Theor Appl Genet (2012) 124:685–695
DOI 10.1007/s00122-011-1739-9

ORIGINAL PAPER

Identification, validation and high-throughput genotyping of transcribed gene SNPs in cassava

Morag E. Ferguson · Sarah J. Hearne · Timothy J. Close · Steve Wanamaker · William A. Moskal · Christopher D. Town · Joe de Young · Pradeep Reddy Marri · Ismail Yusuf Rabbi · Etienne P. de Villiers

Received: 5 May 2011 / Accepted: 18 October 2011 / Published online: 9 November 2011

Abstract The availability of genomic resources can facilitate progress in plant breeding through the application of advanced molecular technologies for crop improvement. This is particularly important in the case of less researched crops such as cassava, a staple and food security crop for more than 800 million people. Here, expressed sequence tags (ESTs) were generated from five drought stressed and well-watered cassava varieties. Two cDNA libraries were developed: one from root tissue (CASR), the other from leaf, stem and stem meristem tissue (CASL). Sequencing generated 706 contigs and 3,430 singletons. These sequences were combined with those from two other EST sequencing initiatives and filtered based on the sequence quality. Quality sequences were aligned using CAP3 and embedded in a Windows browser called HarvEST:Cassava which is made available. HarvEST:Cassava consists of a Unigene set of 22,903 quality sequences. A total of 2,954 putative SNPs were identified. Of these 1,536 SNPs from 1,170 contigs and 53 cassava genotypes were selected for SNP validation using Illumina's GoldenGate assay. As a result 1,190 SNPs were validated technically and biologically. The location of validated SNPs on scaffolds of the cassava genome sequence (v.4.1) is provided. A diversity assessment of 53 cassava varieties reveals some sub-structure based on the geographical origin, greater diversity in the Americas as opposed to Africa, and similar levels of diversity in West Africa and southern, eastern and central Africa. The resources presented allow for improved genetic dissection of economically important traits and the application of modern genomics-based approaches to cassava breeding and conservation.

Communicated by G. Bryan.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-011-1739-9) contains supplementary material, which is available to authorized users.

M. E. Ferguson (✉) · S. J. Hearne · I. Y. Rabbi
International Institute of Tropical Agriculture (IITA),
c/o ILRI, P.O. Box 30709, Nairobi, Kenya
e-mail: mferguson@cgiar.org

Present Address:
S. J. Hearne
International Maize and Wheat Improvement Center
(CIMMYT), Apdo. Postal 6-641, 06600 Mexico, D.F., Mexico

T. J. Close · S. Wanamaker
Department of Botany and Plant Sciences,
University of California, Riverside, CA 92521-0124, USA

W. A. Moskal · C. D. Town
The J. Craig Venter Institute, 9704 Medical Center Drive,
Rockville, MD 20850, USA

Background

Cassava, *Manihot esculenta* Crantz. ($2n = 36$) is a highly adaptable starchy root crop and the primary staple food for

J. de Young
The Southern California Genotyping Consortium,
University of California Los Angeles, 695 Charles E. Young
Dr. South, Los Angeles, CA 90095, USA

P. R. Marri
University of Arizona, B105 Institute,
1657 E Helen St., Tucson, AZ 85719, USA

Present Address:
I. Y. Rabbi
International Institute of Tropical Agriculture (IITA),
Oyo Road, Ibadan, Nigeria

E. P. de Villiers
International Livestock Research Institute (ILRI),
P.O. Box 30709, Nairobi 00100, Kenya

more than 800 million people, largely in sub-Saharan Africa (Lebot 2009). Apart from being a staple food it is also a source of cash income from fresh and processed food, the production of starch-based products, biofuels and animal feed (Dixon et al. 2003; Sriroth et al. 2000; Tonujari 2004). Cultivated in tropical and sub-tropical regions of Asia, Latin America and Africa production reached 233 million tons in 2008 (FAO 2008) with over 50% of this being in Africa. Cassava is often grown in marginal environments with erratic rainfall, poor soils and under low intensity management (El Sharkawy 2004). The difference between the potential yield and the average farmer's yield is more than sixfold, indicating tremendous scope for yield improvement (Lebot 2009).

National and international cassava breeding efforts have made significant impact on cassava production both in terms of disease tolerance, yield and quality improvements. Breeding for these diversified uses under adverse climatic conditions is a challenge, particularly as cassava is a highly heterogeneous and heterozygous vegetatively propagated crop with variable flowering, low seed set, and a long breeding cycle (Jennings and Iglesias 2001). It has been recognised that the application of advanced technologies could substantially increase the efficiency and success of orphan crop breeding programs (Nelson et al. 2004). This does however rely on the availability of genomic tools, including molecular markers.

In recent years, the availability of genomic resources for cassava has increased substantially, most notable through the sequencing of the cassava genome (http://www.phytozome.net/cassava). cDNAs (generally expressed sequence tags; ESTs) can assist with gene discovery, the study and characterisation of plant expressed genes and the isolation of nucleotide sequences of genes with known function (Lopez et al. 2004; Luo et al. 2005). Anderson et al. (2004) provides an overview of EST resources available for cassava and other Euphorbiaceae species. Since then additional EST resources have emerged for cassava (Lokko et al. 2007; Sakurai et al. 2007), although the Unigene set is still limited compared to some other crops such as for maize and the model plant *Arabidopsis* in which over 2 million and 1.5 million ESTs are available, respectively (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html).

The availability of molecular markers is critical if molecular tools are to be applied to cassava breeding. In cassava, the majority of markers that are widely used are simple sequence repeat (SSR) loci (Chavarriaga-Aguirre et al. 1998; Kunkeaw et al. 2010; Mba et al. 2001; Raji et al. 2009; Sraphet et al. 2011; Tangphatsornruang et al. 2008), although Diversity Array Technology (DArT) markers are available (Xia et al. 2005). A high density of single nucleotide polymorphic (SNP) markers would dramatically facilitate progress in cassava genomics and breeding. SNPs and small insertions and deletions (indels) represent the most frequent form of naturally occurring genetic variation in populations (Cho et al. 1999). SNPs are generally biallelic due to their low mutation rate and evolutionary stability rendering them less informative than multi-allelic SSRs (Syvänen 2001). This drawback, however, is offset by their suitability to ultra-high throughput genotyping techniques (Appleby et al. 2009) and their sheer abundance, thus making them the marker of choice for dense genotyping (Rafalski 2002). In addition, the utilisation of multi-SNP haplotypes can offset the relatively low information content of single SNP loci (Brumfield et al. 2003).

