ng genomic DNA, 200 μ*M* each of dATP, dCTP, dGTP, dTTP, 0.2 μ*M* of a primer, 0.5 U *Taq* polymerase and 1× incubation buffer containing 1.5 m*M* magnesium chloride (Boehringer Mannheim). The mixture was gently mixed and centrifuged prior to adding 2 drops of mineral oil. The amplification was performed in a thermocycler (Hybaid-omnigene) programmed as follows: 1 cycle of 94°C for 5 min followed by 45 cycles of 30 s at 94°C, 1 min at 35°C and 2 min at 72°C and finally 1 cycle of 72°C for 5 min. Aliquots of 15 μl of amplification products were loaded in 1.2% (w/v) agarose gels for electrophoresis in 1× TBE buffer (Sambrook et al. 1989). Gels were stained with ethidium bromide and photographed under UV light using Polaroid 667 film.

Data analysis

For analysis of morphological traits, a coordinate data matrix was subjected to the agglomerative hierarchical clustering method (UPGMA, unweighted pair group method using arithmetic averages). Before performing the cluster analysis on coordinate data, the morphological variables were standardised to mean 0 and standard deviation 1 (SAS 1990; MINITAB 1994).

RAPD products were scored as present (1) or absent (0) for each of the primer-accession combinations. Pairwise comparisons of ac-

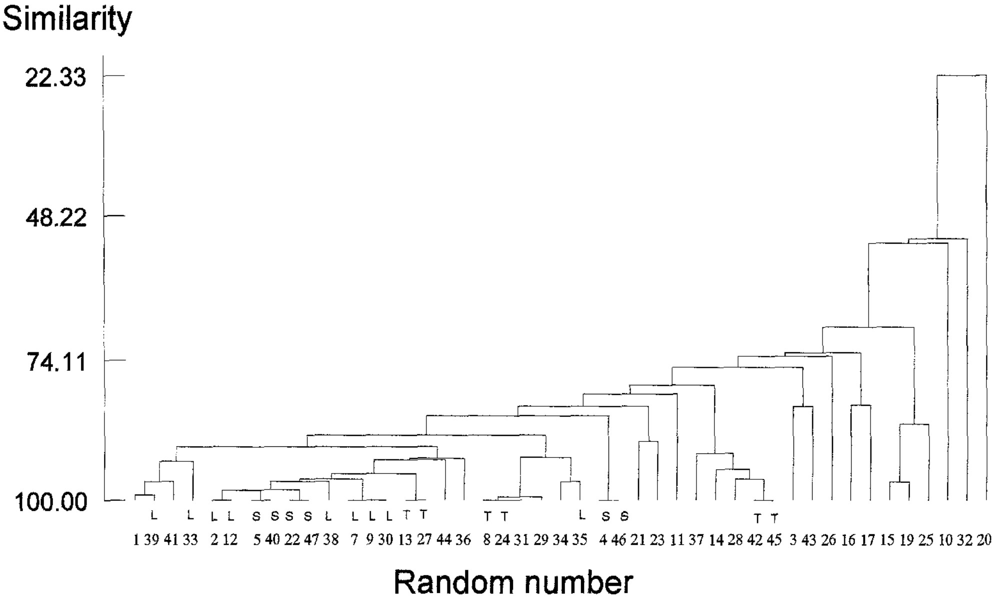

Fig. 1 Clustering of 45 samples of *O. sativa*, as designated in Table 1, based on morphological data for 22 characters (*S* and *T* represent suspected and true duplicates, respectively, while *L* depicts material generated from a landrace)

cessions, based on the presence or absence of unique and shared products, were used to generate similarity coefficients (see Virk et al. 1995 for details). The similarity coefficients were then used to construct a dendrogram by the UPGMA method (SAS 1990; MINITAB 1994).

Results

Morphological markers

A dendrogram showing the relationships between accessions is presented in Fig. 1. For convenience we shall refer to accessions by their allocated random numbers (RN) and not the IRGC numbers (Table 1). The three pairs of true duplicates (RN 13 and 27; 8 and 24; 42 and 45) and three pairs of suspected duplicates (RN 4 and 46; 5 and 40; 22 and 47) paired together at 100% similarity (Fig. 1). The 9 accessions derived from a landrace could be placed into six groups that were clearly related (>80% similarity). Of these, accessions RN 2 and 12, RN 7, 9 and 30 were found to be related at 100% similarity, whilst accession RN31 clustered with a pair of true duplicates (RN8 and 24).

