Theor Appl Genet (2012) 125:329–342
DOI 10.1007/s00122-012-1836-4

ORIGINAL PAPER

An EST-derived SNP and SSR genetic linkage map of cassava (*Manihot esculenta* Crantz)

Ismail Yusuf Rabbi · Heneriko Philbert Kulembeka · Esther Masumba · Pradeep Reddy Marri · Morag Ferguson

Received: 17 October 2011 / Accepted: 24 February 2012 / Published online: 15 March 2012

Abstract Cassava (*Manihot esculenta* Crantz) is one of the most important food security crops in the tropics and increasingly being adopted for agro-industrial processing. Genetic improvement of cassava can be enhanced through marker-assisted breeding. For this, appropriate genomic tools are required to dissect the genetic architecture of economically important traits. Here, a genome-wide SNP-based genetic map of cassava anchored in SSRs is presented. An outbreeder full-sib (F1) family was genotyped on two independent SNP assay platforms: an array of 1,536 SNPs on Illumina's GoldenGate platform was used to genotype a first batch of 60 F1. Of the 1,358 successfully converted SNPs, 600 which were polymorphic in at least one of the parents and was subsequently converted to KBiosciences' KASPar assay platform for genotyping 70 additional F1. High-precision genotyping of 163 informative SSRs using capillary electrophoresis was also carried out. Linkage analysis resulted in a final linkage map of 1,837 centi-Morgans (cM) containing 568 markers (434 SNPs and 134 SSRs) distributed across 19 linkage groups. The average distance between adjacent markers was 3.4 cM. About 94.2% of the mapped SNPs and SSRs have also been localized on scaffolds of version 4.1 assembly of the cassava draft genome sequence. This more saturated genetic linkage map of cassava that combines SSR and SNP markers should find several applications in the improvement of cassava including aligning scaffolds of the cassava genome sequence, genetic analyses of important agro-morphological traits, studying the linkage disequilibrium landscape and comparative genomics.

Communicated by A. Bervillé.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-012-1836-4) contains supplementary material, which is available to authorized users.

I. Y. Rabbi (✉)
International Institute of Tropical Agriculture (IITA),
PMB 5320 Oyo Road, Ibadan, Nigeria
e-mail: I.Rabbi@cgiar.org

H. P. Kulembeka
Ukiriguru Agricultural Research Institute,
P.O. Box 1433, Mwanza, Tanzania

E. Masumba
Sugarcane Research Institute, P.O. Box 30031,
Kibaha, Tanzania

P. R. Marri
BIO5 Institute, University of Arizona,
Tucson, AZ, USA

M. Ferguson
IITA c/o ILRI, P.O. Box 30709 00100,
Nairobi, Kenya

Introduction

Cassava, (*Manihot esculenta* Crantz) ($2n = 36$), which originates from Latin America, is one of the most important food crops with a worldwide production of 233 million tons, of which more than 50% occurs in Africa (FAOSTAT 2010). Asia and the Americas contribute about 33 and 15% of the world production, respectively. Numerous traits of the crop such as drought and heat tolerance, and little requirement for agricultural fertilizers make it an attractive crop. Though traditionally grown by subsistence farmers, the use of cassava for agro-industrial processing such as starch and biofuel is increasing (Thresh 2006; Jansson et al. 2009). Despite its monoecious flowering nature, self-pollination in cassava is mainly prevented by protogyny

(Alves 2002) thus rendering the crop highly heterozygous. The crop is propagated via stem cuttings and seed-based production is limited.

Cassava is one of the least studied major crops of the world and the potential for genetic improvement is substantial (Okogbenin and Fregene 2003). Understanding the genetic architecture of economically important traits in cassava can be assisted by the development and utilization of modern genomic tools such as single nucleotide polymorphism (SNP) markers that are distributed throughout the crop's genome and ideally anchored on a genetic linkage map. Genetic maps offer a framework for carrying out evolutionary and comparative genomic studies (Ahn and Tanksley 1993). They are also crucial in the search for Mendelian and quantitative trait loci (QTL) (Lander and Botstein 1989), and understanding the organization and dynamics of organism genome such as landscape of linkage disequilibrium (LD) (Flint-Garcia et al. 2003). Knowledge of LD, defined as the correlation of alleles at different loci, is required to determine the density of markers needed for association mapping (Buckler and Thornsberry 2002) and genomic selection (Meuwissen et al. 2001). High-resolution genetic linkage maps can also assist in aligning scaffolds in the cassava genome assembly.

To date, several linkage maps having between 100 and 510 markers comprising amplified fragment length polymorphism (AFLPs), restriction fragment length polymorphism (RFLPs), randomly amplified polymorphic DNA (RAPDs), simple sequence repeats (SSRs) and expressed sequence tags (EST)-derived SSRs markers have been published (Sraphet et al. 2011; Chen et al. 2010; Kunkeaw et al. 2010, 2011; Fregene et al. 1997; Okogbenin et al. 2006). With this limited genomic resource, QTLs controlling cyanogenic glucosides and dry matter content (Whankaew et al. 2011; Kizito et al. 2007), plant architecture and productivity (Boonchanawiwat et al. 2011; Okogbenin and Fregene 2003; Okogbenin et al. 2008), bacterial blight (Wydra et al. 2004; Lopez et al. 2007; Jorge et al. 2000), and cassava mosaic disease (Akano et al. 2002) have been studied. Hitherto, no SNP-based genetic map of cassava has been developed, yet SNPs are increasingly becoming markers of choice for dense genotyping owing to their abundance (Rafalski 2002) occurring at a frequency of one SNP every 100–300 bp (Edwards et al. 2010). Unlike other traditional markers such as RFLP, AFLP and SSRs, SNPs are highly amenable to ultra-high throughput genotyping technologies (Syvänen 2001, 2005; Akhunov et al. 2009; Appleby et al. 2009), thus compensating for their short-coming of being predominantly bi-allelic (Syvänen 2001) relative to SSRs that are multi-allelic. SNP genotyping can be performed using predefined SNP arrays or de novo using techniques such as 'genotyping-by-sequencing' (Elshire et al. 2011).

A resource of 2,954 putative SNPs has been identified from ESTs generated from the transcriptome of cassava. From these, a set of 1,536 SNPs were converted to Illumina's BeadArray technology-based GoldenGate assay and genotyped on a set of diverse cassava accessions (Ferguson et al. 2011). In the present study, we used this customary 1536 oligo pool assay (OPA) to genotype a bi-parental full-sib (F1) population with the objective of developing a first-generation SNP-based genetic linkage of cassava. In addition, a subset of these SNPs which segregated in the parents was also converted to competitive allele-specific PCR (KASPar) platform (KBioscience—LGC Genomics) and used to genotype additional F1 from the same family.

Materials and methods

Plant materials

An outbreeder full-sib family (F1) was developed via an intra-specific cross involving two cassava (*Manihot esculenta*) parents from Eastern-Africa: Namikonga (female) is one of the best sources of resistance to cassava brown streak disease (CBSD) (Kanju et al. 2010) while Albert, the male parent is susceptible to CBSD. In addition, Albert was selected due to its high flowering ability, tolerance to cassava mosaic disease and genetic divergence from Namikonga based on preliminary screening of several potential parents with SSRs. Crosses were carried out in Tanzania. Embryo rescue was employed to germinate some of the seeds at CIAT, Colombia (CIAT 2003). DNA was extracted from ~0.5 g of young leaf tissue using a modification of the miniprep protocol as described by Dellaporta et al. (1983). The concentration and quality of extracted DNA were assessed using NanoDrop™ ND-1000 Spectrophotometer (Thermo Fisher Scientific, USA) in addition to electrophoresis on a 0.8% agarose gel.