Sakurai et al. (2007) reported the identification, but no details, of 2,356 putative SNPs; through the University of Maryland's Cassava Genome Database 384 and 371, putative SNP containing sequences derived from genes and the cassava physical map, respectively, were made available. Apart from these studies the analysis of DNA sequence variation in cassava has been mainly confined to single genes or DNA fragments with the goal of defining gene structure, function or evolutionary relationships. These studies include Kawuki et al. (2009) who studied sequence diversity in nine genes involved in cyanogenesis, starch metabolism, stress and/or defense related pathways and identified 26 SNPs; Lopez et al. (2005) exploited ESTs to detect SNPs in five cultivars of cassava.

The large amount of plant sequence data available in public databases represents a rich resource for SNP discovery using bioinformatics approaches (Chao et al. 2009). Several methods for identifying SNP markers from EST sequence databases have been reported (Picoult-Newberg et al. 1999). The use of EST databases has the advantage of possibly identifying SNPs associated with changes in phenotype. EST databases have been mined for large-scale SNP discovery in several species including humans (Garg et al. 1999), *Arabidopsis* (Schmid et al. 2003), maize (Batley et al. 2003) and sugarcane (Grivet et al. 2002).

The objectives of this study were to (1) expand the available ESTs in relation to drought response in cassava, (2) consolidate quality cassava sequences into an EST database, (3) identify a substantial number of putative SNPs from aligned ESTs, (4) validate putative SNPs in a high-throughput genotyping platform in cultivated cassava and (5) use gene-based SNPs to elucidate genetic relationships and diversity of cassava varieties on a regional basis.

Methods

Normalised cDNA library preparation, EST sequencing and characterisation

Twelve plants of four cassava farmer varieties from Malawi ('Sauti', 'Gomani', 'Mbundumali' and 'Mkondezi') and

TME 1 from the International Institute of Tropical Agriculture (IITA) were grown in pots from stem cuttings. These varieties show varying responses to drought including typical susceptibility, rapid leaf loss and stay-green. Plants were grown in a screenhouse at Chitedze Research Station, Malawi, under normal light and temperature regimes, in a sandy-loam soil. No additional fertilizer was provided. Pots were randomly split into two treatments of six plants. Ten weeks post planting one treatment received limited irrigation to induce drought stress. Once significant wilting was observed, 3 weeks post the initiation of drought stress, root, leaf, stem and stem meristem tissues were removed from each variety under each treatment and frozen in liquid nitrogen before being stored in a −80°C freezer. Samples of (1) root, (2) leaf and (3) stem/stem meristem tissue from drought stressed and non-stressed plants were pooled for all cultivars on a weight basis. RNA was extracted from each of these three pools using Concert™ Plant RNA Reagent (Invitrogen™ by Life Technologies™) following manufacturer's protocol. After extraction the leaf and stem meristem RNA was pooled on a 1:1 quantity basis. Evrogen constructed two normalised cDNA libraries from these samples; a leaf/stem meristem library (CASL) and a root library designated (CASR). Both libraries had titers of approximately one million clones. Sequencing was performed at The Institute of Genomic Research (TIGR).

Data analysis

Trace files were assessed for quality using Phred and Cross match. Poor quality sequences were discarded. To assign a putative function to the ESTs, specific Uniprot identifiers were assigned to cassava Unigene sequences using Uniprot/Arabidopsis database. GO annotations were retrieved using the GORetriever tool from AgBase (http://www.agbase.msstate.edu) (McCarthy et al. 2006). A summary of GO terms for each of the three main categories (biological process, molecular function and cellular component) was obtained using a plant specific GO Slim ontology using GOSlimViewer in AgBase (McCarthy et al. 2006).

HarvEST database and putative SNP identification

Trace files of EST sequences from five cDNA libraries were used to compile a database of high quality sequences. These included 18,633 EST sequences from CV01 and CV02 (Anderson et al. 2004; Lokko et al. 2007), 34,955 EST sequences (Sakurai et al. 2007) and 5,019 EST sequences from CASL and CASR (this article). The full-length cDNA library from Sakurai et al. (2007) was derived from one cassava variety from Thailand (MTAI16) under normal, heat, drought, high aluminium/low pH and post-harvest physiological deterioration conditions. cDNA libraries from Anderson et al. (2004) were derived from dehydration stressed and control well-watered tissues of three varieties of West African origin; TME117, TME3 and TMS30572. Trace files were obtained from EST developers or NCBI Trace DB using a query string species code '*Manihot esculenta*'. These were assessed for quality using Phred and Cross match exactly as in Close et al. (2004). Poor quality sequences were discarded. Remaining sequences were aligned using CAP3 using settings appropriate for SNP detection (Close et al. 2009). The EST assembly was embedded into a database and packaged within the HarvEST software to form a Windows browser called HarvEST:Cassava (version 1.00–1.06). Specific Uniprot identifiers were assigned to cassava Unigene sequences using Uniprot/Arabidopsis database. GO annotations were retrieved and a summary of GO terms was derived as previously described.

SNPs were identified from sequence alignment with CAP3 as described above. Those with a minimum of two supporting ESTs for each allele were classified as 2E, whereas those with a minimum of three supporting ESTs for each allele were classified as 3E. This was continued up to the maximum of 6E. SNPs were named as follows: Me for *M. esculenta* Crantz., MEF for the first author of this article, c. for complementary DNA, followed by a unique number e.g. Me.MEF.c.0545.

