Journal of Root Crops
Indian Society for Root Crops
ISSN 0378-2409, ISSN 2454-9053 (online)
Journal homepage: https://journal.isrc.in

SNP marker development in cassava for cassava mosaic disease resistance using bioinformatics tools

Ambu Vijayan, C. Mohan, M.N. Sheela, and J. Sreekumar*

ICAR-Central Tuber Crops Research Institute, Sreekariyam, Thiruvananthapuram-695017, Kerala, India

Abstract

Cassava (*Manihot esculenta* Crantz), originated in Latin America is one of the most important food crops with a worldwide production of 314.81 million tonnes. The advancements in sequencing ability and less cost involved allow for effective genome-wide discovery of SNPs. The present study was undertaken to computationally develop SNPs for cassava mosaic disease resistance and to understand the effectiveness of molecular markers in cassava for biotic stress response (cassava mosaic virus). The preliminary data set for the work was obtained from the EST section of NCBI (http://www.ncbi.nlm.nih.gov/nucest). The draft cassava genome sequence and transcript sequences (variety AM560-2, JGI annotation v4.1) from the Phytozome website (http://phytozome.jgi.doe.gov/pz/portal.html) were also utilized. The SNP prediction tools, viz., Quality SNP and Auto SNP were used to predictthe SNPs. Quality SNP predicted about 56 SNPs, in which 30 were non-synonymous and 26 were synonymous SNPs. Primers were designed for five selected SNPs associated with CMD resistant genes. These primers were validated using 5 resistant and 5 susceptible cassava genotypes. Among the primers, after validation one SNP (SNP896) primer was able to clearly differentiate between the resistant and susceptible genotypes. This is the first report of SNPs computationally identified and verified in wet lab. The results showed that the sequence with SNP1043 did not show any variation in the predicted SNP site, but SNP896 in the variety, MNga showed SNP at the 1493th position with a variation in the base. The same SNP896 did not show any variance in that position for the susceptible variety CI732.

Keywords: Cassava mosaic disease, SNP marker, Bioinformatics, Molecular markers

Introduction

Cassava, (*Manihot esculenta* Crantz) (2n = 36), which originated in Latin America is an important food crop with a worldwide production of 314.81 million tonnes, with the highest production of 203.57 million tonnes in Africa, followed by Asia (84.25 million tonnes)(FAOSTAT 2022). Cassava is an essential staple food for over 700 million people all over the tropical and sub-tropical regions of the world. It can be grown all year round and provides food in periods of scarcity. Various traits of the crop such as drought tolerance, heat tolerance and less requirement for agricultural fertilizers make it an attractive crop. Cassava has monoecious flowering nature and so self-pollination is mainly prevented by protogynywhich renders the crop highly heterozygous (Alves, 2002). The high starch content (20-40%) makes cassava a desirable energy source both for human consumption and industrial biofuel applications (Schmitz et al., 2009). Cassava is one of the most used raw materials to produce starch. High purity, low production costs, distinctive characteristics like clear viscous paste has made many industries adopt cassava starch as an alternative to more traditional sources like potato and maize.

* Corresponding author:
E-mail: Sreekumar.J@icar.gov.in

Received: 21 March 2022; Revised: 02 May 2022; Accepted: 05 May 2022

Cassava mosaic disease (CMD) is the single most important disease affecting cassava cultivation. Economic losses due to CMD is estimated at US $1.5 billion annually in Africa (Legg et al.,2006; Thresh et al., 1997). CMD is caused by gemini viruses of the genus Begomovirus (Family Geminiviridae) transmitted by a vector, white fly [*Bemisia tabaci, (Gennadius)*]. The begomoviruses represent distinct species such as African cassava mosaic virus (ACMV), East African cassava mosaic virus (EACMV), East African cassava mosaic Cameroon virus, East African cassava mosaic Zanzibar virus and South African cassava mosaic virus (Berrie et al., 2001). The causative agent of CMD in India is Indian cassava mosaic virus, ICMV (Hong et al., 1993). Complete nucleotide sequencing of two cloned ICMV DNAs, one from the state of Kerala and another from Maharashtra showed that they were highly like each other, indicating them to be isolates of the same virus (Saunders et al., 2002).

