Theor Appl Genet (2005) 110: 1092–1098
DOI 10.1007/s00122-005-1937-4

ORIGINAL PAPER

Ling Xia · Kaiman Peng · Shiying Yang · Peter Wenzl
M. Carmen de Vicente · Martin Fregene · Andrzej Kilian

DArT for high-throughput genotyping of Cassava (*Manihot esculenta*) and its wild relatives

Received: 7 August 2004 / Accepted: 6 January 2005 / Published online: 2 March 2005

Abstract Understanding the distribution of genetic diversity within and among individuals, populations, species and gene pools is crucial for the efficient management of germplasm collections. Molecular markers are playing an increasing role in germplasm characterization, yet their broad application is limited by the availability of markers, the costs and the low throughput of existing technologies. This is particularly true for crops of resource-poor farmers such as cassava, *Manihot esculenta*. Here we report on the development of Diversity Arrays Technology (DArT) for cassava. DArT uses microarrays to detect DNA polymorphism at several hundred genomic loci in a single assay without relying on DNA sequence information. We tested three complexity reduction methods and selected the two that generated genomic representations with the largest frequency of polymorphic clones (*Pst*I/*Taq*I: 14.6%, *Pst*I/*Bst*NI: 17.2%) to produce large genotyping arrays. Nearly 1,000 candidate polymorphic clones were detected on the two arrays. The performance of the *Pst*I/*Taq*I array was validated by typing a group of 38 accessions, 24 of them in duplicate. The average call rate was 98.1%, and the scoring reproducibility was 99.8%. DArT markers displayed fairly high polymorphism information content (PIC) values and revealed genetic relationships among the samples consistent with the information available on these samples. Our study suggests that DArT offers advantages over current technologies in terms of cost and speed of marker discovery and analysis. It can therefore be used to genotype large germplasm collections.

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00122-005-1937-4

Communicated by H.C. Becker

L. Xia · K. Peng · S. Yang · P. Wenzl · A. Kilian (✉)
DArT P/L, GPO Box 3200, Canberra,
ACT 2601, Australia
E-mail: a.kilian@DiversityArrays.com
Tel.: +61-2-62464519
Fax: +61-2-62464501

M. Carmen de Vicente
Office for the Americas,
International Plant Genetic Resources Institute (IPGRI) ,
c/o CIAT, AA 6713 Cali, Colombia

M. Fregene
Centro Internacional de Agricultura Tropical
(CIAT), AA 6713 Cali, Colombia

A. Kilian
Center for the Application of Molecular Biology to
International Agriculture (CAMBIA), GPO Box 3200,
Canberra, ACT 2601, Australia

Present address: K. Peng
John Curtin School of Medical Research,
Australian National University, Canberra,
ACT 0200, Australia

Introduction

Ex situ conservation of plant genetic resources ensures the supply of raw material for plant breeding and hence the continuous availability of food. To date, approximately 1,400 gene banks exist worldwide. They hold more than one million accessions. The efficient management and utilization of these germplasm collections critically depends on our understanding of the existing genetic diversity and its distribution within and among individuals, populations, species and gene pools. In the past, analysis of genetic diversity usually has been performed by evaluating morphological data and agronomic performance. More recently, the use of genetic and molecular markers has accelerated this process and provides more extensive and reliable data.

Cassava (*Manihot esculenta* Crantz) is the largest per-capita source of calories in Sub-Saharan Africa and a major food, feed, and industrial crop in Latin America and South East Asia (FAOSTAT 2001). Half of the world's cassava production of 180 million tons in 2001

was produced in Africa. Because of its excellent adaptability to erratic rainfalls and low-fertility soils, it is a staple and famine-reserve crop as well as a cash source for resource-poor farmers.

There are several germplasm collections with hundreds to thousands of accessions in national programs of Brazil, Thailand, Nigeria, Mozambique and Tanzania. In addition, two International Agricultural Research Centers, the Centro Internacional de Agricultura Tropical (CIAT) in Colombia and the International Institute for Tropical Agriculture (IITA) in Nigeria, hold large germplasm collections. The CIAT collection comprises more than 5,000 accessions from 23 countries and has been evaluated for pest and disease resistance and for novel starch quality traits (Bonierbale et al. 1995; Belloti and Arias 2001; CIAT 2002). Although a core collection has been assembled from the CIAT collection based upon morphological traits and origin, the systematic exploitation and targeted conservation of the main collection remain yet to be tackled.