SSR genotyping

A total of 163 genomic and EST-derived SSR markers that were polymorphic in the parents were genotyped in the mapping population. Capillary electrophoresis was used for high-resolution fragment analysis of the SSR markers. For multiplexing the SSR markers during the capillary electrophoresis, the PCR primers were either (i) directly labeled with fluorescent dye or (ii) indirectly during PCR setup using fluorescent universal primer sequence derived from octopus specific DNA sequence (sequence: 5′-GCTACAGAGCATCTGGCTCACTGG-3′) concatenated with locus-specific forward primer (Schuelke 2000). In the direct method, the dye is covalently attached

to the 5′-end of, usually, the forward primer. The indirect method provides the flexibility to arbitrarily choose a specific dye during the reaction setup. The SSRs were amplified in a 10 µl reaction mixture containing 10 ng of total genomic DNA, 1× PCR buffer, 2 mM $MgCl_2$, 0.2 mM dNTPs, 0.375 U *Taq* DNA-polymerase, 0.8 mM of each primer using Dual 384-Well GeneAmp® PCR System 9700 (Applied Biosystems, Foster City, CA, USA). The amplification protocol comprised an initial denaturation for 2 min at 95°C then 30 cycles of denaturing at 95°C for 30 s, annealing at between 57 and 62°C for 1 min, extension at 72°C for 1 min and a final extension of 72°C for 30 min. One microlitre of three to four pooled PCR products, mixed with 9 µl formamide-standard mix (0.11 µl GS500 LIZ and 8.89 µl Hi-Di Formamide, Applied Biosystems), was denatured at 95°C for 3 min before capillary electrophoresis on Applied Biosystem's 3730 DNA Analyzer. Allele sizes and genotyped scoring were determined using Genemapper Version 3.7 software (Applied Biosytems, Foster City, CA, USA) and resulting genotypic data converted to JoinMap 4 (van Ooijen 2006) format.

Ascertainment of F1 authenticity

Prior to SNP genotyping the authenticity of the F1 population was ascertained through examination of the SSR genotypic data. The presence of off-types or admixtures (i.e. progenies whose parents are other than Namikonga and Albert) was readily detected by the occurrence of alleles not found in the parents. In addition, markers that are polymorphic and with unique allele in the female parent (***ab*** × *cd*, ***lm*** × *ll*, and ***ef*** × *eg*—JoinMap nomenclature) were useful for detection of self-progenies. Occurrence of these unique alleles in homozygous states suggested selfing. The model-based cluster analysis implemented in STRUCTURE v2.3.2 (Pritchard et al. 2000) was also applied to determine the genetic composition of the mapping population. Since it does not require a priori grouping of progenies, this method is suitable to identify progenies derived from contaminating pollen that are expected to show distinct non-parental ancestry. The program was run using default parameters (i.e. correlated allele frequencies and admixture model) and the number of subpopulations (K) was varied from 2 to 8 with five independent runs for each K. The number of MCMC replications was 100,000 after a burn-in period of 50,000. The most likely number of genetic clusters was estimated using the ad hoc statistic (deltaK) (Evanno et al. 2005) calculated from the average of the five independent runs. The proportion of individual's ancestry assigned to each cluster (Q) for the most likely number of clusters was summarized using bar plots.

SNP genotyping (Illumina)

A custom array of 1,536 expressed sequence tag (EST)-derived SNPs, recently developed and validated in cassava (Ferguson et al. 2011), was used. SNP genotyping was carried out on two commercial platforms: Illumina's GoldenGate™ assay (Fan et al. 2003; Gunderson et al. 2005) and KASPar assay (KBioscence—LGC Genomics) (http://www.kbioscience.co.uk/). A first batch of samples consisting of 60 authenticated F1 progenies and the parents were genotyped with 1,536 SNPs on the Illumina platform in accordance with the manufacturer's protocol (Shen et al. 2005) at the Southern California Genotyping Consortium of the University of California, LA (SCGC, website: http://scgc.genetics.ucla.edu/services). The GoldenGate™ technology is based on highly multiplexed allele-specific primer extension (ASPE) with two allele-specific oligonucleotides (ASO) and one locus-specific oligo (LSO) to preferentially amplify correctly matched allele. After ligation, the amplicon is further amplified with universal primers using sequences concatenated to the ASOs and LSO. The PCR products are labeled with Cy3- and Cy5-fluorescent dye depending on the SNP allele and the product is hybridized to its complementary bead type through a unique barcode address sequences attached to the locus-specific oligo (LSO) (Gunderson et al. 2005). The genotypes are read on universal capture-beads assembled into 96 sample arrays. Each bead is replicated 30 times and this inherent high redundancy ensures robustness and genotyping accuracy. After normalization of allele signal intensities, each SNP was assigned a cluster position (and resulting genotype) with the BeadStudio software (Illumina, San Diego, CA). The quality of the SNP data was examined using GeneCall (GC) score, a reliability statistic calculated for each genotype (data point) and ranges from 0 to 1 (Shen et al. 2005). The score measures the relative positions of a SNP data point *vis a vis* the center of its genotypic cluster. A GC value below 0.25 generally indicates a failed SNP assay and the respective genotype occurring far from the center of their cluster while scores above 0.7 are considered to be reliable SNP scores (Anithakumari et al. 2010).

SNP genotyping (KBioscience)

From the 1,536 SNPs genotyped on Illumina's OPA, a subset of 600 informative SNPs was selected to genotype an additional set of 70 F1 progenies using KBiosciences' KASPar assay. This SNP genotyping chemistry uses a novel proprietary form of competitive allele-specific PCR system. KASPar is flexible, and like in the GoldenGate assay, and is based on competitive allele-specific PCR. However, the allele detection is based on a FRET quencher

cassette with fluorescence of either VIC or FAM fluorophores. During the PCR process, the quencher is released when its complementary flour-labeled oligo is incorporated in the amplification process and the appropriate signal is generated and read using a FRET-capable plate reader. The main difference is that the KASPar assay is designed to query one SNP at a time, while the GoldenGate is designed to assay multiplex of 96, 384, 1,536, or 3,072 SNPs per sample, 12–96 samples in parallel.

Construction of genetic linkage map

Genetic linkage maps were constructed using JoinMap version 4 (van Ooijen and Voorrips 2001) via two approaches as described by Tavassolian et al. (2010). Firstly, separate linkage maps were constructed from the male and female datasets (created with the function 'Create Maternal and Paternal Population Nodes' of JoinMap 4.0) followed by generation of an integrated parental map. The two parental maps were integrated using the 'Combine Groups for Map Integration' function of JoinMap 4. This is referred to as two-step method and maps produce are designated 'II'. Fully informative markers with three or four segregating alleles (*ab* × *cd* and *ef* × *eg*, respectively), double heterozygous markers (*hk* × *hk*) and markers segregating only in the female parent (*lm* × *ll*) were used to construct the female map. The same was done for the male map, except that *lm* × *ll* markers were replaced by *nn* × *np* markers which segregate in the male parent.

Secondly, a one-step approach, using the CP option, was used to construct a linkage map from the combined dataset, regardless of whether they were segregating in one or both parents and the resulting maps are denoted as 'I'. This approach has been used to produce most recent cassava genetic maps (Kunkeaw et al. 2010, 2011; Chen et al. 2010; Sraphet et al. 2011). For both two-step and one-step methods, Chi-square tests were performed to test for deviation from the expected Mendelian segregation ratio for each marker. Slightly skewed markers ($0.005 \leq P \leq 0.05$) were retained in the linkage analysis unless their presence affected marker order or greatly changed linkage distance. The grouping of linked markers was evaluated using independence LOD and regression mapping was used for map construction using default parameters (recombination frequency <0.4, LOD >1, and jump = 5). Map distances were calculated using the Kosambi mapping function.

To identify scaffolds associated with the markers used in this study, SNPs and SSRs were queried against the cassava genome assembly Version 4.1 (Prochnik et al. 2011; http://www.phytozome/cassava) using BLAST. A 122-bp nucleotide sequence spanning the SNP site and SSR primer sequences were used to carry out the search.

Results

Composition of the mapping population

From an initial mapping population of 195 individuals, SSR data revealed 130 progenies from true crosses between Namikonga and Albert and 52 off-types or admixtures with allele(s) that were not present in either of the parents, and 14 possible selfs were indicated by the presence of unexpected allele combinations. The results from cluster analysis using STRUCTURE revealed presence of 5 unique clusters ($K = 5$) in the mapping population (Fig. 1) and further corroborated with the ΔK statistic of Evanno et al. (2005). The clusters in this case correspond to parents of the mapping population. Admixtures resulting from pollen contamination from three unique male parents were also detected.