SNP validation

For SNP validation purposes 53 cultivated cassava (*M. esculenta* Crantz.) accessions were genotyped using 1,536 putative SNPs using Illumina's GoldenGate assay (Illumina Inc., San Diego, CA) at The Southern California Genotyping Consortium (SCGC), University of California Los Angeles (UCLA) (http://scgc.genetics.ucla.edu) (Table 1). The genotype set included 22 accessions from the Americas, 23 accessions from West Africa, 11 accessions from southern, eastern and central (SEC) Africa and two accessions from Asia. Putative duplicate accessions were also included in the set from the IITA Genebank, to determine the potential of high-density genotyping to identify duplicates in cassava germplasm collections. These accessions were TMe539 and TMe3187, 'Kaleso' and 'Namikonga', TMe589 and TMe3209, and TMe153 and TMe2929. Three accessions from West Africa were of unknown identity.

A set of 1,536 SNPs were selected from all putative SNPs identified. Initially SNPs with an Illumina primer score of 0.6 and above were considered. This provides a quality score for the design of Illumina primers. In

Table 1 Cultivated cassava germplasm used for SNP validation

ID	Other identifier	Country of origin	Region of origin
AR37-80		Unknown	Americas
AR40-6		Unknown	Americas
MCOL 1734		Colombia	Americas
VEN77		Venezuela	Americas
CIAT6	BRA206	Brazil	Americas
CIAT56	COL2459	Colombia	Americas
CIAT80	GUA59	Guatemala	Americas
CIAT819	BRA200	Brazil	Americas
CIAT322	BRA125	Brazil	Americas
CIAT834	BRA436	Brazil	Americas
CIAT857	BRA785	Brazil	Americas
CIAT1226	BRA1016	Brazil	Americas
CIAT1212	BRA842	Brazil	Americas
CIAT880	BRA990	Brazil	Americas
CIAT1303	COL233	Colombia	Americas
CIAT391	ARG12	Argentina	Americas
CIAT567	PAR23	Paraguay	Americas
CIAT543	CR19	Costa Rica	Americas
CIAT694	COL2638	Colombia	Americas
CIAT584	PER458	Peru	Americas
CIAT1135	USA7	USA	Americas
CIAT1370	BRA1001	Brazil	Americas
CIAT759	TAI1	Thailand	Asia
CIAT558	MAL60	Malaysia	Asia
Nachinyaya		Tanzania	SEC Africa
Kiroba		Tanzania	SEC Africa
NDL06/132		Tanzania	SEC Africa
Muzege		Tanzania	SEC Africa
TMe3187	Bao (T1)	Uganda	SEC Africa
TMe3288	Ex Mwachande	Kenya	SEC Africa
TMe539	Bao (T1)	Uganda	SEC Africa
Kaleso		Kenya	SEC Africa
Albert		Tanzania	SEC Africa
Namikonga		Tanzania	SEC Africa
I96/1089A		Unknown	West Africa
Unknown1		Unknown	West Africa
TME7		Unknown	West Africa
I96/1632		Unknown	West Africa
TMS30572		Unknown	West Africa
97/3200		Unknown	West Africa
94/0026		Unknown	West Africa
I92/0326		Unknown	West Africa
TMe5	Bagi Wawa	Nigeria	West Africa
Unknown2		Nigeria	West Africa
TMe153	82/00290	Cameroon	West Africa
TMe3002	TOMA 36	Togo	West Africa
TMe3209	Bassa Girl	Liberia	West Africa
TMe3445	Kolia 3	Guinea Conakry	West Africa
TMe3082	Toma 175	Togo	West Africa
TMe2929	82/00290	Cameroon	West Africa
TMe125	Ikpaki	Nigeria	West Africa
Unknown3		Unknown	West Africa
TMe589	Bassa Girl	Liberia	West Africa

addition, each putative SNP selected had to be at least 60 bp from the nearest SNP. Initially SNPs classified as 3E and higher were selected, then 2E SNPs were selected so that the maximum number of unigenes were represented. SNP olicode primers were designed using the Assay Design Tool (ADT) (Illumina, Inc.).

Raw data were transformed to genotype calls using Illumina's BeadStudio version 3 with the genotyping module. Cassava is highly heterozygous so there was no need to include any 'synthetic heterozygotes' to anchor heterozygote cluster positions to enable the identification of true heterozygotes, as in other highly inbred crops such as barley (Close et al. 2009). The spatial positions of heterozygote and homozygote data clusters were confined to areas of high certainty so that data points with less certainty fell outside the boundaries of heterozygotes and homozygotes and were scored as 'no call'. Genotype calls were exported as spreadsheets from BeadStudio and converted to create input files for PowerMarker v3.25 (Liu and Muse 2005) and DARwin5 (Perrier and Jacquemoud-Collet 2006).

Initially loci with more than 85% missing data were deleted from the analysis. Allele frequencies were calculated using PowerMarker v3.25 (Liu and Muse 2005) and monomorphic loci recorded. The remaining SNPs were considered validated for use in cultivated cassava. Results from the 53 germplasm selections were used to estimate minor allele frequencies (MAF). Observed heterozygosity, gene diversity (expected heterozygosity) and polymorphism information content (PIC) were calculated on a per locus basis, using PowerMarker v. 3.25 (Liu and Muse 2005). Validated SNPs were located on scaffolds of the cassava genome (v.4.1.) using BLASTN with the short Illumina SNP primer as a query. The length of the query and subject matches were compared to determine the length of the SNP region in the cassava genome, and to determine whether any insertions exist. In the case of a regular SNP, the 122 bp SNP sequence would correspond to the same length on a cassava scaffold, however, if there is an insertion in the genome within the SNP primer sequence, the corresponding region on the cassava genome would be longer.

Genetic diversity assessment

To provide insights into the genetic relationships of cassava from SEC Africa, Asia, West Africa and the Americas the simple matching coefficient was used to calculate a dissimilarity matrix with a pairwise variable deletion of 70%. Weighted neighbour-joining was used to construct a tree with 1,000 bootstrap iterations. All analyses were performed using DARwin5 (Perrier and Jacquemoud-Collet 2006). Allele frequencies were calculated on a regional basis using PowerMarker v3.25 (Liu and Muse 2005), and those with allele frequency differences $\geq$0.5 among regions ($\geq$0.45 in the case of 'Americas' and 'Africa'), showing maximum discrimination among regions, were identified. Mean observed heterozygosity (Ho) in the entire population and Nei's unbiased estimate of gene diversity (Nei 1987) within the regions, 'Americas', and 'Africa' and sub-regions 'West Africa' and 'Southern eastern and central Africa' were calculated using PowerMarker v3.25 (Liu and Muse 2005).