Several attempts have been made to characterize the genetic diversity in cassava collections to identify useful variability for breeding and to eliminate duplicates by using biochemical, isozyme, and molecular markers such as random amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP) and simple sequence repeat (SSR) (Ocampo et al. 1992; Bonierbale et al. 1995; Second et al. 1997; Chavarriaga-Aguirre et al. 1998, 1999; Fregene et al. 2000). More recently, an SSR diversity study of samples from Africa and the Neotropics has provided insight into the organization of diversity in the primary and secondary centers of diversity (CIAT 2002, 2003; Fregene et al. 2003). The cost of available marker technologies, however, remains an obstacle to the exploitation of the germplasm collections.

Microarray-based genotyping technologies offer the highest throughput available to date. Diversity Arrays Technology (DArT) is sequence-independent, yet enables the simultaneous typing of several hundred polymorphic loci spread over the genome (Jaccoud et al. 2001; Wenzl et al. 2004). With a proper setup and software, hundreds to thousands of individual samples could be processed in a cost-effective and speedy manner.

Here we report on the development of DArT for cassava. The objective of this pilot study was: (1) to test different complexity reduction methods used for DArT fingerprinting; (2) to quantify the quality of data produced by DArT; (3) to validate the genetic relationships revealed by DArT.

Materials and methods

Source of DNA

The cassava accessions used to prepare DArT libraries represent the genetic diversity present in the cultivated species (*Manihot esculenta*), a wild progenitor (*M. esculenta* subsp. *flabellifolia*) and three more distantly related wild species (*M. walkerae*, *M. carthagenensis* and *M. filamentosa*) (Supplementary Table 1).

Leaves were dried at 65°C for 24 h and stored in a desiccator until DNA was extracted according to Dellaporta et al. (1983). The DNA was further purified by phenol/chloroform and chloroform extraction, followed by ethanol precipitation. It was dissolved in 0.1× TE and adjusted to 50 ng l^{-1}. DNA samples were shipped from CIAT, Colombia, to Australia for array development and genotyping, which resulted in partial DNA degradation for a number of samples. We excluded the samples with the most significant degradation from genotyping, but we included nearly all samples in array development, since library creation is not significantly affected by DNA degradation in a subset of samples.

Microarray preparation

DArT arrays were produced from individualized clones of libraries prepared from *Pst*I-based genomic representations as defined in Jaccoud et al. (2001). The representations were generated by cutting 100 ng of mixtures of DNA samples with 2 U *Pst*I and a frequent cutter (*Apo*I, *Bst*NI or *Taq*I) (NEB) in a buffer containing 10 m*M* Tris-OAc, 50 m*M* KOAc, 10 m*M* $Mg(OAc)_2$ and 5 m*M* DTT. A *Pst*I adapter (5′-GTT CAG TCA AGT TAG ATG GTG CA-3′ annealed with 5′-CCA TCT AAC TTG ACT G-3′) was simultaneously ligated with T4 DNA ligase (NEB). The ligation reaction was diluted 20-fold, and a 1-µl aliquot was used as a template in 50-µl amplification reactions using a primer complementary to the adapter sequence (5′-CAG TCA AGT TAG ATG GTG CAG-3′). The annealing temperature was 58°C. Other cycling conditions were as in Jaccoud et al. (2001).

Libraries were prepared from amplified fragments essentially according to Jaccoud et al. (2001). Individual clones were grown in 384-well plates containing LB medium supplemented with 100 mg l^{-1} ampicillin and a "freezing mix" (D. Kudrna, personal communication). Aliquots of the cultures were used as templates to amplify inserts according to Jaccoud et al. (2001).