Marker polymorphism

The results were obtained from 1,358 (88.4%) of the 1,536 SNPs in the GoldenGate Assay. Six hundred (44%) SNPs were informative in at least one of the mapping parents. These were selected for genotyping the second batch of 70 F1 along with their two parents using the KASPar assay. Eleven of the selected SNPs were not successfully converted to the KASPar assay. Sixty-one

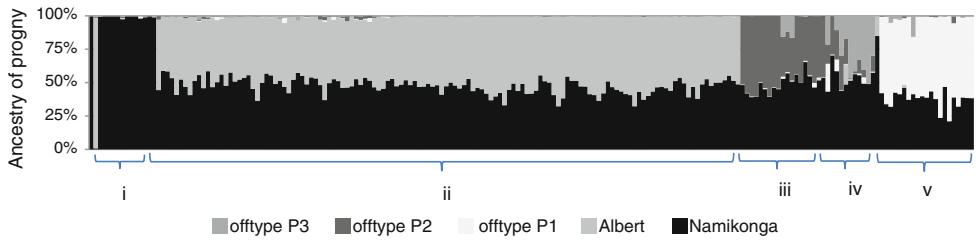

Fig. 1 Genetic structure of the putative F1 progenies of the Namikonga × Albert cross. *Each bar* represents an individual and the *different color* the ancestry of individual. *i* selfed Namikonga progenies, *ii* authentic Namikonga × Albert progenies ($n = 130$), *iii iv*, and *v* are off-type progenies from three other unknown parents (color figure online)

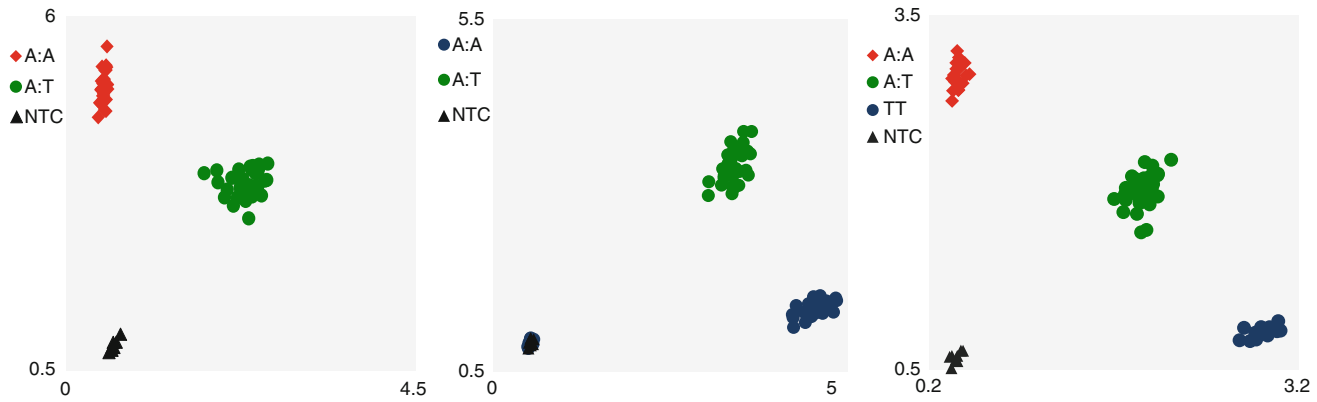

Fig. 2 An example of genotyping data plotted with Microsoft Excel for three different SNP segregation types in our mapping population. Genotypes are color-coded according to their nucleotide alleles. Genotyped samples *marked red* are homozygous for the allele reported with HEX, those *marked blue* are homozygous for the FAM allele; those *marked green* are heterozygous. Genotype cluster in *lower-left corner* are negative controls (color figure online)

SNPs gave inconsistent results across the two platforms. For example, 31 SNPs that were informative in the GoldenGate assay (AA × AC) were found to be fixed-differences between the two parents in the KASPar assay (e.g. AA × CC). A set of 528 good quality cross-platform SNPs data was identified. An example of genotype cluster separation is given in Fig. 2.

In addition to the SNP data, high quality genotypic data were obtained for 163 SSRs, thus brought the total number of markers used for linkage mapping to 691. Of these, 289 (42%) segregated in both parents: 9 with four alleles (*ab* × *cd*); 65 with three alleles (*ef* × *eg*); 215 with two alleles (*hk* × *hk*); 165 (24%) were heterozygous only in Namikonga (*lm* × *ll*); while 237 (34%) were heterozygous only in Albert (*nn* × *np*) (Table 1). Being bi-allelic, SNPs were scored as *lm* × *ll*, *nn* × *np* and *hk* × *hk*.

Segregation distortion

Of the 691 markers segregating in the mapping population, 620 (90%) were conformed to the expected Mendelian segregation ratios. Severe distortion was detected in 16 markers ($P \leq 0.001$) while 55 were moderately distorted ($0.005 \leq P \leq 0.05$). A very similar proportion of markers showed skewed segregation in the female parent, Namikonga (8%) and the male parent, Albert (11%). The proportion of distorted markers was similar for SSRs and SNPs (Fig. 3). Sixteen markers (11 SSRs and 5 SNPs) showing a highly skewed segregation ($P \leq 0.001$) were discarded from the analysis.

Table 1 A summary of number of markers and segregation of the SSR and SNP markers used in the present study

Marker type	abxcd	efxeg	hkxhk	lmxll	nnxnp	Total
SNP	–	–	213	119	196	528
SSR	9	65	2	46	41	163
Total	9	65	215	165	237	691

Linkage maps constructed using one-step and two-step

One-step method

A one-step linkage map of 1,837 cM containing 568 markers (435 SNPs and 133 SSRs, Table 2) distributed across 19 linkage groups (LG) was constructed using the cross-pollinator (CP) option of JoinMap®. Five of the 19 LGs (2-I, 5-I, 9-I, 10-I and 15-I) consisted of two subgroups each, which were found to be linked in the two-step mapping approach. The group sizes ranged from 15 cM (LG 18-I) to 143.43 cM (6-I) and average number of markers per linkage group ranged from 5 (19-I) to 45 (1-I). Individually, the average marker distance varied from 2.38 (1-I) to 11.05 (18-I) cM, and generally, larger groups with more markers had smaller inter-marker distances compared to shorter and sparse linkage groups (Table 2).

The one-step map contained 46 markers with distorted segregation ratios and are marked with *, **, ***, ****,

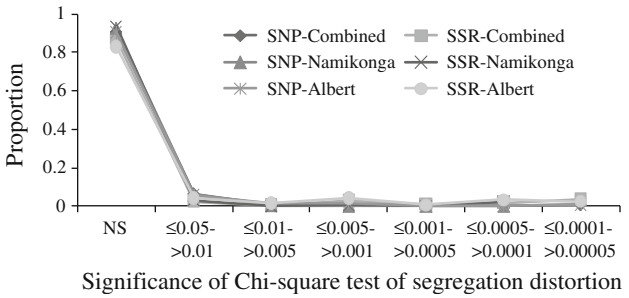

Fig. 3 Proportion of SSR and SNP markers showing various degrees of skewed segregation ratio for individual parents as well as combined set datasets

Table 2 Main characteristics of the one-step map

LG	SNP	SSR	Combined	Marker interval (cM)	Size (cM)
1	36	9	45	2.38	104.94
2	36	8	44	2.98	125.1[a]
3	37	6	43	2.82	118.51
4	35	8	42	2.49	104.45
5	33	7	40	2.86	114.5[a]
6	32	10	42	3.59	143.43
7	33	6	39	3.20	121.43
8	27	11	37	3.20	118.39
9	24	11	35	3.36	121.0[a]
10	17	16	33	4.11	107.4[a]
11	16	15	31	4.38	131.54
12	20	8	28	3.68	99.26
13	23	4	27	2.90	75.50
14	16	8	24	4.41	101.43
15	15	1	16	3.38	78.8[a]
16	12	1	13	4.27	51.29[b]
17	11	1	12	5.57	61.24
18	8	–	8	2.14	14.99[b]
19	2	3	5	11.05	44.19
Average	435	133	568	3.35	96.70

[a] Linkage groups represented as two subgroups and shown to be linked when aligned with two-step map

[b] Linkage groups having two corresponding subgroups in the two-step map

***** where $P = 0.05$, $P = 0.01$, $P = 0.005$, $P = 0.001$, and $P = 0.0005$, respectively (Fig. 4). Inclusion of these markers did not result in significant changes in marker order or inflation of map distances and they were therefore retained in the final map. These markers were randomly distributed throughout the genome, except for LG 7-I with seven moderately skewed loci and LG 8-I, with 14 skewed loci that cluster at one end of each linkage group. The significance levels of the Chi-square test for skewed markers in LG8-I showed a bell-shaped distribution with markers in the middle having more distortion than those at the both sides. Many linkage groups were free from skewed markers.