Results

EST development and characterisation

A total of 5,046 fragments were successfully sequenced from two cDNA libraries with 2,396 from the CASL library and 2,650 from the CASR library. All EST data are publically available through the National Center for Biotechnology Information [NCBI, Bethesda, MD, USA; GenBank dbEST accession nos FF379626 to FF382021 (CASL) and FF534207 to FF536856 (CASR)]. This comprised 706 contigs and 3,430 singletons.

Of the 2,404 sequences from CASR that had UniProt accession identifiers and that were subjected to GO annotation, putative functions were assigned to 1,291 unique sequences. A total of 3,485 annotations contributed to 'biological process', 3601 to 'metabolic function' and 1,968 to 'cellular component'. Of the 2,265 sequences from CASL that had UniProt accession identifiers and that were subjected to GO annotation, putative functions were assigned to 1,285 sequences. A total of 3,337 GO annotations were assigned to 'biological process', 3,656 were assigned to 'metabolic function' and 1,865 to 'cellular component'. GO annotations and GO Slim summaries for CASR and CASL are provided as supplementary material in Online Resource 1.

A cassava EST database

HarvEST:Cassava (http://harvest.ucr.edu/ and http://harvest.ucr.edu/HCassava106.exe) consists of 58,607 ESTs from 42,970 clones, assembled into 9,471 contigs and 13,432 singletons providing a Unigene set of 22,903 sequences. The contribution to HarvEST:Cassava by cDNA library is provided in Table 2. This includes 2,383 sequences from CASL and 2,636 sequences from CASR. Through this database it is possible to conduct a number of searches, including a search of ESTs by expression pattern, select a sequence or sequences using GenBank #, EST Name and Unigene #, and use NCBI blast for all protein sequences (blastx nr), all plant ESTs (tblastx est_others Embryophyta [orgn]) and cassava M01 ESTs blastn. The HarvEST:Cassava Unigene set with annotations from UniProt and the *Arabidopsis* database is provided in Online Resource 2.

From the Unigene set of 22,903 sequences 20,027 were assigned with specific UniProt identifiers using a cut-off of e−4. As a result of GO Slim annotation 8,501 were classified as 'cellular components', 23,129 were 'metabolic function' and 20,787 were 'biological processes'. Online Resource 2 provides details of GO annotations of the HarvEST:Cassava Unigene set and GO Slim annotation summaries.

SNP identification, characterisation and annotation

After low-quality masking and intron avoidance steps 3,380 putative SNPs were identified from the HarvEST:Cassava assembly. Of these, 426 were deleted for having many SNPs in close proximity indicating alignment error, and two putative SNPs also showed three alleles and were removed. As a result of the 'cleaning' process 2,954 putative SNPs remained from 1,234 unigenes. Of these 92 were likely to be within 30 bp of an intron junction and 100 were within 50 bp at the end of the sequence. These SNPs remain in the database of putative SNPs, but appropriate annotations have been made. The database of putative SNPs is available in Online Resource 3 and includes the relationship of SNP source sequences to HarvEST:Cassava unigenes.

SNPs were classified according to the minimum number of ESTs supporting each allele. For example, 2E SNP had two supporting ESTs for each allele, and a 3E SNP had three ESTs supporting each allele. Of the 2,954 SNPs, 1,829 were 2E, 509 were 3E, 247 were 4E, 106 were 5E and 263 were 6E. Confidence of being a real SNP, as opposed to a false positive is much higher in 3E SNPs and above than 2E SNPs. Of these SNPs, 1,653 were transitions (C/T or G/A) and 1,301 were transversions (C/G, A/T, C/A, or T/G). These details are provided for each putative SNP in Online Resource 3.

SNP validation

A set of 1,536 putative SNPs, representing 1,170 contigs from a possible 1,234 contigs, were selected for validation

Table 2 Contributions to HarvEST:Cassava by cDNA library

Library	Tissue	Treatment	# of clones	Total # ESTs	Forward	Reverse	Unknown	In # Cap3 contigs	In # Cap3 contigs	Contig members	Singles	% Unique
CASL	Leaf, stem, stem meristem	Normal and drought stressed, norm'd	2383	2383	2166	90	127	1358	54	117	739	35.9
CASR	Adventitious root	Normal and drought stressed, norm'd	2636	2636	2419	84	133	1366	64	131	993	42.6
CV01	Mixture	Normal, norm'd	9208	9208	8584	125	499	3595	290	674	2512	34.6
CV02	Mixture	Water stressed, norm'd	9425	9425	8927	110	388	3529	240	555	2013	27.2
CAS01	Leaf, root		19318	34955	18633	16322	0	7606	3300	10390	7175	50.3
Total			42970	58607	40729	16731	1147		3948	11867	13432	

using the Illumina's GoldenGate assay on a diverse collection of 53 cassava accessions (Table 1). Of the 1,170 contigs, 861 were represented by 1 SNP, 252 were represented by 2 SNPs and 57 were represented by 3 SNPs. Of these putative SNPs, 806 were 2E and 730 were 3E and above. Eleven SNPs had an Illumina score below 0.6. Details of the SNPs selected for inclusion in the Illumina GoldenGate oligonucleotide pool assay (OPA) are provided in Online Resource 3, under column 'Illumina selection'.

Of the 1,536 SNPs present in the Illumina GoldenGate assay, 178 (11.6%) failed to pass initial quality assurance threshold of the GC score (i.e. GC score <0.2). Six SNPs had more than 85% missing data points and were discarded. Most of the remaining SNPs had GC score >0.5 leaving 1,351 that were technically validated. Of these SNPs, 161 (12%) were monomorphic in the diversity panel leaving 1,190 polymorphic SNPs mostly with intermediate allele frequencies. Eighty-three SNPs had a minor allele frequency (MAF) of less than 0.05, with a mean MAF of 0.27. MAF, observed heterozygosity, gene diversity and PIC for each SNP locus are provided in Online Resource 4. Mean observed heterozygosity across loci was 0.3531, mean gene diversity was 0.3566 and mean PIC was 0.2836 with a maximum of 0.3746. Figure 1 shows the frequency distribution of PIC values across validated loci. By using the 121 bp SNP primer sequences, 1,116 of the 1,190 validated SNPs were located on scaffolds on the cassava genome v. 4.1 (http://www.phytozome.net/cassava). This data is provided in Online Resource 5. A total of 908 SNPs were uniquely present in the genome assembly, whereas the remaining 208 match the genome assembly at more than one location.