Amplification reactions were dried, dissolved in 50% DMSO and spotted in triplicate on polylysine-coated slides [Menzel or CEL Associates, Los Angeles (Calif.)] using either a GMS 417 arrayer (Affymetrix) or a MicroGrid II arrayer (Genomics Solutions; Lincoln, Neb.). After printing, slides were heated to 80°C for 2 h, incubated in hot water (95°C) for 2 min and dried by centrifugation.

Each clone (marker) was given a preliminary name, which will be revised in the future using a more generally applicable naming system. For the markers listed in Supplementary Tables 2 and 3, marker names contain information on array type (A, B and T, for *Pst*I/*Apo*I,

*Pst*I/*Bst*NI and *Pst*I/*Taq*I, respectively) and plate location (plate number + well position).

Genotyping of DNA samples

Genomic representations were generated from 20 cassava varieties and wild accessions using the same complexity reduction method used for library construction (*Pst*I/*Apo*I, *Pst*I/*Bst*NI or *Pst*I/*Taq*I). They were concentrated tenfold by precipitation with one volume of isopropanol, denatured and labeled with 1 µl 500 µ*M* cy3-labeled random decamers using the exo$^-$ Klenow fragment of *E. coli* DNA polymerase I (NEB). Labeled representations, called "targets", were added to 50 µl of a 50:5:1 mixture of ExpressHyb buffer (Clonetech, Palo Alto, Calif.) and 10 g l^{-1} herring sperm DNA, and the cy5-labeled polylinker fragment of the plasmid used for library preparation (Jaccoud et al. 2001). After denaturing, labeled targets were hybridized to microarrays overnight at 65°C. Slides were washed according to Jaccoud et al. (2001) and scanned with an Affymetrix 48 confocal laser scanner.

Automatic image analysis and polymorphism scoring

DARTSOFT, a software package developed in-house, was used to automatically analyze each batch of TIF image pairs generated in an experiment, typically comprising a few dozens of slides. Two versions of the software were used in this report. The earlier version of DARTSOFT, exclusively based on a variance partitioning algorithm (ANOVA), was used in the initial tests of complexity reduction methods. A more advanced version was used in the validation experiments with full-sized arrays. The improved software version localized spots, rejected those with a weak reference signals (if specified by the user), computed and normalized relative hybridization intensities (= log[cy3target/cy5reference]), calculated the median value for replicate spots, identified polymorphic clones by using a combination of ANOVA and fuzzy C-means clustering at a fuzziness level of 1.5 (Bezdek 1981), and classified polymorphic clones as being present (1) or absent (0) in the representation hybridized to a slide. The clustering algorithm also provided a probability estimate for each individual genotype call (Cayla et al. in preparation).

Production and validation of a full-size array

The *Pst*I/*Taq*I library was expanded by 3,072 clones to produce a first-generation genotyping array. This array was used to fingerprint 38 cassava varieties and wild accessions. Polymorphic clones were identified by analyzing hybridization patterns with DARTSOFT by variance-partitioning (Cayla et al. in preparation). The DARTSOFT-generated 0 and 1 scores were used as input for the RESTDIST and NEIGHBOR programs of the PHYLIP V3.6 software package to build a UPGMA dendrogram based on the Nei/Li restriction fragment distance (Nei and Li 1979). Clade strength was tested by 1,000 bootstrap analyses performed with the SEQBOOT program (Felsenstein 1989).

To test reproducibility of scoring, we typed a subset of the samples a second time. The experiment was analyzed using settings previously validated in barley, with the rejection of weak spots and scoring by clustering (Wenzl et al. 2004). Thirty-one wild accessions were analyzed using the *Pst*I/*Taq*I array expanded to 4,608 clones, and the same DArTsoft settings were applied as for the 3,072-clone array.

Results and discussion

Testing of complexity reduction methods

A critical step in DArT is the genome complexity reduction to prepare genomic representations. Our work with several other plant genomes has taught us that digestion with the *Pst*I restriction enzyme (RE) together with a more frequently cutting RE, combined with adapter ligation-based amplification of intact *Pst*I fragments, is an efficient method. Therefore, we started by evaluating several combinations of *Pst*I and a frequent cutting RE (*Bst*NI, *Apo*I and *Taq*I) using a mixture of genomic DNA from 22 cassava genotypes (Supplementary Table 1).