To better understand the underlying selection processes, whether gametic and/or zygotic, that resulted in the observed pattern of segregation distortion, the maximum likelihood framework of Bechsgaard et al. (2004) as outlined by Leppälä et al. (2008) was used to analyze the segregation data. For this analysis, only fully informative microsatellite markers were used. Deviation due to zygotic selection was assessed by comparing the full model against the gametic model using a likelihood ratio test ($-2\Delta L$) (for details see Leppälä et al. (2008). Similarly, gametic selection was tested by comparing the segregation of the alleles from each parent separately (gametic model) against Mendelian expectations (Mendelian model). The gametic selection was tested independently for each parent, considering two independent parental meioses that resulted in the F1 population.

Out of the total of 74 loci segregating in both parents (9 as $ab \times cd$ and 65 as $ef \times eg$), zygotic selection test was significant for only one locus (SSRY179) while gametic selection test revealed that 27 (36.5%) were significantly distorted based on parental allele combination tests (Allele test). Most of these loci were also significant under the initial genotypic tests. Thus gametic selection explains the observed pattern of segregation distortion in the present mapping population. Indeed, linkage group 8 markers with a cluster of skewed markers showed biased distortion towards specific allele in Albert. The microsatellite markers located on the distorted region of LG-7 and LG8-I were significant only for the allelic tests, particularly those originating from 'Albert'. This suggests that a gametic selection in favor of one of the paternal alleles (pollen). For example, the frequency of allele 'd' from 'Albert' is double (an average of 0.66) that of allele 'c' (average of 0.33) for four SSRs, namely, NS176, NS300, SSRY76 and NS909 that occur in the distorted region in LG8.

The distribution of marker distance is illustrated using a boxplot (Fig. 5). The average marker distance was slightly lower for the one-step map (3.35 cM) compared to the two-step map (3.51 cM). Maps developed using both approaches have several large gaps that are depicted as outliers in the boxplot that is skewed to the right. Frequency distribution of inter-marker distance (cM) indicates that more than 80% of the marker-pairs are separated by up to 6 cM.

Two-step method

Individual parental maps were developed and subsequently integrated. Using this two-step method, a map of 1,541 cM and 483 markers (348 SNPs and 128 SSRs) was created. The markers were distributed across 18 linkage groups with an average map distance of 90 cM per group (range 5.8 cM in group 17-II to 135.2 cM in group 2-II).

The one-step and two-step mapping approaches produced consistent mapping results for most markers, with linkage groups being largely collinear as shown in Fig. 4. Few linkage groups had minor inversions in the order of markers, often in clusters of closely linked markers (LG 5-I, II and LG 11-I, II). Very few markers showed larger shifts of more than 20 cM i.e. SNP MEF_c_3082 in LG7-I and LG7-II.

Correspondence with earlier maps

Correspondence between our map and a preceding SSR-based map of cassava developed by Whankaew et al.

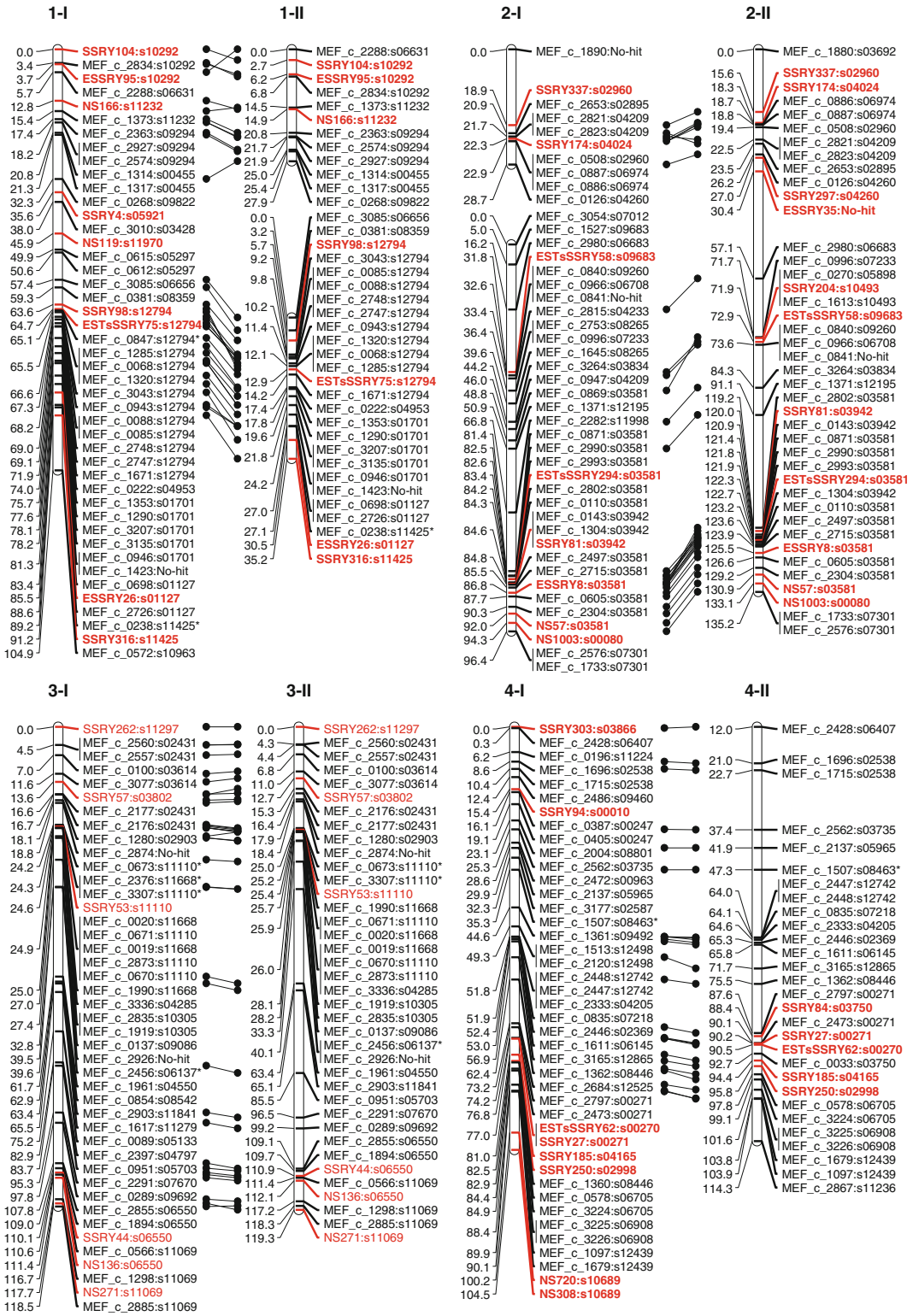

Fig. 4 An alignment of SNP and SSR genetic linkage maps of cassava derived from a 'Namikonga' × 'Albert' F1 population constructed using two-step and one-step approach. *Map on the left* (designated as I) was constructed from the combined dataset while the *corresponding map on the right* was derived from integrating the homologous parental linkage groups (designated as II). The SSR markers are shown in *bold*. The marker names are concatenated with their corresponding scaffolds (name starting with 's' followed by 5-digit identity)