Availability of Nucleotide sequence information of cassava has effectively helped in the discovery of several genes including disease resistant genes. Expressed sequence tags (ESTs), which are short (300–500 bp) single read sequences from random cDNA clones, have a wide range of applications including the use as gene cloning reservoirs, evaluation of expression of tissue-specific gene, molecular markers for map based cloning and genomic sequence annotation. The EST data have also led to a better understanding of both the existence and expression patterns of alternative transcripts and of coordinated gene expression and itrepresents a potentially significant resource for the detection of single nucleotide polymorphism (SNPs) in plants (Batley et al., 2003). Analysis of ESTs has advanced into an economical and capable gene discovery methodology (Ohlrogge et al., 2003). About 74,316,793 million ESTs are available at the EST database of National Center for Biotechnology Information (NCBI). The genome of cassava is approximately 770 Mb (Awoleye et al., 1994), and the draft genome sequence of cassava was created through the whole genome shotgun strategy. The whole genome of cassava is available in Phytozome, which was developed as part of the global cassava partnership in the year 2003. The cassava genome is predicted to contain 30,666 genes (Prochnik et al., 2012). However, the function of many of the genes remains unclear.

Plant genetic and physical mapping resources along with breeding programes in different agricultural crops led to the development of various plant databases like *Brassica.info, PlantGDB(13), Plaza(14), Ensemblgenomes(16), GSAD(17), NCBI, PGDJ, Phytozome, TAIR, Plant DNA C-values database, Gramene(15), Plant rDNA Database, SGN, The Plant GDB Genome Browser, GDR, LIS, and PTG Base*. These databases consist of genomic data which can be downloaded and used for further analysis and they provide a set of automated analysis tools within the web portals.

In plant genetic research, molecular markers are used for effective marker assisted selection, population structure analysis, evolutionary relationship study and whole genome studies. Molecular markers such as Single Nucleotide Polymorphisms (SNPs) and Single sequence repeats (SSRs) has high potential to help plant breeders. SNPs are markers of choice for high-density genetic mapping due to their sheer abundance in the genome (Rafalski, 2002). SNPs are known to occur at a rate of one per 100-500 bp in plant genomes, depending on the species. The advancements in sequencing ability along with the savings in sequencing cost allow for effective genome-wide discovery of SNPs. RNA-seq has been successfully applied to large-scale SNP discovery and EST- derived SNP development in various plant species (Ferguson et al., 2012; Paritosh et al., 2013). The present study was aimed at developing computationally predicted SNP makers for cassava mosaic disease resistance. The SNP development tools were also evaluated to understand their performance.

Materials and Methods

The detailed workflow used for the SNP prediction is given in Fig. 1.

Primary cassava dataset

The preliminary data set which consists of a total of 86310 ESTs was obtained from the EST section of NCBI (http://www.ncbi.nlm.nih.gov/nucest) and the transcript sequences consisting of 34151 sequences of cassava (variety AM560-2, JGI annotation v4.1) was downloaded from the Phytozome website (http://phytozome.jgi.doe.gov/pz/portal.html). Together, a total of 1,20,461 sequences were taken as the primary dataset. The sequences were pre-processed using the Seq Clean script (http://sourceforge.net/projects/seqclean/files/) with the default runtime options for eliminating contamination or simple repeats. Vector sequences in these ESTs were then trimmed using the UniVec_Core database (http://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/) of NCBI. Sequences with more than 96 percentage similarity to a contaminant was removed. Based on cultivars, the sequences were classified into 19 categories and one category with unclassified sequences. All the phytozome transcript sequences used were of a cassava cultivar named Am560-2 with 34151 sequences. Most sequences in NCBI were from MTai-16 with 35400 sequences and sequences with the least number was of the cultivar H-226 which had only 21 sequences.

