Development and Characterization of Microsatellite Markers, Genetic Diversity and Population Structure Analysis in Sapota (*Manilkara Zapota* (L.) P. Royen)

Kundapura Venkataramana Ravishankar (kv_ravishankar@yahoo.co.in)
Indian Institute of Horticultural Research https://orcid.org/0000-0001-9935-763X

Pavithra Sathanandam
IIHR: Indian Institute of Horticultural Research

Prakash Patil
IIHR: Indian Institute of Horticultural Research

Ajitha Rekha
IIHR: Indian Institute of Horticultural Research

Iyyamperumal Muthuvel
Tamil Nadu Agricultural University

Amrutlal Patel
China Agricultural University

Ramesh Boggala
APHU: Dr YSR Horticultural University

Adiveppa Shirol
Hue University of Sciences: Hue University University of Sciences

Research Article

Keywords: Manilkara sapota, Illumina sequencing, Microsatellites, Polymorphic Information Content, Genetic diversity, Population structure

Posted Date: July 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-609051/v1

Abstract

Manilkara zapota (L.) P. Royen, a widely adaptable and popular tree meant for its appetizing fruits in tropics with no genomic resources like microsatellite markers. In order to develop genomic markers primarily for sapota, we sequenced partial genomic DNA using next generation sequencing technology on the Illumina HiSeq 2500 platform. We analyzed a total of 3.3 Gb data that were assembled into 6396224 contigs. From these contigs, 3591 simple sequence repeats were identified. Among the different type of repeats mononucleotide repeats (59.1%) were predominant followed by dinucleotide (28.6%) and trinucleotide repeats (8.2%). Primers were designed for 1285 microsatellite regions from which 30 randomly selected primers were standardized and employed for amplification in 53 genotypes of sapota. We observed 692 alleles from 30 loci with a polymorphic information content ranged from 0.85 to 0.96 with a mean of 0.9118. The probability of identity ranged from 0.002 to 0.043 with a mean of 0.012. Genetic diversity assessed by neighbour-joining and STRUCTURE assignment tests showed admixed population with 3 groups. Analysis of molecular variance revealed a significant F st value of 0.69659 indicating high genetic differentiation among the 53 genotypes. The developed microsatellites will be advantageous in assessing genetic diversity, developing linkage map and also molecular characterization of genotypes

Introduction

Sapota (*Manilkara zapota* (L.) P. Royen) is a delectable fruit of tropical and sub-tropical region belongs to the family Sapotaceae. It is native to Central America and has now spread to many tropical countries. In Asia, it was first introduced to Philippines by the Spanish and later spread to other Asian countries (Meghala et al. 2005). In India it is commonly known as “chikku” and was introduced in the 1800’s from Mexico via Sri Lanka (Rekha et al. 2011). The hardy nature of the tree and its wide range of adaptability to different conditions have made this tree popular in India, and it is mainly meant for its fruit value. In some countries, it is commercially grown for a gum like substance chickle, extracted from the latex of unripe fruit which is used in the preparation of chewing gum. The fruit has many health beneficial ingredients in sufficient quantities such as dietary fibres, sugars like fructose and sucrose, phenolics viz., gallic acid, catechin, chlorogenic acid, leucodelphinidin, leucocyanidin and leucopelargonidin, carotenoids, ascorbic acid, minerals like potassium, calcium, phosphorous and iron, and antioxidant compounds (Siddiqui et al. 2014; Rastegar 2015).

Now DNA markers are widely used in genetic studies especially in diversity analysis, DNA fingerprinting, mapping, identification of genes, disease diagnostics, pedigree analysis, hybridity confirmation, identification of sex types and marker assisted breeding (Bhat et al. 2010).Markers have become part of crop improvement program.