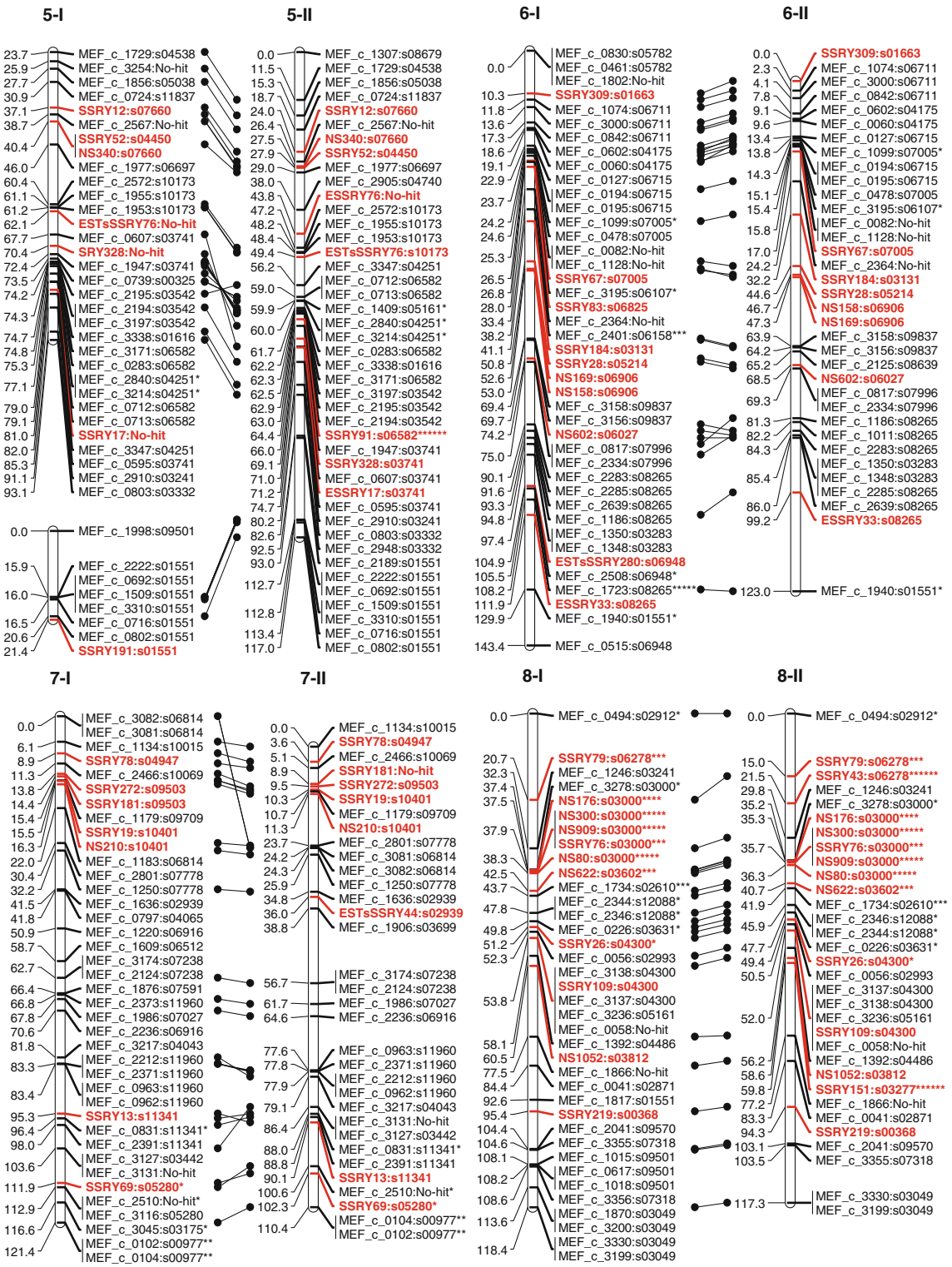

Fig. 4 continued

(2011) was checked. Co-linearity between the two maps was identified by aligning six pairs of linkage groups each having between 3 and 5 SSRs in common. This alignment revealed clear conservation in grouping and order of markers between the two maps (Online Resource 1).

Distribution of markers in scaffolds

A cassava genome assembly from whole genome shotgun sequencing is available at Phytozome (Cassava Genome Project 2010, http://www.phytozome.net/cassava). This

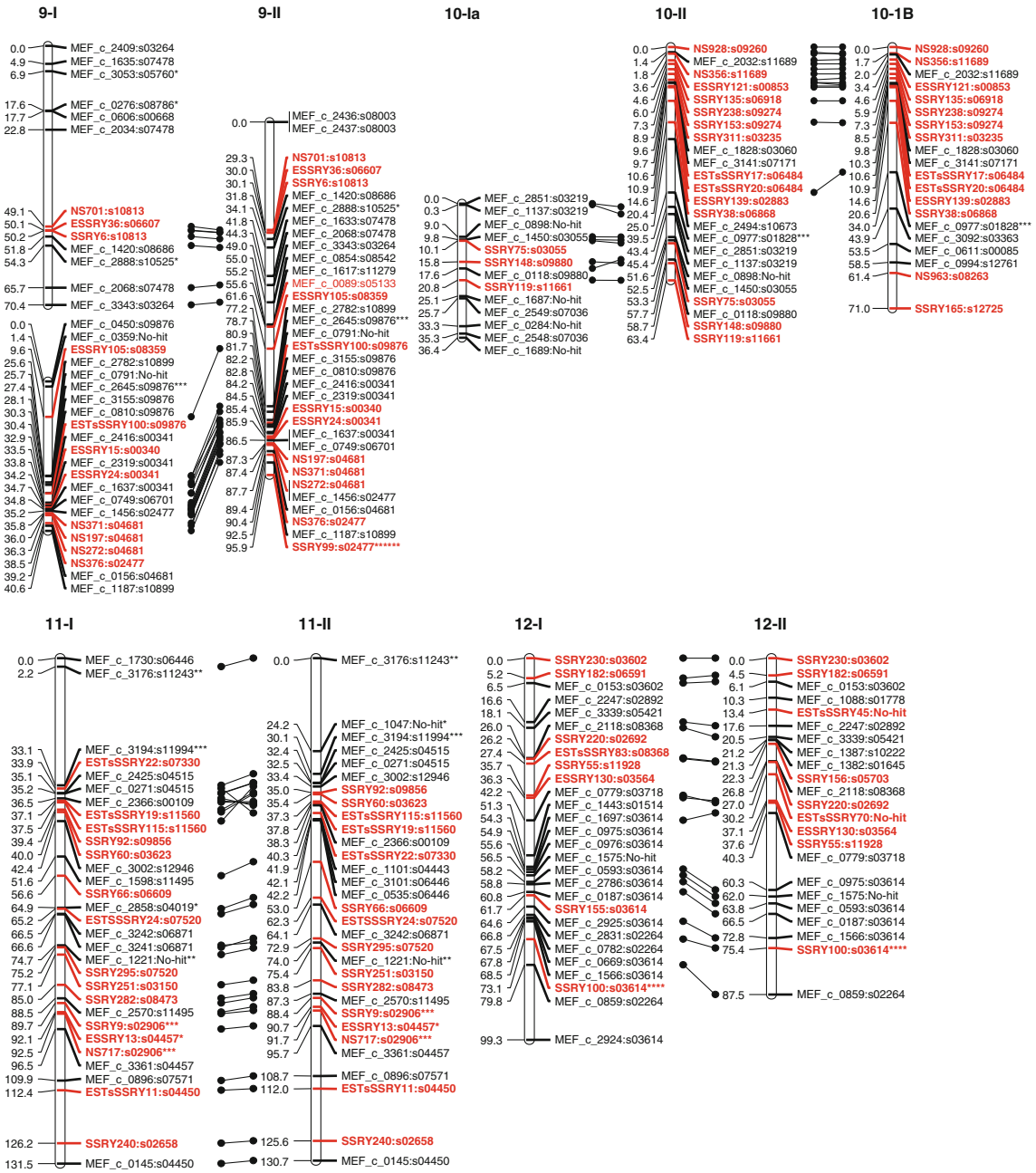

Fig. 4 continued

assembly consists of 12,977 scaffolds spanning 533 Mb, where half of the assembly is represented in 487 of the largest scaffolds (Prochnik et al. 2011). The locations of 535 markers, representing 94.2% of the markers in the one-step map were identified in 285 unique scaffolds of the cassava genome (Fig.4). Figure 6 shows the distribution of markers in scaffolds. There were 172 single-marker scaffolds, 104 scaffolds with 2–4 markers, and 12 scaffolds with more than four markers. Scaffold03614 had 14 markers followed by Scaffold03581 and Sacffold12794 with 13 markers each.