Virus resistant gene database

Virus resistant gene database including CMD resistant gene database was manually created and compiled from uniprot KB. The UniProt Knowledgebase (UniProt KB) is the central access point for extensively curated protein information, including function, classification,

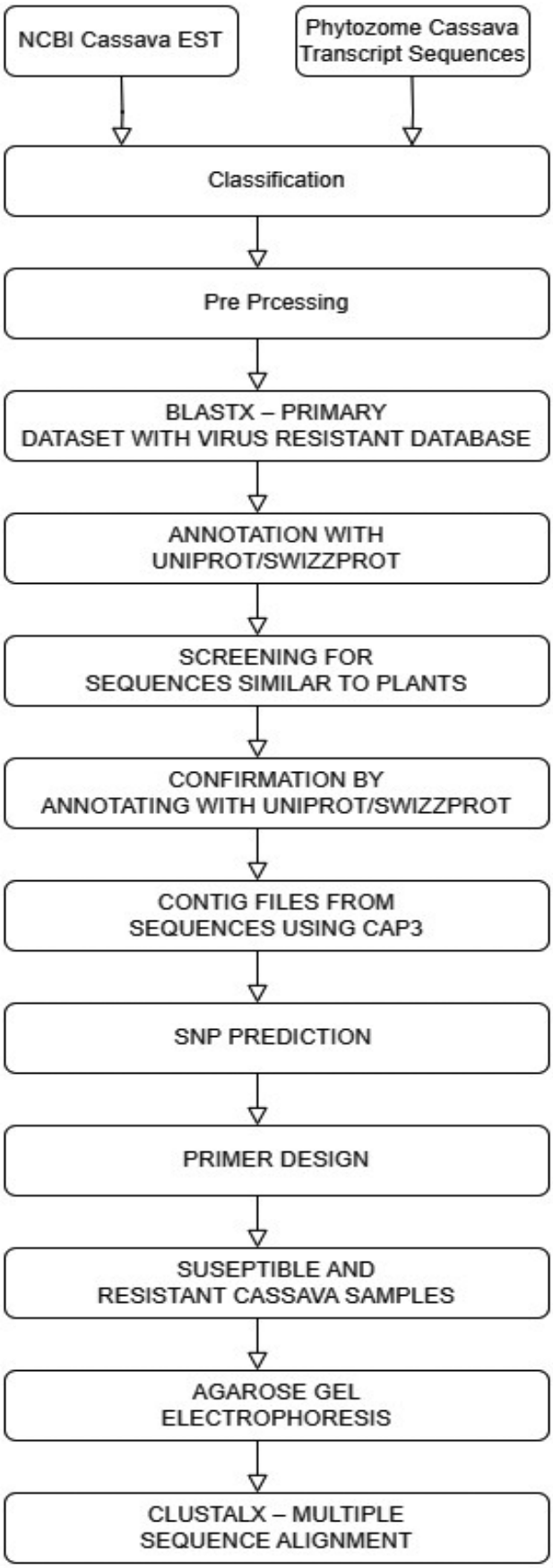

Fig. 1. Workflow for the prediction and validation of SNP

and cross-references. R-gene or resistant genes related to cassava and mosaic diseases was screened from it and were used for database creation. The virus resistance protein database consisted of 730 resistant genes.

Processing of primary dataset

For screening of primary dataset with virus resistant protein sequences, 'BlastX- Search protein database using a translated nucleotide query' was employed. Klast was used for doing sequence comparison (http://koriscale.inria.fr/). The cassava ESTs and transcript sequences were screened against resistant genes using BLASTX. Resulting cassava ESTs and transcript sequences were annotated using Uniprot/SwissProt database and only sequences which have functional annotations were retained. Nucleotide sequenceswith similarity to organisms other than plants were eliminated. The resulting screened sequence datasetwas used for DNA polymorphism discovery.

DNA polymorphism studies

Assembling of the screened sequence datasetof sequences was carried out using the Perl script CAP3 program (http://seq.cs.iastate.edu/CAP3.html) with default runtime options. SNP and InDelpolymorphisms were discovered from the contigs obtained. Quality SNP pipelinewas used for the discovery of SNP and InDels (http://www.bioinformatics.nl/tools/snpweb/). Analysis of the alignment information to select cluster size of four was done using the Perl script 'Getalignmentinfo'. The Cprograms 'Getavailcontigseq' and 'Getavailcontigqual' extracted the sequences from the contigs and delivered the quality information of contigs. Using C program 'QualitySNP' predicted SNPs and InDels. Another C program named 'Getnonsy SNP fasty' was used to analyze the FASTY results, detect the ORFs and find non-synonymous SNPs. For the analysis of non-synonymous SNPs, Viridiplantae database was used(tp://141.161.180.197/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/uniprot_sprot_plants.dat.gz).