Microsatellites, also known as Simple Sequence Repeats, are part of genomic sequences with a special characteristics of tandem repeats of short nucleotide motifs (1 to 6 bp) (Ellegren 2004). SSRs are widely used in plant genetic research as they are co-dominant, with multi-allelic nature, high reproducibility and high polymorphic information content. Earlier, microsatellites detection and isolation has been based on enrichment of genomic libraries by selective hybridization or by primer extension, another approach is to identify microsatellite repeats in DNA databases such as EST sequences. Currently Next Generation Sequencing (NGS), a robust revolutionizing technology is used for microsatellites development (Ravishankar et al. 2015b, c; Araya et al. 2017).

In this study, we used next generation sequencing technology to partially sequence sapota genome which is promising approach to generate high throughput genome data at reduced cost and at lesser time. Further this helps in the detection of thousands of microsatellite sites in the genome of target species. The generated data was used to identify and standardize microsatellite markers in Sapota. Further we analyzed genetic diversity and population structure of sapota germplasm available at ICAR - Indian Institute of Horticultural Research, Bengaluru, India.

Materials And Methods

Plant material and DNA extraction

Fifty three genotypes of Sapota (supplementary material 1) were utilized in this study. Young leaf samples were collected from the sapota germplasm collection maintained at ICAR - Indian Institute of Horticultural Research, Bengaluru, India. Total genomic DNA was isolated from the leaves using modified CTAB method (Ravishankar et al. 2000). The quality of extracted DNA was examined using agarose gel electrophoresis (0.8%) and the concentration of DNA was determined by UV- spectrophotometer (Gene Quanta, Amersham Biosciences) at 260/280 nm.

Next Generation sequencing and de novo assembly

Total genomic DNA of variety ‘Cricket ball’ was used to perform paired end HiSeq Illumina sequencing using 2500 platform following manufacturer’s protocol (QTLomics, Technologies P Ltd. Bengaluru). The FastQ files containing raw data were submitted to sequence read archive at National Centre for Biotechnology Information (NCBI) with accession ID SRP127995. The quality of paired end data was checked using the FastQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) at the default parameters followed by quality trimming of the reads with trimmomatic program (Bolger et al. 2014). High-quality data was used for analysis i.e. bases having Phred score ≥ 40. Filtered reads were used to get optimal k-mer for assembly using VelvetOptimizer (Zerbino and Birney 2008). Based on the optimization results assembly was performed using velvet using k-mer 31. The assembled contigs with a length longer than 1kb were taken for SSR mining.

Microsatellites mining and primer designing

A Perl based SSRs motifs scrutinizing tool, MISA software was used for the identification and localization of microsatellites (http://pgrc.ipk-gatersleben.de/misa/). The SSR contigs were screened for dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, hexanucleotide and complex repeats. The primers were designed for the flanking regions of SSR contigs using a web based application Batchprimer3 (You et al. 2008).

From the generated primers, 75 primer pairs were chosen and tailed with M13 sequence 'GTAAAACGACGGCCAGT' at the 5´ end of forward primers and sequence 'GTTTCTT' at the 5´ end of reverse primers. The M13 sequences were labelled with four dyes FAM, VIC, NED and PET at their 5´ end and used as probes (Schuelke 2000).

PCR and SSRs validation

PCR conditions were standardized with a 15 μl volume of reaction containing 50–100 ng of DNA, 1.5 μl of 10X *Taq* Buffer, 0.8 μl of 10 mM dNTPs, 0.5 μl of 5 μM locus-specific forward primer, 1.0 μl of 5 μM reverse primer, 1.0 μl of 5 μM M13 probe and 0.5 unit of *Taq* DNA polymerase (Genei). PCR was carried out in BioRad Thermocycler with the following temperature profile: 94°C for 4 minutes followed by 35 cycles at 94^{o} C for 45 seconds, an annealing temperature of 55°C for 45 seconds, and 72°C for 1 minute. A final extension reaction was proceeded at 72°C for 7 minutes. For confirmation, the amplified products were separated on 2% agarose gel. Then the PCR samples were mixed by combining four PCR products, labelled with four different fluorophores (FAM, VIC, NED and PET) into a single tube. These samples were separated on the 96 capillary-automated DNA Sequencer (Applied Biosystem Company, USA) at M/SBioserve facility, Hyderabad.