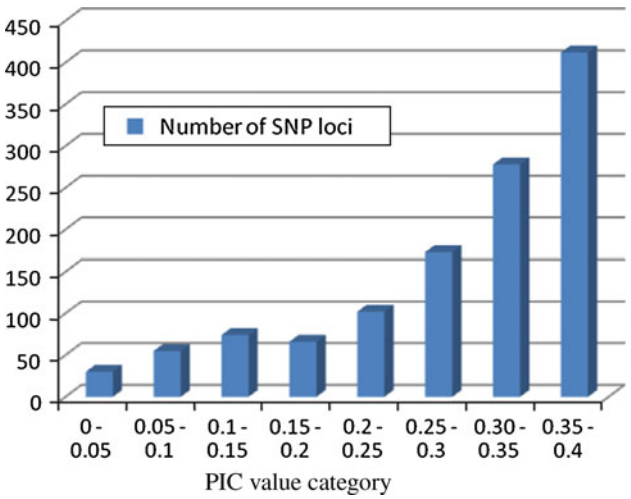

Fig. 1 Frequency distribution of PIC values calculated across 53 varieties genotyped for 1,190 SNPs

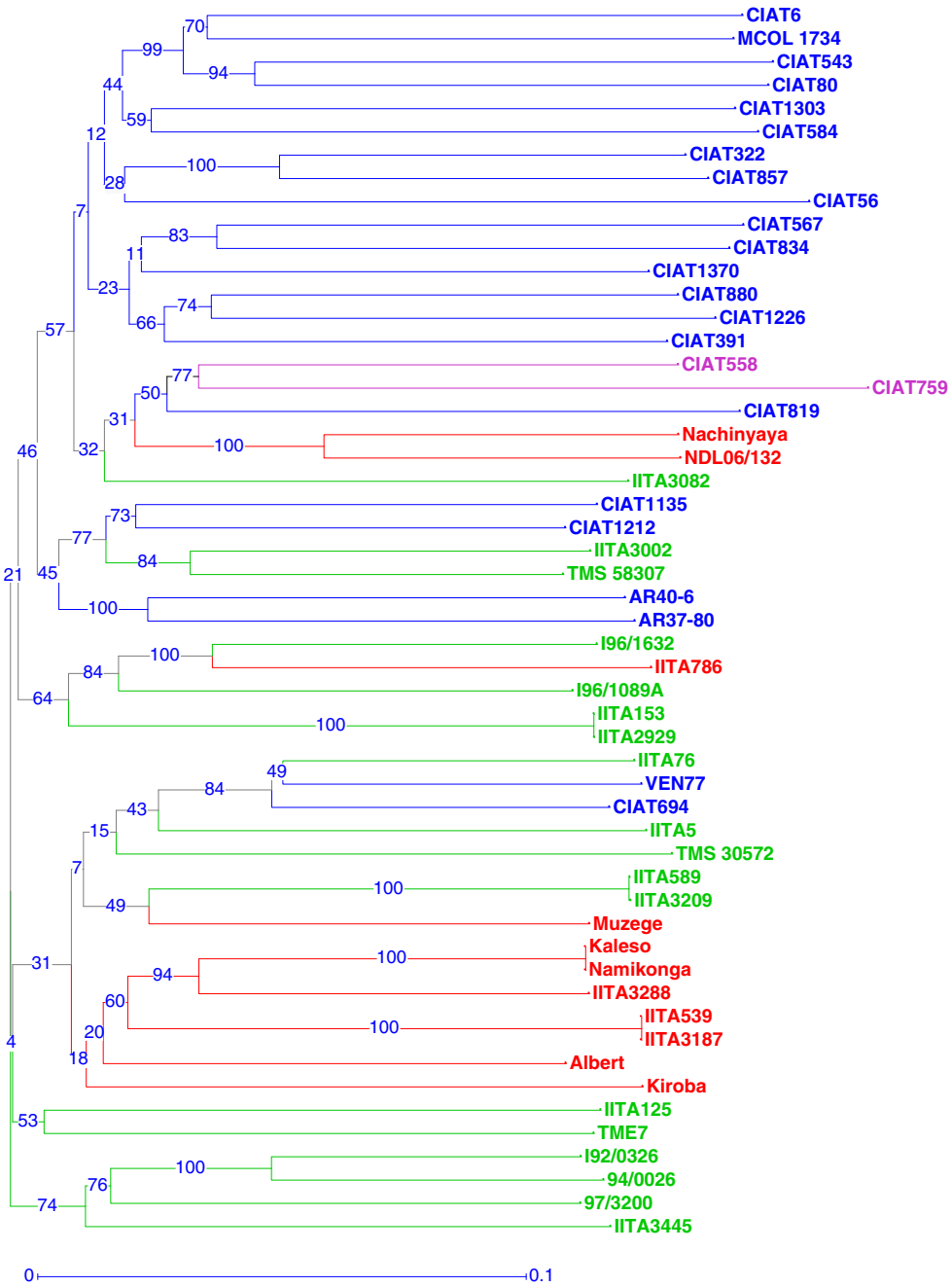

Fig. 2 Dendrogram illustrating relationship among 53 cassava varieties from the Americas (blue), Asia (purple), West Africa (green) and southern eastern and central Africa (red), based on 1,190 SNPs. The value on the branch refers to the bootstrap value (color figure online)