Verification in wet lab

Primer designing for predicted SNPs

Primer pairs were designed to amplify the genomic region around each discovered SNP site. Sequences were selected for primer designing based on the hit percentage of contigs containing SNP with the resistant genes. SNP containing contigs with a hit percentage between 80-100% were selected. Primer pairs were designed using Primer3plus tool with the parameters set as (i) GC content above 50% and (ii) Melting temperature between 55 and 60°C.

Plant materials

A total of 10 cassava varieties which included 5 CMD resistant and 5 susceptible varieties were selected based

on field trials conducted at ICAR-Central Tuber Crop Research Institute (CTCRI), Thiruvananthapuram. Fresh young leaves were collected, and DNA was isolated from these leaf samples using the method described by Dellaporta et al., (1983) with some modifications. The concentration and purity of all the DNA samples was determined using a UV spectrophotometer by taking absorbance at 260 nm. The amount of DNA was quantified using the following formula:

$$\text{DNA concentration } (\mu g\ ml^{-1}) = \frac{OD260 \times 100 \text{ (dilution factor)} \times 50}{1000}$$

According to the reading obtained after quantification, genomic DNA was diluted to a concentration of 50 ng μl^{-1} and stored at 4°C. The stock DNA was then stored in -20°C.

Amplification of designed primers

Primer amplification was done in a BioRad C1000™ thermal Cycler with respective parameters for SNP and SSR primers.

Validation of SNP markers

Validation of SNP markers was done by running the marker with the DNA isolated from the five resistant and five susceptible cassava varieties in agarose gel electrophoresis and then eluting the bands, sequencing it, and comparing it with the reference genome of cassava. Reference genome of cassava is available in the Phytozome database. ClustalX was used for aligning the sequences and to validate the SNP.

Results and Discussion

Cassava dataset and screening process

The preliminary data consisted of 86310 sequences from the EST section of NCBI and 34151 transcript sequences from the Phytozome database. From 19 cassava cultivars (arg7, Cas 36.04, Cm 523-7, cm 21772, crantz, Iac 12.829, ku 50, mper183, mbra 685, mcol 1522, h 226, mirassol, mnga 2, mtai 16, 'Sauti, Gomani, Mbundumali, TME1 and Mkondezi', g 107-35, w 14, Cas 36.01 and Am 560-2), 97921 sequences and from unknown varieties, 22540 sequences were obtained. The sequences were cleaned for contaminants and a total of 120398 sequences were obtained as the starting data for the study. The primary dataset was screened against virus resistance gene database using BLASTX and 16299 sequences were obtained with similarity to R-genes. About 86% *i.e.*, 104099 sequences were screened out by this process.The sequence with similarity to R-genes were annotated using Uniprot/Swissprot database and the number of sequences after annotation was 15796 and about 3% *i.e.*, 503 sequences were screened out. About 1460 sequences were eliminated because of the presence of sequence similarity to organisms other than plants.

DNA polymorphism study

The sequences after screening were aligned and assembled using cap 3. 2088 contigs and 5236 singlets were obtained from 14336 sequences with similarity to virus resistant genes. Quality SNP was able to identify 128 SNPs. From 2088 contigs, a total of 3297227 sequences were examined. In this study, about 204 SNPs were predicted which are exclusievely related to CMD resistance in cassava. These can be validated and screened for effective markers against CMD resistance. More than 56 SNPs were confirmed in the coding region which makes them candidate SNPs for screening for resistance against CMD. More than 30 SNPs were nonsynonymous which can result in change in the transcription product.

A total of 121 SNPs were identified using Quality SNP. The total number of transitions (67) was marginally greater than the total number of transversions (54) yielding a transition-to-transversion ratio of 1.24 (Table 1). Based on the annotation data, these SNPs were classified as SNPs in coding region, non-coding region and in untranslated region. About 56 SNPs were found in the coding region and these can alter proteins. About 65 SNPs were predicted from non-coding region and 76 from untranslated region. Based on the type of SNPs, these were further classified into synonymous SNPs and nonsynonymous SNPs. About 30 SNPs were nonsynonymous SNPs (Table 2), which means that they will effect a change in the translated protein. About 26 SNPs were synonymous (Table 3), *i.e.*, the mutations will not cause any change in the system. Again, based on the type of polymorphism these SNPS can be classified into SNPs and InDels. About 72 SNPs and 56 InDels are obtained.