Prior to cloning the genomic representations, we evaluated their complexity by amplification with primers containing two or three selective bases (extending beyond the *Pst*I site) at the 3′ end (Table 1). This approach resembles the AFLP technique, but it was used here not to identify DNA sequence variation, but to estimate the approximate number of *Pst*I fragments in each representation. The number of fragments amplified with various primers suggested that the approximate complexity of the representations tested was between 20,000 and 30,000 unique fragments (data not presented). Our experience with other genomes, for which we carried the complexity estimation using more precise methods, suggested that the AFLP-like method overestimates the true complexity three- to fourfold (Wenzl et al. unpublished observations).

We also estimated the number of fragments in the representations in silico. We used all of the cassava sequence data available at Genebank (1.7 Mbp) to estimate the cutting frequencies of *Pst*I, *Taq*I, *Bst*NI and *Apo*I and to predict the number of intact *Pst*I fragments in the whole genome based on a genome size of 760 Mbp (Awoleye et al. 1994). The estimates obtained for the *Pst*I/*Taq*I and *Pst*I/*Bst*NI representations were greater than 16,000 (Table 1). In silico calculations, however, do not account for the methylation of *Pst*I sites, which is expected to reduce the number of fragments several-fold (Wenzl et al. 2004). Even though the precise correction

Table 1 Polymorphic clones identified in libraries prepared for protocol optimization

Library	DNA samples		Estimated no. of fragments in representations[a]	DArT clones	
	Varieties	Wild accessions		Total number	Polymorphism frequency (%)[b]
*Pst*I/*Apo*I	13	7	8,600	768	9.0
*Pst*I/*Bst*NI	13	7	16,900	768	17.2
*Pst*I/*Taq*I	13	7	16,600	768	14.6

[a] In silico estimates based on randomly distributed recognition sites and empirical cutting frequencies measured for a 1.7-Mbp sample of genomic sequence of *Manihot esculenta*

[b] Polymorphic clones were identified with DArTsoft using the variance-partitioning algorithm for scoring (see Materials and methods)

factors for both methods of complexity estimation are not known, we can assume from these two estimates that the *Pst*I/*Taq*I and *Pst*I/*Bst*NI libraries contain at least 5,000–7,000 unique fragments. These numbers were deemed sufficient to progress to cloning the libraries from the three complexity reduction methods (*Pst*I/*Bst*NI, *Pst*I/*Apo*I, and *Pst*I/*Taq*I).

Genomic *Pst*I fragments lacking the recognition site of the frequently cutting RE (*Bst*NI, *Apo*I or *Taq*I) were ligated into a plasmid vector and individualized by transformation into *E. coli*. The amplified inserts were micro-arrayed. Seven hundred sixty clones from each library were assayed by using matching genomic representations prepared separately from each of a group of 20 genotypes. Polymorphic clones were identified and scored using the variance-partitioning algorithm (see Materials and methods).

There were substantial differences in the percentage of polymorphic clones revealed by each of the three complexity reduction methods tested (Table 1). The highest percentage of clones showing a binary distribution of relative fluorescence intensity among accessions was observed for the *Pst*I/*Bst*NI method (132 clones), followed by the *Pst*/*Taq*I (112) and the *Pst*/*Apo*I (69) method, with a total of 313 clones from all representations. We excluded 17 clones for which we obtained incomplete scoring across the 20 genotypes analyzed and used the scores for remaining 296 polymorphic clones to create a UPGMA dendrogram based on the Nei/Li restriction fragment distance (Supplementary Table 2; Fig. 1a). The dendrogram showed the expected separation of cultivated cassava accessions and their progenitor species (*M. esculenta* subsp. *flabellifolia*) from their wild relatives (*M. carthagenensis* and *M. walkerae*). The relationship between the provenience of cultivated accessions and their proximity on the dendrogram was also apparent.