Data analysis

The raw data generated through genotyping were analyzed using GeneMarker V2.6.3 software (Softgenetics, LLC) for detecting the fragment size of the alleles. The generated fragment size data was used for genetic analysis using CERVUS 3.0 software (Kalinowski et al. 2007) for calculating the number of alleles (k), observed heterozygosity (H_{obs}), expected heterozygosity (H_{exp}) and polymorphic information content (PIC). The probability of identity (PI) were also calculated using IDENTITY 1.0 software (Wagner and Sefc 1999). The genetic relationship among 53 sapota genotypes were analyzed and examined by dendrogram analysis using neighbour-joining method with DARwin 6 software (Perrier and Jacquemoud-Collet 2006). Further to confirm, the genetic structure analysis was inferred with STRUCTURE V2.3.4 software (Kaeuffer et al. 2007) with a data set of 30 SSR markers evaluated. An assignment test of individuals was carried out using admixture model with 200,000 burn-ins and 200,000 iterations and a total of 10 independent simulations were run for each value of k tested ranging from K=1 to K=20. Further delta k values using Evanno method and graphical representation were deduced using the STRUCTURE HARVESTER program (Earl and VonHoldt 2012; http://taylor0.biology.ucla.edu/structureHarvester/) and STRUCTURE PLOT (Ramasamy et al. 2014; http://omicsspeaks.com/strplot2/) respectively. AMOVA analysis was performed based on the population clusters obtained by STRUCTURE analysis using Arlequin V 3.5 software with 10000 permutations (Excoffier et al. 2005). AMOVA derived genetic differentiation values (F_{ST}) between pairs of populations analogous to F-statistics were calculated.

Results

Genome sequencing and de novo assembly

The Next Generation Sequencing technology Illumina HiSeq 2500 platform was used to sequence genomic DNA from cultivar Cricket ball. The sequencing run yielded 3,326,257,143 bases from 22,028,193 reads. Low quality reads were filtered and finally 21654953 (98.31%) paired end reads were obtained. Assembly optimization was done with k-mer 31 as it had optimal readings for N50 which generated 6396224 contigs in the assembly. The total length of assembled contigs was 839591847 bp. A total of 16359 contigs had sequences longer than 1000 bp. The longest contig had a length of 11136 bp and minimum length of the contig was 61 bp. The average GC content of the genome was 40.91%.

Microsatellites Mining

A total of 16416 sequences were examined out of which 2591 sequences contained SSRs. Among the 2591 sequences a total of 3591 SSRs were identified. There were 710 sequences containing more than one SSR and 252 SSRs were found in compound formation. Among 3591 SSRs there were 2122 monorepeats (59.1%), 1027 direpeats (28.6%), 294 trirepeats (8.2%), 63 tetrarepeats (1.75%), 10 pentarepeats (0.25%) and 3 hexarepeats (0.075%). Monorepeats were found to be more copious class compared to di, tri, tetra, penta and hexa repeats (Fig. 1). However, we were able to design primers using Batchprimer3 web based software, only for 1285 microsatellite regions (supplementary material 2). From that set we have randomly selected 75 primers and screened for PCR from which 30 successfully amplified primers with clear one or two bands were used for genotyping. SSRs with 6 tandem repeats (25.5%) were more common, followed by 5 tandem repeats (15.5%), 3 tandem repeats (14.6%), 4 tandem repeats (12.7%), 7 tandem repeats (11.5%), 8 tandem repeats (8.6%), 9 tandem repeats (6.8%), 10 tandem repeats (3%) and 11 tandem repeats (1.7%) (Fig. 4). Among the direpeats, AT/AG repeat motifs (33.3%) were the most common motif. The most frequent trirepeat motifs were TCT/AAT (10.1%) and tetrarepeat motif was AAAT (12.5%) (Fig. 3).