Table 1. Distribution of transition and transversion SNPs from QualitySNP

Characterization	Type of SNP	SNPs	Total
Transition	C/T	33	67
	G/A	34	
Transversion	A/C	14	54
	A/T	11	
	C/G	17	
	T/G	12	

Quality SNP showed more promising SNPs than Auto SNP, where a huge number of SNPs including false positive SNPs were also predicted. Quality SNP showed unique ability to annotate and classify SNPs based on their polymorphism, the type of annotation data and the type of SNP. However, in Auto SNP, classification was entirely based on the type of SNPs. Quality SNP gave a more detailed and precise information whereas Auto SNP predicted thousands of SNPs with difficulty in identifying the viable ones from the enormous list of identified SNPs (Table 4).

Table 2. List of Nonsynonymous SNP coding data identified by QualitySNP

Contig no.	Position	SNP	Length	Normal sequence	Sequence with base change	Transcribed protein	
260	388	TC	10	CACCAGAATTTATCATCAAGC	CACCAGAATTCATCATCAAGC	HQNLSSS	HQNSSSS
344	509	AT	10	AAATCAGCTTATGCATTGTGT	AAATCAGCTTTTGCATTGTGT	KSAYALC	KSAFALC
385	683	GC	10	AACAGTGAGAGCAAACAAGAG	AACAGTGAGACCAAACAAGAG	NSESKQE	NSETKQE
401	630	GT	11	TTGCGCAAGCAGTACGGACCT	TTGCGCAAGCATTACGGACCT	LRKQYGP	LRKHYGP
401	833	GC	10	CGGAATCCAAGGAAAAGGCTA	CGGAATCCAACGAAAAGGCTA	RNPRKRL	RNPTKRL
401	836	GA	10	AATCCAACGAGAAGGCTATCA	AATCCAACGAAAAGGCTATCA	NPTRRLS	NPTKRLS
468	1143	AG	9	GCTGCATTCAATATGCCACCC	GCTGCATTCGATATGCCACCC	AAFNMPP	AAFDMPP
732	82	CA	11	GTTCAATCTCACCCCAGAAGC	GTTCAATCTCAACCCAGAAGC	VQSHPRS	VQSQPRS
896	1495	CA	10	GTGCTATATACGCACCCAGCA	GTGCTATATAAGCACCCAGCA	VLYTHPA	VLYKHPA
1043	635	CT	10	TCTCAAACAACGATTTATGTG	TCTCAAACAATGATTTATGTG	SQTTIYV	SQTMIYV
1053	1044	TG	11	TGTCAGGGAGATTATGTGGTG	TGTCAGGGAGAGTATGTGGTG	CQGDYVV	CQGEYVV
1073	76	CT	10	CGTGAACAACCTCCCTCCATC	CGTGAACAACTTCCCTCCATC	REQPPSI	REQLPSI
1073	79	TC	10	GAACAACCTCTCTCCATCCTC	GAACAACCTCCCTCCATCCTC	EQPLSIL	EQPPSIL
1073	126	TC	9	TTTGGCTCTTTTTCTCCCTTG	TTTGGCTCTCTTTCTCCCTTG	FGSFSPL	FGSLSPL
1228	2528	AG	10	TACAGCATCGAACTTCCAAGC	TACAGCATCGGACTTCCAAGC	YSIELPS	YSIGLPS
1238	415	AG	9	TTTCTCGTGATTTTGCTTTTG	TTTCTCGTGGTTTTGCTTTTG	FLVILLL	FLVVLLL
1889	668	AG	10	ACACCCGGCCAGGAATTTACT	ACACCCGGCCGGGAATTTACT	TPGQEFT	TPGREFT
1889	685	AG	9	ACTTTTACAATTCGTAGGGGA	ACTTTTACAGTTCGTAGGGGA	TFTIRRG	TFTVRRG
1889	881	GA	10	CTAAATGTTAGAGGAAAAAGC	CTAAATGTTAAAGGAAAAAGC	LNVRGKS	LNVKGKS
1930	1379	AC	9	GAGGTTAGTAACCTTACAGCC	GAGGTTAGTCACCTTACAGCC	EVSNLTA	EVSHLTA
2023	574	GT	9	AGCTACACTGTGGCTTATGGA	AGCTACACTTTGGCTTATGGA	SYTVAYG	SYTLAYG
2023	602	CG	10	CCAGAACCTACTTGTCCTTGT	CCAGAACCTAGTTGTCCTTGT	PEPTCPC	PEPSCPC
2055	1540	CT	10	AAAAAATATGCTGAGGTTCTT	AAAAAATATGTTGAGGTTCTT	KKYAEVL	KKYVEVL
2055	1560	GC	9	AGACTGATAGGGAGACTTACG	AGACTGATACGGAGACTTACG	RLIGRLT	RLIRRLT
2055	1563	AG	9	CTGATAGGGAGACTTACGTTG	CTGATAGGGGGACTTACGTTG	LIGRLTL	LIGGLTL
2055	1617	GC	9	CAAGACTCCGAGCTAGACCAA	CAAGACTCCCAGCTAGACCAA	QDSELDQ	QDSQLDQ
2055	1680	GA	9	AGTCTGGTTGCTTTAGCACCA	AGTCTGGTTACTTTAGCACCA	SLVALAP	SLVTLAP
2055	1725	GA	9	ATCACGTTGGAAGTGTTGAAA	ATCACGTTGAAAGTGTTGAAA	ITLEVLK	ITLKVLK
2055	1987	TA	10	GTAACTGTGATGCAATGCCCC	GTAACTGTGAAGCAATGCCCC	VTVMQCP	VTVKQCP
2064	625	GC	9	TCAAATCAGGCTTCAGTTACT	TCAAATCAGCCTTCAGTTACT	SNQASVT	SNQPSVT