Discussion

Efficiency of marker-assisted breeding depends on the degree of saturation of genetic maps: denser map presents a higher likelihood of detecting polymorphic markers in any genetic background and genomic region of interest. The recent availability of SNP markers coupled with the development of high-throughput genotyping technologies and decreasing costs makes the utilization of SNP markers in cassava breeding a realistic goal. The objective of the present study was to develop a SNP-based genetic map of

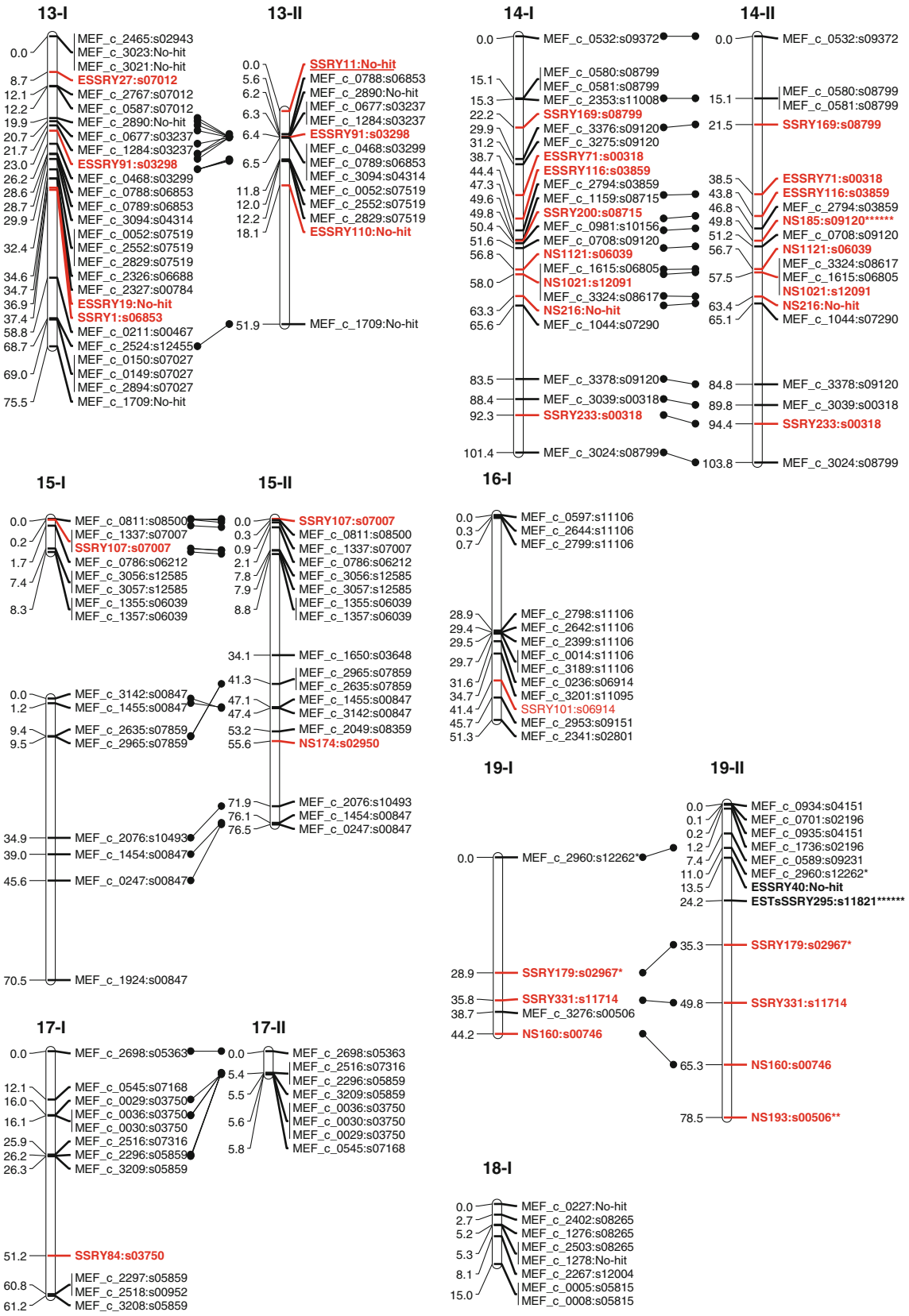

Fig. 4 continued

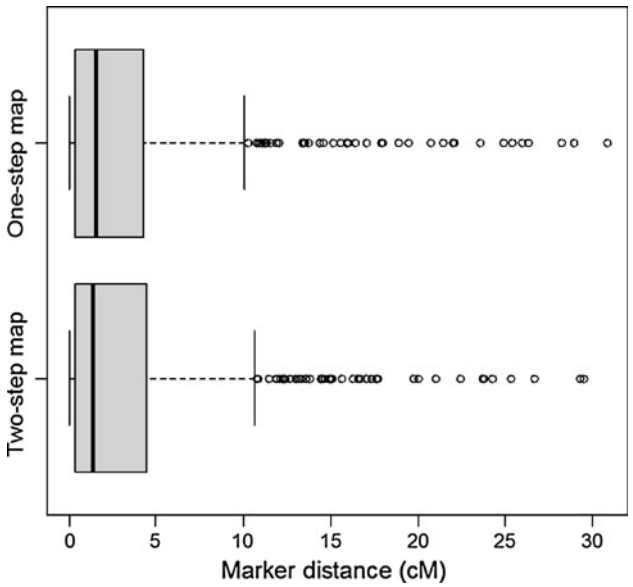

Fig. 5 A *boxplot* of distance (cM) between adjacent markers in the one-step and two-step linkage maps. The *bottom and top of the box* represent the 25th and 75th percentiles, respectively. Distances >10 cM are depicted as outliers in the *boxplots*

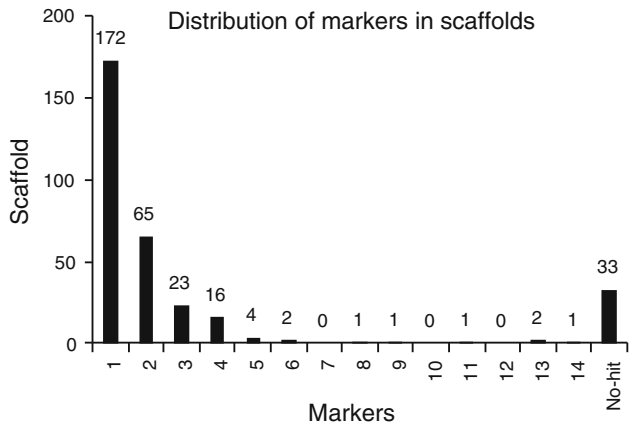

Fig. 6 Distribution of the number of markers occurring in various cassava genome assembly v4.1 scaffolds

cassava anchored in SSRs. Two independent and complementary SNP assay platforms (Illumina's GoldenGate and KBiosciences' KASPar chemistry) were used to genotype a full-sib family of heterozygous cassava from East-Africa. The GoldenGate platform has fixed arrays of markers and genotypes, which was 1,536 SNPs × 96 samples in the present study, while the Kbioscience's platform has the flexibility of handling any combination of samples × markers. This flexibility enabled use only a subset (i.e. 600) of the 1,536 SNPs on the GoldenGate's assay that was informative in the mapping population to genotype the second batch of F1.

Concordance of the GoldenGate and KASPar genotyping methodologies was determined by comparing results of parental genotypes that were genotyped redundantly on both platforms. Inconsistencies, including fixed polymorphisms were noted in 60 out of 600 SNPs (representing 10%) when we used KASPar assay. These rather high error rates could come from a number of factors, including incorrect primer synthesis and genotyping errors and merit further investigation. Nevertheless, having parents in both platforms have helped to identify and exclude data from these problematic SNPs.