Genetic relationship among cassava varieties

The genetic structure of 53 diverse cassava accessions was analysed using DARwin and a dendrogram displaying genetic relationships among the cassava varieties from the Americas, Asia, West Africa and SEC Africa produced (Fig. 2). As expected the germplasm is mainly structured according to geographical origin, with the majority of varieties from the Americas clustering together. Another group containing IITA125, IITA 3445 and TME7, amongst others from West Africa group together. Genotypes from SEC Africa including 'Namikonga', 'Albert', 'Kiroba', IITA 3187, amongst others, also cluster together. Two accessions from Asia are closely related to one another, and to the group from the Americas. Several groups also contain germplasm from both West Africa and the Americas. In addition, two accessions from SEC Africa, 'Nachinyaya' and 'NDL06/132', cluster fairly closely to the Asian accessions and CIAT 819 from Brazil. Putative duplicate accessions included in the study to demonstrate the effectiveness of the technology for identifying duplicates within a genebank were found to be identical for all SNPs. Two farmer varieties, Namikonga and Kaleso, from East Africa, were also found to be identical. Observed

Table 3 SNPs with major differences in allele frequency among regions

Region	SEC Africa	West Africa	Africa
West Africa	Me.MEF.c.2351 (AG)		
	Me.MEF.c.3070 (CT)		
	Me.MEF.c.0854 (GT)		
	Me.MEF.c.2693 (AG)		
	Me.MEF.c.2353 (AC)		
	Me.MEF.c.3011 (AC)		
Americas	Me.MEF.c.0379 (AT)	Me.MEF.c.2948(AG)	Me.MEF.c.0047(AG)
	Me.MEF.c. 2907(AG)	Me.MEF.c.0047(AG)	Me.MEF.c.0379(AT)
	Me.MEF.c.0047(AG)	Me.MEF.c.0732(AG)	Me.MEF.c.0222(CT)*
	Me.MEF.c.2351(AG)	Me.MEF.c.1734(AG)	Me.MEF.c.2747(AG)*
	Me.MEF.c.0444(AT)	Me.MEF.c.2490(AG)	Me.MEF.c.2907(AG)*
	Me.MEF.c.3011(AC)	Me.MEF.c.2326(CT)	Me.MEF.c.3347 (CT)*
	Me.MEF.c.2951(AG)	Me.MEF.c.2327(AG)	Me.MEF.c.1734(AG)*
	Me.MEF.c.3124(CT)		
	Me.MEF.c.2351(AG)		

*SNPs with allele frequency differences among regions ≥ 0.45; all other SNPs allele frequency differences are ≥ 0.5

heterozygosity (Ho) across all loci was 0.3531. Unbiased estimates of gene diversity were 0.3488 for 'Americas' ($n = 22$), 0.3357 for 'Africa' ($n = 22$), 0.3230 for 'West Africa' ($n = 11$) and 0.3221 for SEC Africa ($n = 11$). Those loci providing maximum discrimination among regions, by allele frequency, are provided in Table 3. Six loci had allele frequency differences $\geq$0.5 between SEC Africa and West Africa, nine between SEC Africa and the Americas, seven between West Africa and the Americas, and just two between Africa as a whole and the Americas.

Discussion

Genomic tools are required to make a significant improvement in the ability to apply molecular markers to cassava research and breeding. Here, for the first time a substantial number (1,190) of biologically and technically validated SNP markers are presented for cassava. The location of the majority of the SNP markers in the cassava genome sequence assembly are reported. In addition, new ESTs and a searchable database of quality EST sequences are presented. It is anticipated that these resources will dramatically improve the quality of molecular marker applications in cassava.

In many databases and applications, sequence quality is a serious issue, particularly for SNP identification and genetic linkage mapping (Close et al. 2009). HarvEST:Cassava version 1.06 (http://harvest.ucr.edu/ and www.harvest-web.org) provides a valuable, easily searchable database of high quality sequence. It is possible that this database could be updated as new ESTs become available. The additional EST sequences reported here were derived from four African farmer varieties and one improved variety. The previous published cassava ESTs have been generated from four improved varieties (Anderson et al. 2004; Lokko et al. 2007; Sakurai et al. 2007). From sequences in HarvEST:Cassava, 2,954 putative SNPs were identified from 1,234 unigenes. Of these, 1,536 from 1,170 unigenes were selected for validation, leaving 1,418 SNPs still to be validated. Sequence information is provided in Online Resource 3.

In this study the suitability of the GoldenGate SNP assay for genotyping a diverse panel of cassava genotypes is demonstrated. Quality genotyping data was obtained from 1,358 out of the 1,536 markers used, representing a success rate of 89%. This rate is similar to that obtained for barley [90% (Rostoks et al. 2005)], 89% each for soybean (Hyten et al. 2008), potato (Anithakumari et al. 2010) and tetraploid wheat (Akhunov et al. 2009). Failed assays can be attributed to various factors, the most plausible are those proposed by Anithakumari et al. (2010), including incorrectly synthesized primers as a result of incorrect sequence data, polymorphism (including indels) in the primer annealing site and a large intron within the SNP primer target sequences. It is possible that initial intron avoidance steps missed some intron junctions. Excluding markers with large amounts of missing data, 12% of the 1,351 SNPs were found to be monomorphic in the panel of 53 genotypes used. This does not necessarily imply that these SNPs failed biologically, as the genotypes from which the SNPs were generated, were not included in the diversity panel. A total of 1,190 SNPs were biallelic and were validated technically and biologically. Of these 1,116 were located on scaffolds on the cassava genome v. 4.1. To date scaffolds on the cassava genome v.4.1. are not anchored on a physical map, so chromosomal locations cannot be assigned. A recent study by Sraphet et al. (2011) anchored

these scaffolds using SSR markers. The fact that 208 SNP primer sequences aligned to multiple locations on the genome sequence could either be due to the presence of homoeologous sequences, duplication in the cassava genome or the presence of several identical copies of some genes. This could confound genome map positions and pose difficulties for designing robust SNPs. Evidence that cassava evolved from an ancient duplication of the castor bean (*Ricinus communis*) is indicated through alignment of the cassava and castor bean genomes (Steve Rounsley per comm.). For genotyping purposes it is preferable to select SNPs whose primers anneal to a single distinct location on the genome and, with the genome sequence now available, it is feasible to design such primers.