Table 3. List of Synonymous SNP coding data identified by Quality SNP

Contig no.	Position	SNP	Length	Normal sequence	Sequence with base change	Transcribed protein
361	358	GA	11	GCTAACCTGAGGCGCGCTGCT	GCTAACCTGAGACGCGCTGCT	ANLRRAA
361	454	CG	11	AGGCAGTTTCTCGGGCTGAGG	AGGCAGTTTCTGGGGCTGAGG	RQFLGLR
361	1053	AT	11	TACGGTTCGGCAGGCTATGCT	TACGGTTCGGCTGGCTATGCT	YGSAGYA
361	1189	TC	11	TCCATGGTGTCTACTGTTGCT	TCCATGGTGTCCACTGTTGCT	SMVSTVA
401	591	CA	11	GATGTTGTTGGCAGTCCATAC	GATGTTGTTGGAAGTCCATAC	DVVGSPY
401	609	AG	11	TACTATGTCGCACCAGAGGTG	TACTATGTCGCGCCAGAGGTG	YYVAPEV
401	618	GA	11	GCACCAGAGGTGTTGCGCAAG	GCACCAGAGGTATTGCGCAAG	APEVLRK
401	633	CT	11	CGCAAGCAGTACGGACCTGAA	CGCAAGCAGTATGGACCTGAA	RKQYGPE
401	678	TC	11	ATTTTGTATATTTTATTATCT	ATTTTGTATATCTTATTATCT	ILYILLS
401	699	AT	11	GGAGTGCCACCATTTTGGGCA	GGAGTGCCACCTTTTTGGGCA	GVPPFWA
401	837	GA	11	AATCCAACGAAGAGGCTATCA	AATCCAACGAAAAGGCTATCA	NPTKRLS
463	141	CT	11	GGAAAGTCGACCACTACTGGT	GGAAAGTCGACTACTACTGGT	GKSTTTG
468	1115	AT	11	ATTTCTACAGGAGCCTTCCTT	ATTTCTACAGGTGCCTTCCTT	ISTGAFL
567	427	CT	11	CCAAAGAAGACCGGCACCTCA	CCAAAGAAGACTGGCACCTCA	PKKTGTS
896	1490	AT	11	GAAAATGTGCTATATACGCAC	GAAAATGTGCTTTATACGCAC	ENVLYTH
899	1299	TC	11	CCCGAGCTTGTTAACAAGCTG	CCCGAGCTTGTCAACAAGCTG	PELVNKL
1136	285	CT	11	AAAATCAGAACCGTGGAGCTG	AAAATCAGAACTGTGGAGCTG	KIRTVEL
1228	2604	GA	11	AAGTCATTCACGTGTACTTTA	AAGTCATTCACATGTACTTTA	KSFTCTL