Molecular markers

Forty-seven samples were subjected to RAPD assay firstly utilising only 12 primers (OPC-07 and -15; OPD-08; OPF-06, -13, -14 and-17; OPK-16, -17 and -20; BFL-12 and -13) to test whether the pattern of diversity revealed reflects known relationships among these accessions. Standard protocol (material and methods) was followed for all 12 primers except for 2, OPC-07 and OPF-17, for which a higher concentration of oligonucleotide (0.8 μ*M*) was used (Virk et al. 1995). The 12 primers yielded 63 strong reproducible and easily scorable bands of which 32 were polymorphic. Of these primers 9 (all except the three K kit primers) were chosen a priori based on their efficiency in producing reliable and easily scorable banding patterns in rice germplasm (Virk et al. 1995). The remaining 3 primers were selected from 20 K kit primers by systematic screening to obtain further resolution within the diversity patterns. For example, 2 accessions, RN 37 and RN 44 (which were neither suspected nor true duplicates), were not separated on the dendrogram on the basis of markers yielded by 9 primers; however, the patterns produced by OPK-20 discriminated between them (data not shown).

Cluster analysis (UPGMA method) was performed on dissimilarity indices, computed from both monomorphic and polymorphic fragments, to generate a dendrogram (Fig. 2). The three pairs of true duplicates (RN 13 and 27; 8 and 24; 42 and 45) and two pairs of suspected duplicates (RN 4 and 46; 5 and 40) clustered together at 100% similarity. However, 2 accessions suspected to be a duplicate pair (RN 22 and 47) could be differentiated from each other at 93% similarity. Nine closely related accessions (derived from a landrace) were placed into four groups. Six of these (RN 12, 30, 33, 35, 38 and 39) were placed in one group, while RN 2 and 7, although different from each other, were more closely related to these 6 than RN9.

The above results demonstrate that the pattern of variation revealed by RAPD data closely reflects the previously understood relationships between the rice samples. This encouraged us to concentrate on our primary objective of developing methods for designating duplicates. We therefore generated further RAPD data only on the three pairs of suspected duplicates and one pair of true duplicates (RN 8 and 24). A set of 20 primers (OPK-01 to OPK-20) was used to generate RAPD data for these 8 samples. Three primers namely, OPK-02, -03 and -05 did not yield reproducible banding patterns, and 3 others (OPK-16, -17 and -20) had also been used on the complete set of 47 samples. In all, 77 bands were scored, of which 25 were