The average polymorphism information content (PIC) values of the DArT markers identified calculated according to Anderson et al. (1993) was 0.42 (SD = 0.09), and the median was 0.46 (Fig. 1b). These values are very high for randomly chosen bi-allelic loci, but only marginally higher than those observed for barley DArT markers developed using similar complexity reduction methods (Wenzl et al. 2004). However, the PIC values are lower than for SSR markers where they range from 0.5 to 0.7 (Fregene et al. 2003). The lower PIC values could be advantageous because they could lead to a more defined genetic structure of accessions from distant geographic regions, information difficult to obtain with SSR markers.

Development of a full-size array

With 15–17% of polymorphic clones and a predicted number of unique fragments between 5,000 and 7,000, we selected the *Pst*I/*Bst*NI and *Pst*I/*Taq*I representations to develop an array containing approximately 1,000 polymorphic clones. To achieve this goal we expanded the two libraries by an additional 3,072 random clones.

Library expansion was carried out using DNA samples from the accessions listed in Supplementary Table 1. Approximately 50% (38) of the samples used for library expansion were also used for the identification of polymorphic clones using a similar algorithm as in the complexity reduction testing stage. In the *Pst*I/*Taq*I array we identified 440 polymorphic clones (14.3%), a rate very consistent with the results obtained with the initial smaller array (14.6%). The list

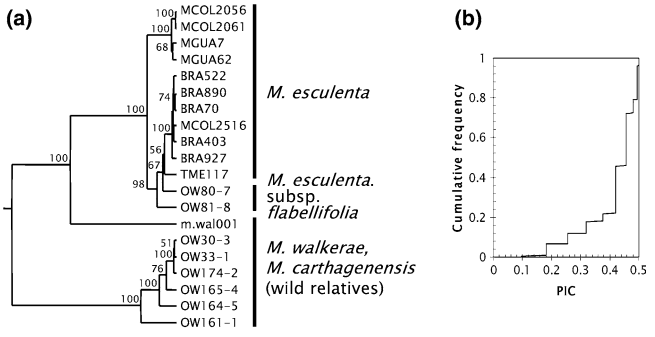

Fig. 1 Genetic relationships among a group of *Manihot* accessions. **a** UPGMA dendrogram constructed from 296 *Pst*I/*Apo*I, *Pst*I/*Bst*NI and *Pst*I/*Taq*I markers based on the Nei/Li restriction fragment distance matrix (Nei and Li 1979; Felsenstein 1989). Bootstrap support values (1,000 replicates) are shown if greater than 50%. **b** Cumulative distribution function of the polymorphism information content (PIC) values of the DArT markers used to construct the dendrogram (Anderson et al. 1993)

of polymorphic clones and their scores is presented in Supplementary Table 3.

The marker scores generated by DARTSOFT were used to construct a UPGMA dendrogram based on the Nei/Li restriction fragment distance (Fig. 2). The separation between the *M. esculenta* accessions and the wild accessions was clear and consistent with the dendrogram in Fig. 1. The separation of the two *M. esculenta* subsp. *flabellifolia* accessions from the cultivated samples was significant in this dataset (100% bootstrap support), but not in the dendrogram in Fig. 1. This difference may be due to a 50% larger number of markers in the expanded *Pst*I/*Taq*I array compared to the combined data from the initial *Pst*I/*Bst*NI, *Pst*I/*Apo*I and *Pst*I/*Taq*I arrays. We note that the variety MCOL1505 and its progenitor, CM6740-7, form a single cluster, similarly to that of the half-sibs NGA19 and NGA2.

Interestingly, the average PIC value calculated from the *Pst*I/*Taq*I array validation experiment (0.27 ± 0.14) was approximately 35% lower than the average value obtained with the discovery arrays (Fig. 2c). It is unlikely that the drop in PIC values was a result of using a single complexity reduction method (*Pst*I/*Taq*I) for the array validation experiment, because the *Pst*I/*Taq*I clones in the initial experiments had an average PIC value that was almost identical to average of all three representations (0.43 and 0.42, respectively). Instead, the lower PIC values can be attributed to the different set of accessions used in the analysis. In our initial tests of arrays with the three complexity reduction methods (*Pst*I/*Bst*NI, *Pst*I/*Apo*I, and *Pst*I/*Taq*I), 35% of the analyzed samples were wild relatives of cassava (*M. carthagenensis* and *M. walkerae*). By contrast, in the array validation experiment only 7.9% of the samples represented wild relatives. We therefore assayed significantly more genetic diversity in the first experiments, which is expected to translate into higher PIC values.