Before commencement of SNP genotyping, the presence of admixtures and self-progenies was checked by visual inspection of SSR data for unexpected alleles or allelic combinations in conjunction with STRUCTURE software that clusters individuals without a priori pedigree information. This approach helped us to identify unwanted admixtures which arose from pollen cross contamination. Out-crossing, one of the main avenues for cross contamination in cassava, is mediated by insects vectors including honey bees (*Apis mellifera*) that are known to travel several kilometers while foraging for pollen (Pasquet et al. 2008).

The present map had a slightly longer map distance of 1,837 cM compared to other cassava maps and the average marker interval was 3.4 cM. This density is higher than those reported for five existing maps incorporating RFLPs, AFLPs, RAPDs, and SSRs. These earlier maps, with between 100 and 510 markers, are mostly derived from two specific and often-reused crosses: TMS30572 × CM 2177-2 (Fregene et al. 1997; Okogbenin et al. 2006) and Huay Bong 60 × Hanatee (Kunkeaw et al. 2011; Sraphet et al. 2011; Whankaew et al. 2011).

Genetic mapping in cassava has traditionally been carried out in outbreeder full-sib (F1) families due to its heterozygous nature and subsequent inbreeding depression during selfing (Rojas et al. 2009), long growing cycle and low seed number per crossing making it difficult to create F2-based populations. The F1 approach has been used in other perennial and clonally propagated crops such as tea (Hackett et al. 2000), Rhodesgrass (Ubi et al. 2004) and banana (Hippolyte et al. 2010). Since an F1 population in an outcrossing species is derived from two independent meiotic events, parental maps were initially created before being integrated. This approach is referred to as two-step mapping. Additionally, a final one-step linkage map was created using the CP option of JoinMap v.4 and consisted of 568 markers (133 SSRs and 435 SNPs) distributed across 19 linkage groups. Both two-step and one-step mapping approaches produced similar results in the Namikonga × Albert population with only a small number of markers showing positional changes between the two maps. Most of the previous mapping efforts in cassava have reported the one-step method, with the exception of the map of Fregene et al. (1997) where individual parental maps were reported.

Both clusters and individually occurring segregation distortions of loci occurred in the present map. The presence of segregation distortion loci (SDL) is usually considered to cause segregation distortion of neighboring markers. The SDLs may act various phases of plant development (pollen abortion, pollen tube competition, gametic selection, or zygotic selection) (Xian-Liang et al. 2006). Genomic regions with distorted segregation have been observed in many crops, including maize (Lu et al. 2002), rice (Xu et al. 1997), and sorghum (Pereira et al. 1994). In the present mapping population, segregation distortion seems to operate mainly at the gametic level as revealed by predominant significance of the gametic selection tests compared to the zygotic selection tests. The main factor influencing the pattern of selection observed seems to be some form of gametic selection, particularly in the male parent, "Albert", where a specific allele coming from a single haplotype-phase is favored. This was observed in the SSRs located on the distorted region of LG-08 where about six SSRs showed a skewed segregation towards one allele. Predominance of gametophyte selection over zygotic selection has been reported previously by Leppälä et al. (2008) in *Arabidopsis lyrata*. The randomly occurring distorted loci could be explained by the presence of mutations within primer binding sites in the genomic DNA causing preferential amplification of one of the alleles.

Gaps of >10 cM between markers exist in several linkage groups (Fig. 4). These gaps may reflect a lack of polymorphic markers detected in the SNP ascertainment population (i.e. the germplasm from which ESTs were derived) using the stringent SNP detection criteria used or the presence of recombination hot-spot that results in the inflation of genetic distance in a short physical distance. This phenomenon has been reported in other crops such as soybean (Hwang et al. 2009). It is likely that map coverage of markers was not complete leaving some substantial gaps. This would also explain why 19 linkage groups were detected as opposed to the expected 18 linkage groups which would correspond to the haploid chromosome compliment in cassava.

To check map reliability, the order of SSR markers found in the present study and that of Whankaew et al. (2011) was assessed in linkage groups with at least three markers in common. A strong co-linearity was found. This adds confidence to map order particularly as different mapping populations and marker types were used. A total of 40 markers present in both maps will permit the development of a consensus genetic map for cassava.

The genetic linkage map presented here provides an opportunity for improvement of the genome assembly through anchoring of previously unplaced scaffolds or orientating existing scaffolds. To ascertain agreement between the genetic and physical map, we located mapped SNP and SSR markers in the version 4.1 of the draft cassava genome. About 94.2% of the 568 markers searched managed to find suitable hits in 285 unique scaffolds of the draft genome. Nearly identical results were obtained by Sraphet et al. (2011), where the authors reported locating 94.4% of their mapped SSR markers in 284 unique scaffolds. Markers from the same scaffolds mostly occurred in clusters on the same linkage groups, indicating good agreement between the genetic and physical genome assembly (Additional Table 1). However, some scaffolds were sometimes interspersed with markers from different scaffolds, indicating possible inaccurate estimations of genetic distance based on recombination frequency and/or errors in physical genome assembly. Similar observation was reported by Sraphet et al. (2011). Additionally, differences of the genotypes used in two studies could also be a confounding factor. Less than 8% of unassigned SNPs and SSRs belonged to unassembled sequences or represented structural variation between the genotypes used in the mapping population and the cassava genotype used in the genome sequencing project. Structural variation, characterized by the presence–absence of genomic sequences and copy number variation is increasingly being recognized as a significant type of genetic variation even among individuals of the same species (Swanson-Wagner et al. 2010). The described genetic map integrating both EST-derived SNPs and SSRs will augment and complement existing maps of cassava that were developed using AFLPs, RFLPs, RAPDs, SSRs and EST-SSRs markers systems and substantially increase marker density. It is anticipated that this will allow for more accurate QTL detection and selection of markers more even distributed across the genome sequence.

Acknowledgments The authors appreciate the financial support from the BioSciences eastern and central Africa Network (BecANet), the Generation Challenge Program (GCP) and the International Institute of Tropical Agriculture (IITA). We would like to thank Jim Lorenzen of IITA and Steve Rounsley (University of Arizona and Dow AgroSciences) for the fruitful discussions that led to the improvement of the manuscript. We sincerely thank the two anonymous reviewers and the associated editor for their useful comments that has also led to further improvement of this paper.

References

Ahn S, Tanksley SD (1993) Comparative linkage maps of the rice and maize genomes. Proc Natl Acad Sci USA 90:7980–7984

Akano AO, Dixon AGO, Mba C, Barrera E, Fregene M (2002) Genetic mapping of a dominant gene conferring resistance to cassava mosaic disease. Theor Appl Genet 105:521–525

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

Alves AAA (2002) Cassava botany and physiology. In: Hillocks RJ, Thresh MJ, Bellotti AC (eds) Cassava: biology, production and utilisation. CABI International, Oxford, pp 67–89

Anithakumari AM, Tang J, van Eck HJ, Visser RGF, Leunissen JAM, Vosman B, van der Linden CG (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. Methods Mol Biol 513:19–39

Bechsgaard J, Bataillon T, Schierup Mh (2004) Uneven segregation of sporophytic self-incompatibility alleles in *Arabidopsis lyrata*. J Evol Biol 17:554–561

Boonchanawiwat A, Sraphet S, Boonseng O, Lightfoot DA, Triwitayakorn K (2011) QTL underlying plant and first branch height in cassava (*Manihot esculenta* Crantz). Field Crops Res 121:343–349

Buckler ES, Thornsberry JM (2002) Plant molecular diversity and application to genomics. Curr Opin Plant Biol 5:107–111

Chen X, Xia Z, Fu Y, Lu C, Wang W (2010) Constructing a genetic linkage map using an F1 population of non-inbred parents in cassava (*Manihot esculenta* Crantz). Plant Mol Biol Report 28:676–683

CIAT (2003) Annual Report IP3. Improved cassava for the developing world. International Centre for Tropical Agriculture (CIAT), pp 8–90, Cali