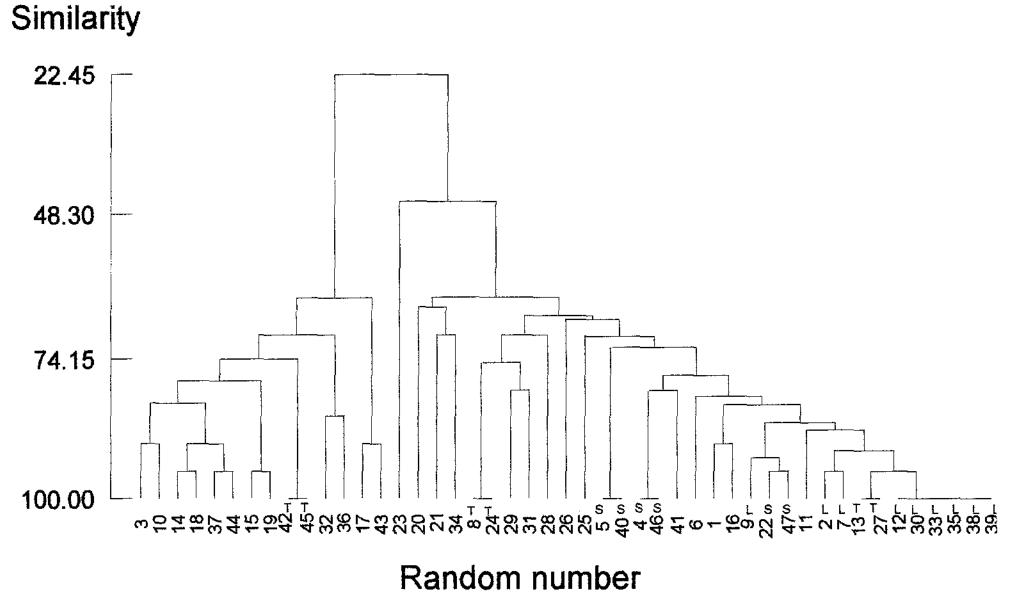

Fig. 2 Dendrogram of 47 samples generated by clustering of dissimilarity coefficient values computed from pairwise comparisons of 63 RAPD fragments (*S*, *T* and *L* are defined in Fig. 1)

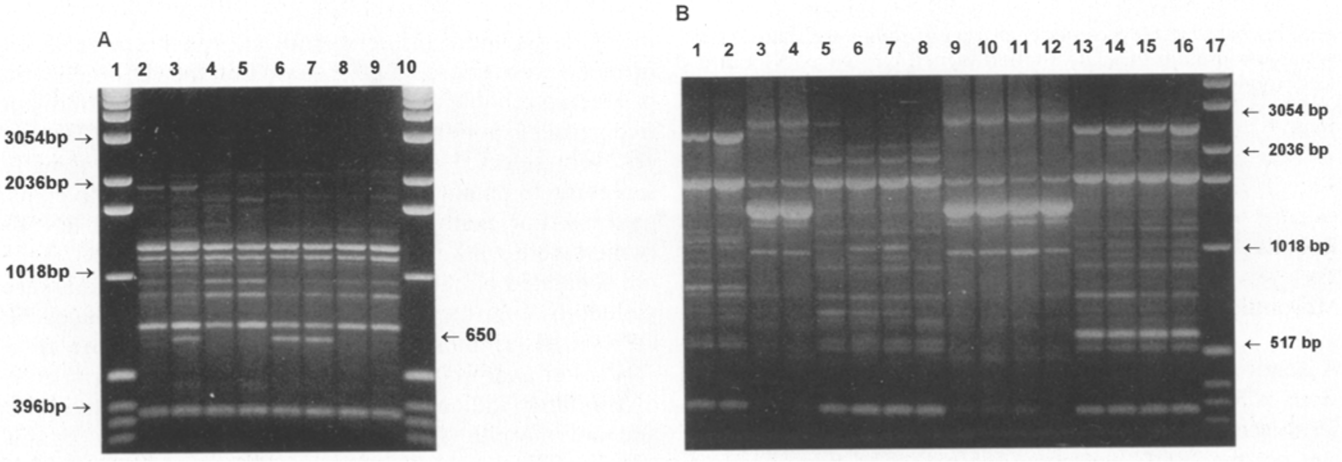

Fig. 3 **A** *Lanes 1* and *10* 1-kb ladder, *Lanes 2–9* amplification products following PCR directed by primer OPK-11 using DNA from RN 22 and 47, 4 and 46, 5 and 40 suspected duplicate pairs and RN8 and 24 true duplicate pair. **B** *Lane 17* 1-kb ladder, *Lanes 1–16* amplification products following PCR directed by primer OPK-17 using DNA from RN 22 and 47, 4 and 46, 5 and 40 suspected duplicate pairs and RN8 and 24 true duplicate pair (2 lanes for each sample)

polymorphic across the 8 samples. However, no differences between each pair of suspected or true duplicates were observed except between one pair of suspected duplicates (RN 22 and 47). Here, 4 bands were found to be polymorphic between RN22 and RN47; these were amplified using OPK-11 (650 bp; Fig. 3a), OPK-17 (2,900 bp; Fig. 3b) and OPK-18 (3 kbp and 3.5 kbp) primers (not shown).

These empirical results were then used to predict the number of primers, total number of markers and number of polymorphic markers required to detect a difference between suspected duplicates. Of the 17 primers utilised, 14 did not discriminate between this pair. Therefore, the number of primers (n) required to detect at least one difference between a pair of suspected duplicates with a 99% confidence can be calculated by solving the following: $(14/17)^n=1-0.99$, which is 24. Similarly, the total number of markers (m) and the number of polymorphic markers (p) required can be calculated from $(73/77)^m=0.01$ (86), and $(21/25)^p=0.01$ (26), respectively. Therefore, for a 99% probability of finding one or more differences between a pair of suspected duplicates we need to screen the material with 24 primers, or for a total of 86 markers, or 26 polymorphic markers.