In the *Pst*I/*Bst*NI array, we identified 554 polymorphic clones (18.0%), a rate very consistent with the 17.2% polymorphism frequency in the smaller *Pst*I/*Bst*NI array in the first project phase (data not presented).

Evaluation of the performance of a full-size array

In order to evaluate the performance of the expanded arrays we analyzed the genetic diversity of a subset of accessions used for array development. We used the *Pst*I/*Taq*I array to hybridize 62 targets derived from 38 cassava accessions. For 24 accessions we used duplicate targets and microarrays, and 14 accessions were analyzed on a single microarray. An improved algorithm was applied to identify polymorphic clones and score individual samples (Cayla et al. in preparation). The main improvements over the previous algorithm included: (1) elimination of array features containing little DNA (as determined by the reference DNA hybridization signal); (2) identification of polymorphic clones based on two complementary approaches (ANOVA and "fuzzy clustering"); (3) filtering of individual scores based on the probability of class membership (0 vs. 1), as determined by the clustering algorithm.

We tested several settings of membership probability to identify the optimum calling rate and scoring

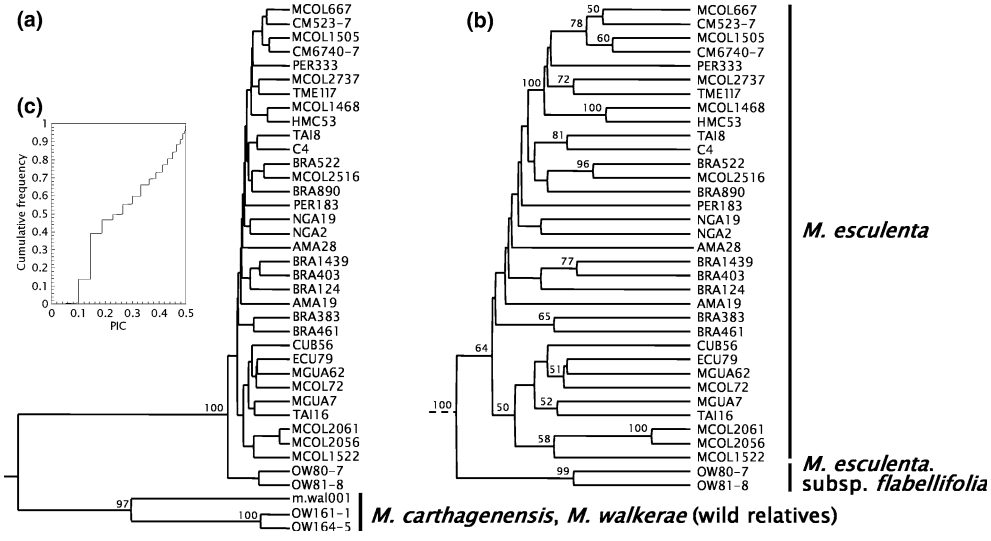

Fig. 2 Genetic relationships among a group of *Manihot* accessions. **a** UPGMA dendrogram constructed from 440 *Pst*I/*Taq*I markers based on the Nei/Li restriction fragment distance matrix (Nei and Li 1979; Felsenstein 1989). **b** Section of the dendrogram containing the cultivated species (*M. esculenta*) and its hypothesized ancestor (*M. esculenta* spp. *flabellifolia*) drawn to a different scale. Bootstrap support values (1,000 replicates) are shown if greater than 50%. **c** Cumulative distribution function of the PIC values of the DArT markers used to construct the dendrogram (Anderson et al. 1993)

Table 2 Properties of DArT markers of different quality

	100 > P[a] > 90	90 > P > 85	85 > P > 80
Number of markers	295	231	208
Call rate	99.3 ± 1.6	97.1 ± 2.8	94.3 ± 3.8
PIC[b]	0.37 ± 0.06	0.37 ± 0.08	0.37 ± 0.10