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

Edwards D, Forster JW, Cogan NOI, Batley J, Chagne' D (2010) Single nucleotide polymorphism discovery. In: Oraguzie NC, Rikkerink EHA, Gardiner SE, De Silva HN (eds) Association mapping in plants. Springer, New York, pp 53–76

Elshire R, Glaubitz J, Sun Q, Poland J, Kawamoto K et al (2011) A robust simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6:e19379. doi:10.1371/journal.pone.0019379

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software Structure: a simulation study. Mol Ecol 14:2611–2620

Fan JB, Oliphant A, Shen R et al (2003) Highly parallel SNP genotyping. Cold Spring Harb Symp Quant Biol 68:69–78

FAOSTAT (2010) Food and agriculture organizations statistics database. FAO, Rome. http://faostat.fao.org/. Accessed Oct 2010

Ferguson ME, Hearne SJ, Close TJ, Wanamaker S, Moskal WA, Town CD, de Young J, Marri PR, Rabbi IY, de Villiers EP (2011) Identification, validation and high-throughput genotyping of transcribed gene SNPs in cassava. Theor Appl Genetics. doi:10.1007/s00122-011-1739-9

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357–374

Fregene M, Angel F, Gomez R, Rodriguez F, Chavarriaga P, Roca W, Tohme J, Bonierbale M (1997) A molecular genetic map of cassava (*Manihot esculenta* Crantz). Theor Appl Genet 95:431–441

Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS (2005) A genome-wide scalable SNP genotyping assay using microarray technology. Nat Genet 37:549–554

Hackett CA, Wachira FN, Paul S, Powell W, Waugh R (2000) Construction of a genetic linkage map for *Camellia sinensis* (tea). Heredity 85:346–355

Hippolyte I, Bakry F, Seguin M, Gardes L, Rivallan R, Risterucci AM et al (2010) A saturated SSR/DArT linkage map of *Musa acuminata* addressing genome rearrangements among bananas. BMC Plant Biol 10:65

Hwang TY, Sayama T, Takahashi M et al (2009) High-density integrated linkage map based on SSR markers in soybean. DNA Res 16:213–225

Jansson C, Westerbergh A, Zhang J, Hud X, Sun C (2009) Cassava, a potential biofuel crop in (the) People's Republic of China. Appl Energy 86:95–99

Jorge V, Fregene MA, Duque MC, Bonierbale MW, Tohme J, Verdier V (2000) Genetic mapping of resistance to bacterial blight disease in cassava (*Manihot esculenta* Crantz). Theor Appl Genet 101:865–872

Kanju E, Mkamilo G, Mgoo V, Ferguson M (2010) Statistical evidence linking the zigzag stem habit with tolerance to cassava brown streak disease. ROOTS 12:4–6

Kizito BE, Ronnberg-Wastljung AC, Egwang T, Gullberg U, Fregene M, Westerbergh A (2007) Quantitative trait loci controlling cyanogenic glucoside and dry matter content in cassava (*Manihot esculenta* Crantz) roots. Hereditas 144:129–136

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. doi:10.1111/j.1439-0523.2009.01623.x

Kunkeaw S, Yoocha T, Sraphet S, Boonchanawiwat A, Boonseng O, Lightfoot DA, Triwitayakorn K, Tangphatsornruang S (2011) Construction of a genetic linkage map using simple sequence repeat markers from expressed sequence tags for cassava (*Manihot esculenta* Crantz). Mol Breed 27:67–75

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

Leppälä J, Bechsgaard JS, Schierup MH, Savolainen O (2008) Transmission ratio distortion in *Arabidopsis lyrata*: effects of population divergence and the S-locus. Heredity 100:71–78

Lopez CE, Quesada-Ocampo LM, Bohorquez A, Duque MC, Vargas J, Tohme J, Verdier V (2007) Mapping EST-derived SSRs and ESTs involved in resistance to bacterial blight in *Manihot esculenta*. Genome 50:1078–1088

Lu H, Romero-Severson J, Bernardo R (2002) Chromosomal regions associated with segregation distortion in maize. Theor Appl Genet 105:622–628

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Okogbenin E, Fregene M (2003) Genetic mapping of QTLs affecting productivity and plant architecture in a full-sib cross from non-inbred parents in cassava (*Manihot esculenta* Crantz). Theor Appl Genet 107:1452–1462

Okogbenin E, Marin J, Fregene M (2006) An SSR-based molecular genetic map of cassava. Euphytica 147:433–440

Okogbenin E, Marin J, Fregene M (2008) QTL analysis for early yield in a pseudo F2 population of cassava. Afr J Biotechnol 7:131–138

Pasquet RS, Peltier A, Hufford MB, Oudin E, Saulnier J, Paul L, Knudsen JT, Herren HR, Gepts P (2008) Long-distance pollen flow assessment through evaluation of pollinator foraging range suggests transgene escape distances. PNAS 105:13456–13461

Pereira MG, Lee M, Bramel-Cox P, Woodman W, Doebley J, Whitkus R (1994) Construction of an RFLP map in sorghum and comparative mapping in maize. Genome 37:236–243

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Prochnik S, Marri PR, Desany B, Rabinowicz PD, Kodira C, Mohiuddin M, Rodriguez F, Fauquet C, Tohme J, Harkins T, Rokhsar DS, Rounsley S (2011) The cassava genome: current progress, future directions. Tropical Plant Biol. doi:10.1007/s12042-011-9088-z

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

Rojas MC, Pérez JC, Ceballos H, Baena D, Morante N, Calle F (2009) Analysis of inbreeding depression in eight S1 cassava families. Crop Sci 49:543–548

Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. Nat Biotechnol 18:234

Shen R, Fan J-B, Campbell D, Chang W, Chen J, Doucet D, Yeakley J, Bibikova M, Wickham Garcia E, McBride C, Steemers F, Garcia F, Kermani BG, Gunderson K, Oliphant A (2005) High-throughput SNP genotyping on universal bead arrays. Mutat Res 573:70–82

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

Swanson-Wagner RA, Eichtenn SR, Kumari S, Tiffin P, Stein JC, Ware D, Springer NM (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res. doi:10.1101/gr.109165.110

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

Syvänen AC (2005) Toward genome-wide SNP genotyping. Nat Genet 37:5–10

Tavassolian I, Rabie G, Gregory D, Mnejja M, Wirthensohn MG, Hunt PW, Gibson JB, Ford CM, Sedgley M, Wu S (2010) Construction of an almond linkage map in an Australian population Nonpareil × Lauranne. BMC Genomics 11:551

Thresh JM (2006) Control of tropical plant virus diseases. Adv Virus Res 67:245–295

Ubi BE, Fujimori M, Mano Y, Komatsu T (2004) A genetic linkage map of rhodesgrass based on an F1 pseudo-testcross population. Plant Breed 123:247–253

Van Ooijen JW (2006) JoinMap4. Software for the calculation of genetic linkage maps in experimental populations Kyazma BV, Wageningen

Van Ooijen JW, Voorrips RE (2001) JoinMap® version 3.0: software for the calculation of genetic linkage maps. Plant Research International, Wageningen

Whankaew S, Poopear S, Kanjanawattanawong S, Tangphatsornruang S, Boonseng O, Lightfoot DA, Triwitayakorn K (2011) A genome scan for quantitative trait loci affecting cyanogenic potential of cassava root in an outbred population. BMC Genomics 12:266

Wydra K, Zinsou V, Jorge V, Verdier V (2004) Identification of pathotypes of *Xanthomonas axonopodis* pv. *manihotis* in Africa and detection of quantitative trait loci and markers for resistance to bacterial blight of cassava. Phytopathology 94:1084–1093

Xian-Liang S, Xue-Zhen S, Tian-Zhen S (2006) Segregation distortion and its effect on genetic mapping in plants. Chin J Agric Biotechnol 3:163–169

Xu Y, Zhu L, Xiao J, Huang N, McCouch SR (1997) Chromosomal regions associated with segregation distortion of molecular markers in F2, backcross, double haploid, and recombinant inbred populations in rice (*Oryza sativa* L.). Mol Gen Genet 253:535–545