To test these predictions, based on the use of 17 K kit primers, we then screened these 8 samples using another set of 24 primers that had not yet been used with this material (OPA-01 to OPA-20; OPF-01, -02 and -09; BFL-02). Using all 24 primers, we found 3 polymorphisms between the pair of suspected duplicates (RN 22 and RN47) from a total of 109 markers scored, of which 41 were generally polymorphic. A close agreement between the predicted and observed number of markers ($\chi^2_{(1)}=0.26$, $P>0.50$), number of polymorphic markers ($\chi^2_{(1)}=0.64$, $P>0.25$) and number of primers ($\chi^2_{(1)}=0.25$, $P>0.50$) was observed. In fact, the first polymorphism between this pair of suspected dupli-

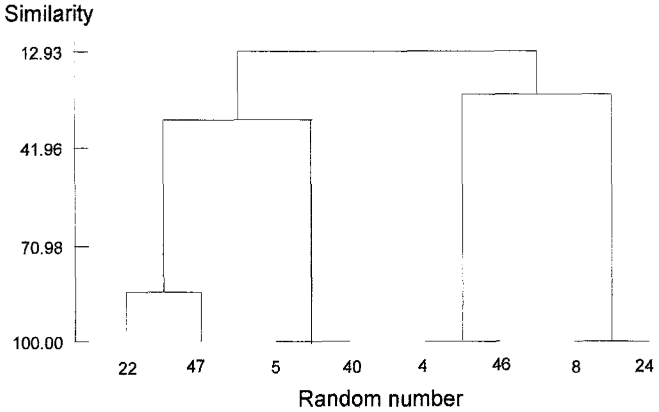

Fig. 4 Dendrogram of 8 samples based on 233 RAPD markers

cates was detected when we were screening with the 13th primer and scoring the 63rd band, which was the 24th polymorphic marker.

Subsequently, collation of all the RAPD data available for the 8 samples using 50 primers allowed the use of 233 marker bands (75 polymorphic). Associations among these 8 samples as revealed by cluster analysis are presented in Fig. 4. As with the previous analysis (Fig. 2) a pair of suspected duplicates (RN 22 and 47) was discriminated, whilst the other suspected duplicates still grouped together. A dendrogram obtained (not shown) from 109 markers (41 polymorphic) amplified from 24 primers (used to test predictions with the set of 8 samples described above) was strikingly similar to the one presented in Fig. 4.

Discussion

In this study, the three sets of true duplicates serve as controls in that they can be regarded as replicate samples. The RAPD banding patterns did not distinguish between them, not only indicating the reliability of the molecular technology, but also the effectiveness of our sampling technique. We have previously been able to detect differences between individuals within a single accession of rice using RAPD. However, pooling material from ten individuals for DNA extraction (Virk et al. 1995) has clearly produced representative DNA samples from these accessions so that no differences between these 'replicates' have been detected during this work, even though 50 primers were employed with one duplicate pair.

Across the 44 accessions, the pattern of diversity revealed using RAPD data reflects closely the pattern of diversity either known (in the case of true duplicates), expected (in the case of the landrace-derived material) or revealed using an analysis of morphological characters (following measurements in the field). This close relationship has previously been reported for rice (Virk et al. 1995), banana (Howell et al. 1994), *Brassica* (Demeke et al. 1992), *Lotus* (Campos et al. 1994) and *Stylosanthes* (Kazan et al. 1993). In our study, the classification of the 8 samples into groups using 233 markers was the same as that obtained using 109 markers. The stability of the classifications (although based on a small sample) demonstrates that no advantage is conferred by using more than a certain number of markers for the analysis of the general pattern of diversity. However, the higher the number of markers used, the more confidence there will be in testing for duplicates.