1233	636	TC	11	GTTTATAAGATTGAAGCTGAA	GTTTATAAGATCGAAGCTGAA	VYKIEAE
1889	608	CT	11	ATGCTTGACACCAAGGGTCCT	ATGCTTGACACTAAGGGTCCT	MLDTKGP
1889	813	AT	11	AAGTCCAAGACAGATGACTCT	AAGTCCAAGACTGATGACTCT	KSKTDDS
1889	912	TG	11	CCTTCCATCACTGAAAAGGAC	CCTTCCATCACGGAAAAGGAC	PSITEKD
2023	369	AG	11	GCTATGTTGTCACGCTCTGCG	GCTATGTTGTCGCGCTCTGCG	AMLSRSA
2023	378	GT	11	TCACGCTCTGCGGCAGGAATA	TCACGCTCTGCTGCAGGAATA	SRSAAGI
2055	1724	GA	11	CTAATCACGTTGGAAGTGTTG	CTAATCACGTTAGAAGTGTTG	LITLEVL
2055	1757	GA	11	TTACTTAGTCTGGTAACATCT	TTACTTAGTCTAGTAACATCT	LLSLVTS

Table 4. Comparative study of SNPs from Auto SNP and Quality SNP

Type of polymorphism	No. of polymorphisms in Auto SNP	No. of polymorphisms in Quality SNP
Transition	8827	67
Transversion	6840	54
Indels	2414	72
Total	18081	193

A similar computational analysis of SNP was carried out by Sakurai et al., (2013). Polymorphisms (SNPs and InDels) were discovered from the contig sequence alignment according to different criteria, Prochnik et al., 2012 fixed the criteria that at the contig should be able to align with the cassava draft genome sequence. Other criteria followed by researchers include that there were fewer than 3 other discontinuous nucleotide polymorphisms around 5 bp of a SNP site (Sakurai et al., 2013).

Validation of SNPs

Table 5. Genotypes used for validation of SNPs associated with CMD

Sl. No.	Resistant	Susceptible
1	Albert	CI732
2	96/1089A	CO2
3	Cr 11/43	Ambakadan
4	TME-3	Sree Vijaya
5	MNga-1	Sree Jaya

The 10 cassava genotypes/varieties used for validation of SNPs associated with CMD are presented in Table 5. Of the selected 5 SNP markers for primer synthesis, only forward primers were fluorescent labelled. Four different fluorescent dyes *viz*., 6-FAM, NED, VIC, PET were used. Validation of SNP was done by eluting the separated bands from the gel and then sequencing it. This sequence was aligned with the corresponding contig sequence from which the respective primer was designed. Multiple sequence alignment was done using ClustalX. The bands were eluted from the gel using the elution kit and were analyzed using 3500 capillary DNA Genetic Analyzer (Applied Biosystem). Three replicate analyses were carried out to avoid sequencing errors. These sequences were then aligned against their respective contigs using ClustalX. Sequence bands from the resistant genotype, MNga and susceptible genotype CI732 which contains the designed primers SNP896 and SNP1043 were sequenced. These sequences were aligned against contig 896 and contig1043 from which the primers were designed. ClustalX was used for multiple sequence alignment. The results showed that the sequence with SNP1043 did not show any variation in predicted SNP site, whereas SNP896 in MNga showed SNP at the 1493th position as designed but with a variation in the base. SNP896 did not show any change in that position for the susceptible variety CI732.

Conclusion

The study was aimed at developing molecular marker development, especially SNPs for cassava mosaicdisease resistance using bioinformatics tools and its validation. The preliminary data set for the identification of SSR/ SNP markers wasobtained from the EST section of NCBI (http://www.ncbi.nlm.nih.gov/nucest) andthe cassava transcript sequences (variety AM560-2, JGI annotation v4.1) from the Phytozome website (http://phytozome.jgi.doe.gov/pz/portal.html). With the help of these SNP prediction tools, we were able to develop novel markers which can be used for differentiating CMD resistant and susceptible genotypes. Primers were designed for both SNPs for CMD resistant genes, these primers were validated using 5 resistant and 5 susceptible cassava genotypes/varieties. Among the primers validated one SNP (SNP896) was able to clearly differentiate between the resistant and susceptible varieties. This is the first report of SNPs and SSRs computationally identified and verified in wet lab. In future, the identified 56 SNPs and 537 SSRs can be validated in wet lab and the resultingpotential markers can be utilized in the breeding program for screening CM Dresistance in cassava.