Although the sample size for a diversity assessment here is small, results demonstrate some divergence in cassava germplasm according to the geographical region of origin, particularly between the Neotropics and Africa, and some sub-structure between germplasm from SEC Africa and West Africa. The results support previous diversity assessments in cassava using SSR markers. From 67 SSR loci and 283 accessions of cassava landraces from Africa (Tanzania and Nigeria) and the Neotropics (Brazil, Colombia, Peru, Venezuela, Guatemala, Mexico and Argentina), Fregene et al. (2003) found a low level of differentiation among country samples, yet sufficient distance between individual genotypes to separate African from Neotropical accessions and to reveal a more pronounced sub-structure in the African landraces. Mean observed heterozygosity across 1,190 loci was 0.3531 which was lower than that observed by Fregene et al. (2003) of 0.5136. The biallelic nature of SNPs as opposed the multi-allelic nature of SSRs explains this difference. A slightly larger gene diversity was found in the Americas (0.3488) as opposed to that in Africa (0.3357), both with $n = 22$. This larger gene diversity is consistent with a centre of origin and domestication in the Americas. Cassava was introduced into Africa, arriving at the western and eastern coasts by Portuguese slave ships from Brazil, during the 1500s until the 1800s (Jones 1969). Germplasm from SEC Africa and West Africa had very similar levels of gene diversity (0.3221 and 0.3230, respectively). Data suggest a larger diversity assessment using the SNP markers presented could be extremely informative for plant breeding and conservation purposes. SSR markers, together with isozymes and AFLP markers have previously identified duplicates in the CIAT core collection (Chavarriaga-Aguirre et al. 1999). The SNP markers presented here are shown to be effective in identifying duplicates within a cassava germplasm collection. The larger number of markers presented here, together with the advent of high-throughput genotyping technologies, enables identification of duplicates with greater confidence than previously possible when SSR markers were available.

Conclusions

A quantum improvement in the application of molecular markers to cassava research and breeding requires a high density of SNPs. Here the first published identification of a substantial number of SNPs (1,190) in cassava that are both technically and biologically validated is presented. Additional ESTs have been identified and compiled with existing sequences into an easily searchable EST database of high quality sequences, HarvEST: Cassava which should facilitate further applications of genomics research. It is anticipated that these resources will facilitate improved dissection of the genetic architecture of economically important traits and the application of modern genomics-based breeding approaches to cassava.

Acknowledgments The authors would like to thank Steve Rounsley and Simon Prochnik for valuable comments. This work was funded by BioSciences eastern and central Africa Network (BecANet) (Goldengate assay development and genotyping), the Generation Challenge Program (GCP) (cDNA library development, sequencing and initial bioinformatics) and the International Institute of Tropical Agriculture (IITA) (HarvEST:Cassava development and SNP identification).

References

Akhunov E, Nicolet C, Dvorak J (2009) Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. Theor Appl Genet 119:507–517

Anderson JV, Delseny M, Fregene MA, Jorge V, Mba C, Lopez C, Restrepo S, Soto M, Piegu B, Verdier V, Cooke R, Tohme J, Horvath DP (2004) An EST resource for cassava and other species of Euphorbiaceae. Plant Mol Biol 56:527–539

Anithakumari A, Tang J, van Eck H, Visser R, Leunissen J, Vosman B, van der Linden C (2010) A pipeline for high throughput detection and mapping of SNPs from EST databases. Mol Breed 26:65–75

Appleby N, Edwards D, Batley J (2009) New technologies for ultra-high throughput genotyping in plants. In: Gustafson JP, Langridge P, Somers DJ (eds) Plant Genomics. Methods in Molecular Biology. Humana Press, New York, pp 19–39. http://dx.doi.org/10.1007/978-1-59745-427-8_2

Batley J, Barker G, O'Sullivan H, Edwards KJ, Edwards D (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiol 132:84–91

Brumfield R, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. Trends Ecol Evol 18:249–256

Chao S, Zhang W, Akhunov E, Sherman J, Ma Y, Luo MC, Dubcovsky J (2009) Analysis of gene-derived SNP marker polymorphism in US wheat (*Triticum aestivum* L.) cultivars. Mol Breed 23:23–33

Chavarriaga-Aguirre PP, Maya MM, Bonierbale M, 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 M, Iglesias C, Bonierbale M, 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. http://dx.doi.org/10.1023/A:1009627231450

Cho RJ, Mindrinos M, Richards DR, Sapolsky RJ, Anderson M, Drenkard E, Dewdney J, Reuber TL, Stammers M, Federspiel N, Theologis A, Yang WH, Hubbell E, Au M, Chung EY, Lashkari D, Lemieux B, Dean C, Lipshutz RJ, Ausubel FM, Davis RW, Oefner PJ (1999) Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. Nat Genet 23:203–207

Close TJ, Wanamaker SI, Caldo RA, Turner SM, Ashlock DA, Dickerson JA, Wing RA, Muehlbauer GJ, Kleinhofs A, Wise RP (2004) A new resource for cereal genomics: 22K barley GeneChip comes of age. Plant Physiol 134:960–968

Close T, Bhat P, Lonardi S, Wu Y, Rostoks N, Ramsay L, Druka A, Stein N, Svensson J, Wanamaker S, Bozdag S, Roose M, Moscou M, Chao S, Varshney R, Szucs P, Sato K, Hayes P, Matthews D, Kleinhofs A, Muehlbauer G, DeYoung J, Marshall D, Madishetty K, Fenton R, Condamine P, Graner A, Waugh R (2009) Development and implementation of high-throughput SNP genotyping in barley. BMC Genomics 10:582

Dixon AGO, Bandyopadhyay R, Coyne D, Ferguson M, Ferris S, Hanna R, Hughes J, Ingelbrecht I, Legg J, Mahungu N, Manyong V, Mowbray D, Neuenschwander P, Whyte J, Hartmann P, Ortiz R (2003) Cassava: from poor farmer's crop to pacesetter of African rural development. Chronica Hortic 43:8–15

El Sharkawy MA (2004) Cassava biology and physiology. Plant Mol Biol 56:481–501

FAO (2008) http://faostat.fao.org/site/567/DesktopDefault.aspx?PageID=567#ancor

Fregene MA, Suarez M, Mkumbira J, Kulembeka H, Ndedya E, Kulaya A, Mitchel S, Gullberg U, Rosling H, Dixon AG, Dean R, Kresovich S (2003) Simple sequence repeat marker diversity in cassava landraces: genetic diversity and differentiation in an asexually propagated crop. Theor Appl Gene 107:1083–1093