In the gene bank at IRRI, accessions are routinely characterised for a list of qualitative and quantitative characters. These data were collected for 42 accessions from ten plants of each accession and were scored in different seasons (dry and wet) as well as in different years (for most of the accessions between 1987 and 1991). The utility of measurements, particularly of those traits that are influenced by environmental variation, is therefore limited for studying the patterns of diversity. However, a number of qualitative characters which are less likely to be affected by environmental variation have been used together with variety name to select the three sets of suspected duplicates used here. Using the 22 morphological characters selected on the basis that they are less likely to be affected by environmental factors, we were unable to distinguish any differences between any of the three pairs of suspected duplicates. However, one pair of suspected duplicates was separated by RAPD analysis with 7 out of a total of 233 marker bands showing a difference.

These results provide information useful in the design of procedures that would allow the routine identification of duplicates within a collection. Discussions concerning the number of marker bands to score before the designation of duplicates are complex. It will not be possible to prove that 2 accessions are genetically identical without sequencing their entire genomes. Given that this is impossible, a decision must be made about the amount of testing that will be performed before 2 accessions are accepted as (or 'designated' as) duplicates. This decision must be influenced by the number of potential duplicates that are to be tested. However, our results indicate that, for one very similar pair of accessions, we can be 99% confident of detecting a difference between them if we examine a total of 86 RAPD markers. It would also be possible to use other types of DNA-based markers for this purpose, although we think it would be important to ensure that the variation defined using alternative markers was biologically valid in terms of taxonomy and genetics. Moreover, the number of bands to be scored may differ depending upon the sequence types represented by the markers.

If it is accepted that, of the material selected at IRRI, two pairs of suspected duplicates be designated duplicate, and that one pair be maintained as two separate accessions, then a rational general strategy would be to designate accessions as duplicate if they showed no variation across 100 RAPD markers. The selection of 100 markers is slightly arbitrary and should certainly be the subject of debate, but operationally genebank managers would at some stage need to select a level of similarity at which accessions are designated as duplicates.

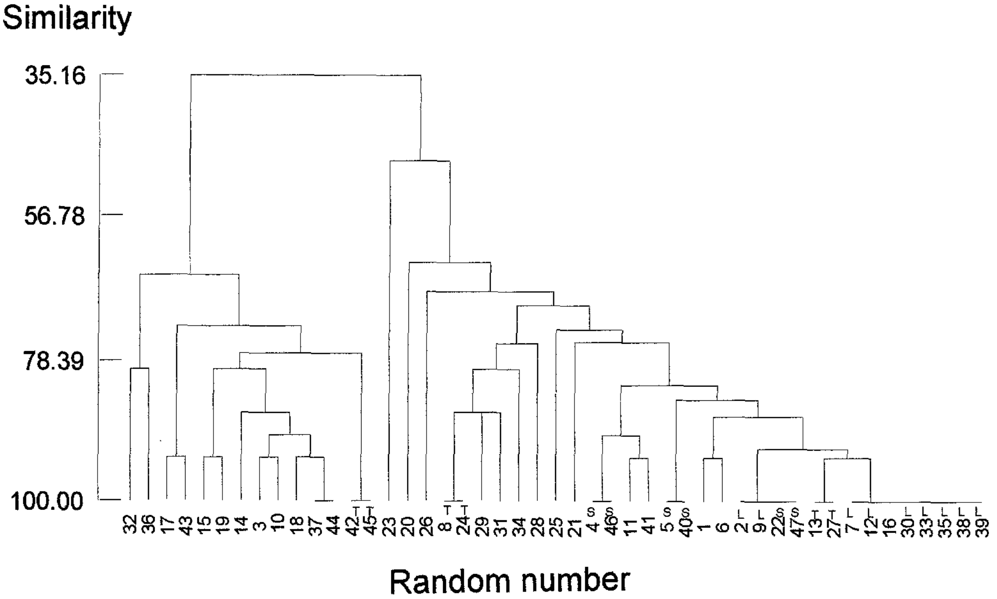

Fig. 5 Average linkage clustering of 47 samples employing data on 22 polymorphic RAPD bands (*S*, *T* and *L* are defined in Fig. 1)

On the basis of our work, we can tentatively propose two procedures for designating duplicates among accessions currently held within the rice collection using RAPD markers. In both, potential duplicates would first be selected from the collection following an examination of passport data. Procedure 1 would then require initial morphological characterisation of suspected duplicates; those accessions that could not be separated using these data would then be subjected to full RAPD analysis (comparison of 100 RAPD bands). Accessions that could not be discriminated would be designated as duplicates. Procedure 2 would involve instead of morphological evaluation a prescreen of pairs of suspected duplicates using RAPD but only employing 2 or 3 primers; those which could not be separated using these data would then be subjected to full RAPD analysis (i.e. 100 RAPD band comparison), and again accessions that could not be discriminated would be designated as duplicates.