References

FAOSTAT 2014.http://www.fao.org/faostat/en/#data/QC

Alves AAA (2002) Cassava botany and physiology. In: *Cassava: biology, production and utilisation*. Hillocks R.J., Thresh M.J. and Bellotti A.C. (Eds.). CABI International, Oxford, pp: 67-89

Schmitz, P.M., Kavallari, A. 2009. Crop plants versus energy plants--on the international food crisis. *Bioorg. Med.Chem.,* **17**(12):4020-1.

Legg J.P., Owor B., Sseruwagi P., Ndunguru J. (2006) Cassava mosaic virus disease in east and central Africa: epidemiology and management of a regional pandemic. *Adv. Virus Res.,* **67**:355-418

Thresh J.M., Otim-Nape G.W., Legg J.P., Fargette D. (1997) African cassava mosaic virus disease: the magnitude of the problem. *AJRTC,* **2**:13-19

Berrie, L.C., Rybicki, E.P., Rey, M.E. 2001. Complete nucleotide sequence and host range of South African cassava mosaic virus: further evidence for recombination amongst begomoviruses. *J.Gen. Virol.,* **82**(Pt 1):53-8.

Hong, Y.G., Robinson, D.J., Harrison, B.D. 1993. Nucleotide sequence evidence for the occurrence of three distinct whitefly-transmitted geminiviruses in cassava. *J.Gen. Virol.,* **74** (Pt 11):2437-43.

Saunders, K., Salim, N., Mali, V.R., Malathi, V.G., Briddon, R., Markham, P.G., Stanley, J. 2002. Characterisation of Sri Lankan cassava mosaic virus and Indian cassava mosaic virus: evidence for acquisition of a DNA B component by a monopartite begomovirus. *Virology,* **293**(1):63-74.

Batley, J., Barker, G., O'Sullivan, H., Edwards, K.J., Edwards, D. 2003. Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiol.*, **132**(1):84-91.

Ohlrogge, J., Benning, C. 2000. Unraveling plant metabolism by EST analysis. *Curr. Opin. Plant Biol.*, **3**(3):224-8.

Awoleye F., Duren M., Dolezel J., Novak F.J. (1994) Nuclear DNA content and in vitro induced somatic polyploidization cassava (Manihot esculenta Crantz) breeding. *Euphytica,* **76**:195-202.

Prochnik, S., Marri, P.R., Desany, B., Rabinowicz, P.D., Kodira, C., Mohiuddin, M., Rodriguez, F., Fauquet, C., Tohme, J., Harkins, T., Rokhsar, D.S., Rounsley, S. 2012. The Cassava Genome: Current Progress, Future Directions. *Trop. Plant Biol.,* **5**(1):88-94.

Rafalski, A. 2002. Applications of single nucleotide polymorphisms in crop genetics. *Curr.Opin.Plant Biol.*, **5**(2):94-100.

Paritosh, K., Yadava, S.K., Gupta, V., Panjabi-Massand, P., Sodhi, Y.S., Pradhan, A.K., Pental, D. 2013. RNA-seq based SNPs in some agronomically important oleiferous lines of Brassica rapa and their use for genome-wide linkage mapping and specific-region fine mapping. *BMC Genom.*, **14**:463.

Ferguson, M.E., Hearne, S.J., Close, T.J., Wanamaker, S., Moskal, W.A., Town, C.D., de Young, J., Marri, P.R., Rabbi, I.Y., de Villiers, E.P. 2012. Identification, validation and high-throughput genotyping of transcribed gene SNPs in cassava. TAG. Theoretical and applied genetics. *Theoretische und angewandte Genetik*, **124**(4):685-95.

Sakurai, T., Mochida, K., Yoshida, T., Akiyama, K., Ishitani, M., Seki, M., Shinozaki, K. 2013. Genome-wide discovery and information resource development of DNA polymorphisms in cassava. *PloS One*, **8**(9):e74056.