[a] P, The percentage of variance between the clusters determined by the fuzzy C-means clustering. The P parameter provides a measure of DArT marker quality
[b] PIC, Polymorphism information content (average ± standard deviation)

reproducibility. A membership probability threshold of 0.7 resulted in a calling rate of 98.85% and a scoring reproducibility of 99.75%. Raising the probability threshold to 0.9 increased the scoring reproducibility to 99.98% but reduced the calling rate to 95.85%. We decided to use a membership probability threshold of 0.8, which resulted in a high calling rate (98.14) and a good scoring reproducibility (99.84; 10 miss-scores out of 6,379 comparisons). Interestingly, these numbers were nearly identical to those obtained in experiments with barley DArT arrays (Wenzl et al. 2004). Application of these stringent quality criteria identified 274 high-quality clones that were polymorphic among the accessions tested (data not presented).

To further evaluate the performance of the full-size *Pst*I/*Taq*I array (in this experiment containing 4,608 clones) we hybridized the targets prepared from wild accessions only. We used 31 samples representing four species of Manihot (*M. esculenta, M. walkerae, M. carthagenensis* and *M. filamentosa*) and a number of hybrids involving *M. esculenta* and *M. tristis* (Supplementary Table 1, Supplementary Fig. 1). We identified 734 markers polymorphic among these accessions using the same DARTSOFT settings as before (membership probability threshold of 0.8). We analyzed the relation between the quality of markers (measured by parameter'' P'', which represents the percentage of total variance which existed between the two clusters) and the two important performance indicators: the PIC value and the call rate (Table 2). As expected, PIC values were independent of marker quality (0.37 for all three classes of markers), while the average call rate was decreasing with the average P value. The markers with highest P values (above 90% of total variance between the clusters) had very high average call rate (99.3%), while the markers with 85 < P > 80 had above 5% of missing scores.

We note that both the number of polymorphic clones (734) and the PIC values (0.37) were higher than in the full-size array experiments involving primarily cultivated materials. This observation does support our suggestion that reduced level of genetic diversity in the sample used for initial validation of the expanded array was a primary cause of the lower average PIC value, as compared to the protocol optimization arrays (above). Importantly, the dendrogram created using the scoring table generated by DARTSOFT from this experiment as the input separates with high confidence all accessions into groups corresponding to the four species analyzed (Supplementary Fig. 1). All the hybrids are grouping together with the accessions of *M. esculenta* subsp. *flabellifolia* and completely separate from the accessions of the remaining three species.

We also tested the *Pst*I/*Bst*NI array with 36 cassava accessions and 45 slides (nine accessions analyzed in duplicate). Using identical DARTSOFT parameters as for the *Pst*I/*Taq*I array (membership probability threshold = 0.8), we identified 299 polymorphic clones with a calling rate of 98.00% and a scoring reproducibility of 99.81%. Detailed results of this analysis will be presented in an upcoming manuscript that will describe the outcome of a comprehensive analysis of a large cassava germplasm collection.

Concluding remarks

A cassava DArT genotyping array was developed containing approximately 1,000 polymorphic clones, which shows the DArT can be an effective tool for exhaustive fingerprinting of germplasm collections. The libraries of clones reported in this paper will be freely available to the research community, thereby offering an alternative approach to cassava genotyping. Compared to other technologies, DArT markers can be developed and typed quickly and cheaply. At the same time, the data quality (measured by the call rate and scoring reproducibility) is at least as good as with other technologies. Data quality benefits from the fact the DArT is amenable to automation, and whole data extraction is done automatically using dedicated software. Our work on other genomes has shown the DArT markers can be used effectively for medium-density mapping, thus enabling applications for quantitative trait locus identification, genome background screening, whole-genome marker-assisted selection or accelerated introgression of selected genomic regions. We are preparing to apply this technology to association mapping in cassava through the combined use of several different complexity reduction methods. We propose the DArT may effectively complement existing technologies in breeding and genomics, especially for crops with limited resources, as in our calculations the cost of DArT datapoint is approximately tenfold lower than the cost of SSR datapoint.