Garg K, Green P, Nickerson DA (1999) Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. Genome Res 9:1087–1092

Grivet L, Glaszmann J-C, Vincentz M, da Silva F, Arruda P (2002) ESTs as a source for sequence polymorphism discovery in sugarcane: example of the Adh genes. Theor Appl Genet 106:190–197

Hyten D, Song Q, Choi IY, Yoon MS, Specht J, Matukumalli L, Nelson R, Shoemaker R, Young N, Cregan P (2008) High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. Theor Appl Genet 116:945–952

Jennings D, Iglesias C (2001) Breeding for crop improvement. In: Hillocks R, Thresh J (eds) Cassava biology, production and utilization. CAB International, Oxon, pp 149–166

Jones W (1969) Manioc in Africa. Stanford University Press, Stanford

Kawuki R, Ferguson M, Labuschagne M, Herselman L, Kim DJ (2009) Identification, characterisation and application of single nucleotide polymorphisms for diversity assessment in cassava (*Manihot esculenta* Crantz). Mol Breed 23:669–684

Kunkeaw S, Tangphatsornruang S, Smith DR, Triwitayakorn K (2010) Genetic linkage map of cassava (*Manihot esculenta* Crantz) based on AFLP and SSR markers. Plant Breed 129:112–115

Lebot V (2009) Tropical root and tuber crops: cassava, sweet potato, yams and aroids. Crop Production Science in Horticulture 17. CABI, Wallingford

Liu K, Muse S (2005) PowerMarker. Integrated analysis environment for genetic marker data. Bioinformatics 21:2128–2129

Lokko Y, Anderson J, Rudd S, Raji A, Horvath D, Mikel M, Kim R, Liu L, Hernandez A, Dixon A, Ingelbrecht I (2007) Characterization of an 18, 166 EST dataset for cassava (*Manihot esculenta* Crantz) enriched for drought-responsive genes. Plant Cell Rep 26:1605–1618

Lopez C, Jorge V, Piegu B, Mba C, Cortes D, Restrepo S, Soto M, Laudie M, Berger C, Cooke R, Delseny M, Tohme J, Verdier V (2004) A unigene catalogue of 5700 expressed genes in cassava. Plant Mol Biol 56:541–554

Lopez C, Piegu B, Cooke R, Delseny M, Tohme J, Verdier V (2005) Using cDNA and genomic sequences as tools to develop SNP strategies in cassava (*Manihot esculenta* Crantz). Theor Appl Genet 110:425–431

Luo M, Dang P, He G, Holbrook C, Bausher M, Lee R (2005) Generation of expressed sequence tags (ESTs) for gene discovery and marker development in cultivated peanut. Crop Sci 45:343–356

Mba REC, Stephensen P, Edwards K, Melzer S, Nkumbira J, Gullberg U, Apel K, Gale M, Tohme J, Fregene M (2001) Simple sequence repeat (SSR) markers survey of the cassava (*Manihot esculenta* Crantz) genome: towards an SSR-based molecular genetic map of cassava. Theor Appl Genet 102:21–31

McCarthy F, Wang N, Magee GB, Nanduri B, Lawrence M, Camon E, Barrell D, Hill D, Dolan M, Williams WP, Luthe D, Bridges S, Burgess S (2006) AgBase: a functional genomics resource for agriculture. BMC Genomics 7:229

Nei M (1987) Molecular Evolutionary Genetics. Colombia University Press, New York

Nelson R, Naylor RL, Jahn MM (2004) The role of genomics research in improvement of "orphan" crops. Crop Sci 44:1901–1904

Perrier X, Jacquemoud-Collet JP (2006) DARwin software. http://darwin.cirad.fr/darwin.http://darwin.cirad.fr/darwin

Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA, Boyce-Jacino M (1999) Mining SNPs from EST databases. Genome Res 9:167–174

Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 5:94–100

Raji A, Anderson J, Kolade O, Ugwu C, Dixon A, Ingelbrecht I (2009) Gene-based microsatellites for cassava (*Manihot esculenta* Crantz): prevalence, polymorphisms, and cross-taxa utility. BMC Plant Biol 9:118

Rostoks N, Borevitz J, Hedley P, Russell J, Mudie S, Morris J, Cardle L, Marshall D, Waugh R (2005) Single-feature polymorphism discovery in the barley transcriptome. Genome Biol 6:R54

Sakurai T, Plata G, Rodriguez-Zapata F, Seki M, Salcedo A, Toyoda A, Ishiwata A, Tohme J, Sakaki Y, Shinozaki K, Ishitani M (2007) Sequencing analysis of 20,000 full-length cDNA clones from cassava reveals lineage specific expansions in gene families related to stress response. BMC Plant Biol 7:66

Schmid KJ, Sörensen TR, Stracke R, Törjék O, Altmann T, Mitchell-Olds T, Weisshaar B (2003) Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. Genome Res 13:1250–1257

Sraphet S, Boonchanawiwat A, Thanyasiriwat T, Boonseng O, Tabata S, Sasamoto S, Shirasawa K, Isobe S, Lightfoot D, Tangphatsornruang S, Triwitayakorn K (2011) SSR and EST-SSR-based genetic linkage map of cassava (*Manihot esculenta* Crantz). Theor Appl Genet 122:1161–1170

Sriroth K, Piyachomkwan K, Wanlapatit S, Oates C (2000) Cassava starch technology: the Thai experience. Starch 52:439–449

Syvänen AC (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat Rev Genet 2:930–942

Tangphatsornruang S, Sraphet S, Singh R, Okogbenin E, Fregene M, Triwitayakorn K (2008) Development of polymorphic markers from expressed sequence tags of *Manihot esculenta* Crantz. Mol Ecol Resour 8:682–685

Tonujari NJ (2004) Cassava and the future of starch. Electron J Biotechnol 7:5–8

Xia L, Peng K, Yang S, Wenzl P, Carmen de Vicente M, Fregene M, Kilian A (2005) DArT for high-throughput genotyping of Cassava (*Manihot esculenta*) and its wild relatives. Theor Appl Genet 110:1092–1098