The procedures differ in the method used for an initial screen before full RAPD analysis, and the relative merits of these methods would have to be balanced by those charged with conserving a collection. Local situations may vary considerably with regard to the relative costs of field work and molecular biology, and the expertise available to carry these out. Care must be taken when utilising morphological data routinely held within collections because, as previously discussed, such data may be subject to variation caused by the environment. Either the characters that are less sensitive to seasonal changes can be used, or further tests need to be carried out. RAPD analysis, along with other PCR-based forms of molecular characterisation, is not subject to environmental effects and offers a high degree of resolution in duplicate screening if only because the number of scorable molecular characters (bands) is high; in fact the number of RAPD markers available is almost infinitely high if one simply continues the screening with more and more primers.

It is possible to make comparisons of the relative average effectiveness of RAPD and morphological markers, and this may be useful when considering the nature of a pre-screen before full RAPD analysis. A close association between these two methods was suggested by a significant correlation (r=0.38±0.09, P< 0.01) between the two matrices (Mantel 1967) used in the cluster analyses. The 22 morphological markers were used here and cluster analysis (Fig. 1) suggests that, following procedure 1 above, 18 accessions, showing 100% similarity, would have been selected for further analysis during duplicate testing. If, on the other hand, an arbitrary set of 22 polymorphic RAPD markers (in this case obtained using 6 primers) were used in procedure 2, the cluster analysis (Fig. 5) suggested that 24 samples should be further tested. The difference in the number of samples for further testing is not large between these two procedures. Furthermore, the number of samples for further testing reduces to 13 and 15 for morphological and RAPD markers, respectively, if the set of 9 closely related accessions derived from a single landrace are removed from the analysis. How such procedures are applied to existing germplasm held in trust versus newly acquired, incoming germplasm may also vary and will also require discussion elsewhere.

The reliable identification of duplicate accessions will provide management options for the germplasm curator. Whether it will lead to the reduction in size of germplasm collections is debatable. In the case of CGIAR centres, their germplasm collections are held in trust, and many accessions are actually duplicate materials of existing national germplasm collections. Clearly, the centres have an obligation to continue to conserve those germplasm accessions already accepted for safety storage. With the acquisition of new germplasm accessions, however, the situation is potentially different. Our study suggests a novel procedure which would allow the level of certainty of identifying duplicate samples to be set before those samples became part of a germplasm collection and before they were assigned unique accession numbers. This option is one which is likely to have a significant impact on germplasm management, provided the PCR-based marker technology can be easily and economically utilised by germplasm curators.

Acknowledgements This work was funded by the Overseas Development Administration, Plant Sciences Programme, Project Number R5059.

References

Campos LP, Raelson JV, Grant WF (1994) Genome relationships among *Lotus* species based on random amplified polymorphic DNA (RAPD). Theor Appl Genet 88:417–422

Demeke T, Adams RP, Chibbar R (1992) Potential taxonomic use of random amplified polymorphic DNA (RAPD) – a case study in *Brassica.* Theor Appl Genet 84:990–994

Howell EC, Newbury HJ, Swennen RL, Withers LA, Ford-Lloyd BV (1994) The use of RAPD for identifying and classifying *Musa* germplasm. Genome 37:328–332

IRRI (International Rice Research Institute) (1980) Descriptors for rice *Oryza sativa L.* IBPGR-IRRI rice advisory committee. Manila, Philippines

Kazan K, Manners JM, Cameron DF (1993) Genetic variation in agronomically important species of *Stylosanthes* determined using random amplified polymorphic DNA markers. Theor Appl Genet 85:882–888

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27:209–220

MINITAB (1994) MINITAB User's guide. Release 10 for windows. State college, Philadelphia, Pen.

Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

SAS (1990) SAS/STAT User's guide. Version 6, 4th edn, vol 1. SAS Institute, Cary, N.C.

Virk PS, Ford-Lloyd BV, Jackson MT, Newbury HJ (1995) Use of RAPD for the study of diversity within plant germplasm collections. Heredity 74:170–179