Acknowledgements We thank our colleagues at DArT P/L and CAMBIA for helpful discussions and colleagues in CIAT for DNA extraction. Special thanks are extended to Cyril Cayla and Grzegorz Uszynski for help with data analysis and to Eric Huttner for his help with drafting the manuscript. Ling Xia and Kaiman Peng contributed equally to the manuscript.

References

Anderson JA, Churchill GA, Autrique JE, Tanksley SD, Sorrells ME (1993) Optimizing parental selection for genetic linkage maps. Genome 36:181–186

Awoleye F, van Duren M, Dolezel J, Novak FJ (1994) Nuclear DNA content and in vitro induced somatic polyploidization cassava (*Manihot esculenta* Crantz) breeding. Euphytica 76:195–20

Bellotti AC, Arias B (2001) Host plant resistance to whiteflies with emphasis on cassava as a case study. Crop Prot 20:813–823

Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York

Bonierbale MW, Maya MM, Claros JL, Iglesias C (1995) Application of molecular markers to describing the genetic structure of cassava gene pools. In: The cassava biotechnology network: Proc 2nd Int Sci Meet. Working document no. 50, Centro International de Agricultura Tropical, 2v, Cali, Colombia

Chavarriaga-Aguirre P, Maya MM, Bonierbale MW, Kresovich S, Fregene MA, Tohme J, Kochert G (1998) Microsatellites in cassava (*Manihot esculenta* Crantz): discovery, inheritance and variability. Theor Appl Genet 97:493–501

Chavarriaga-Aguirre P, Maya MM, Tohme J, Duque MC, Iglesias C, Bonierbale MW, Kresovich S, Kochert G (1999) Using microsatellites, isozymes and AFLPs to evaluate genetic diversity and redundancy in the cassava core collection and to assess the usefulness of DNA-based markers to maintain germplasm collections. Mol Breed 5:263–273

CIAT (2002) Assessing and utilizing agrobiodiversity through biotechnology: Annual Report: Project SB-02. Centro Internacional de Agricultura Tropical, Cali, Colombia

CIAT (2003) Assessing and utilizing agrobiodiversity through biotechnology: Annual Report: Project SB-02. Centro Internacional de Agricultura Tropical, Cali, Colombia

Dellaporta SL, Woods J, Hicks JR (1983) A plant DNA minipreparation: version II. Plant Mol Biol Rep 1:19–21

FAOSTAT (2001) FAO statistical databases. Rome, Italy

Felsenstein, J (1989) PHYLIP—phylogeny inference package. Cladistics 5:164–166

Fregene MA, Bernal A, Dixon A, Roca W, Tohme J (2000) AFLP analysis of African cassava (*Manihot esculenta* Crantz) germplasm resistant to the cassava mosaic disease (CMD). Theor Appl Genet 100:678–685

Fregene M, Suarez M, Mkumbira J, Kulembeka H, Ndedya E, Kulaya A, Mitchel S, Gullberg U, Rosling H, Dixon A, Kresovich S (2003) Simple sequence repeat (SSR) diversity of cassava (*Manihot esculenta* Crantz) landraces: genetic diversity and differentiation in a predominantly asexually propagated crop. Theor Appl Genet 107:1083–1093

Jaccoud D, Peng K, Feinstein D, Kilian A (2001) Diversity arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Res 29:e25

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:5269–5273

Ocampo C, Hershey C, Iglesias C, Iwanaga M (1992) Esterase isozyme fingerprinting of the cassava germplasm collection held at CIAT. In: Roca W, Thro AM (eds) Proc 1st Int Sci Meet Cassava Biotechnol Network CIAT. Cali, Colombia, pp 81–89

Second G, Allem A, Emperaire L, Ingram C, Colombo C, Mendes R, Carvalho L (1997) AFLP based *Manihot* and cassava numerical taxonomy and genetic structure analysis in progress: implications for dynamic conservation and genetic mapping. Afr J Root Tuber Crops 2:140–147

Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinhofs A, Kilian A (2004) diversity arrays technology (DArT) for whole-genome profiling of barley. Proc Natl Acad Sci USA 101:9915–9920