Validation Of Ssrs And Genetic Analysis

A set of 75 primers was chosen for experimental validation includes di, tri and tetra repeats. From the tested primers, 61 (81.33%) generated high quality, reproducible amplicons of expected size and 14 primers failed to amplify. From the amplified 61 primers 30 were used for genotyping of 53 sapota accessions.

In this study, we report the data on 30 polymorphic SSR markers. All these 30 markers show PIC values more than 0.8. The analysis of overall heterozygosity for 30 markers revealed that the number of alleles ranged from 11 to 38 with a total of 692 alleles and a mean of 23.10 alleles per locus. The PIC values ranged from 0.85–0.96 with an average of 0.9118. The mean PIC of loci with tri repeats were high 0.924 compared to that of di and tetra repeats (Tables 1 and 2). Among the 30 loci, 14 markers showed PIC between 0.85–0.89, 13 markers with 0.90–0.95 range and 3 markers with PIC greater than 0.96 (Fig. 2). The PI value ranged from 0.002–0.043 with an average PI of 0.0145 and the total probability of identity was $2.215230e^{-59}$. The locus SapSSR_34 showed highest PIC value 0.962 with 38 alleles and lowest PI value 0.002 (Table 1). The twelve markers, SapSSR_34 (0.002), SapSSR_32 (0.003), SapSSR_4 (0.004), SapSSR_2 (0.005), SapSSR_16 (0.006), SapSSR_21 (0.006), SapSSR_23 (0.006), SapSSR_54 (0.006), SapSSR_1 (0.007), SapSSR_15 (0.008), SapSSR_3 (0.009) and SapSSR_8 (0.009) were found to be having very low PI values. These informative markers will be a useful tool in Sapota breeding programs, diversity analyses and DNA fingerprinting.

Table 1
Genetic analysis of thirty novel sapota microsatellite markers developed.

Locus	Primer Sequences	Repeat Motif	Allele size range (bp)	Number of Alleles (k)	Observed Heterozygosity (H_{Obs})	Expected Heterozygosity (H_{Exp})	Polymorphic Information Content (PIC)	Probility of Identity (PI)
SapSSR_1	F: AACATTTATCAGGTGCCAATA R: GCGAACACAAAAAGACAGTTA	(CT)6	91–153	18	0.094	0.887	0.868	0.007
SapSSR_32	F: TCAGATGGGATTGGGATTCT R: TGAATCAAACTCAAGCGAGGT	(AT)7	174–252	34	0.113	0.962	0.951	0.003
SapSSR_15	F: CTGTTAAAGGGTCAGAGTCAG R: CAATACAAAAGACCAATTTGC	(TC)9	93–149	37	0.642	0.969	0.959	0.008
SapSSR_2	F: ATGTATTGCCTTCTCACTTCC R: TGTCCCACATCATTAAGAAAC	(AC)8	90–155	15	0.170	0.909	0.892	0.005
SapSSR_34	F: GAACAGCCAGATCGAGAAC R: CTGCAGCCGTCCGAACTC	(CAG)5	180–248	38	0.151	0.973	0.962	0.002
SapSSR_16	F: ATCTCCTTTCTTCCATGTAGC R: AAATAGTCTAAGTGGGGTTCG	(CT)7	92–158	38	0.434	0.969	0.959	0.006
SapSSR_3	F: TCCCTTTATGTGAACCTATCA R: GCTGTATCCAAAGAAAAATGA	(TA)7	94–154	13	0.038	0.905	0.887	0.009
SapSSR_4	F: CCAGACTCGCAATCTAATATG R: TATAAACCCTTTTCCTTTCGT	(TG)6	95–150	32	0.509	0.963	0.952	0.004
SapSSR_23	F: AAGAAGATGAAGCTAGGGAAA R: ATGCAGACAGAAAAGAGTGAA	(AT)9	93–147	31	0.491	0.963	0.952	0.006
SapSSR_26	F: CAATTTGACAAACACCCTATC R: TCATTTTACTACTCAAGGTGTCA	(AG)8	100–162	29	0.925	0.952	0.941	0.013
SapSSR_5	F: CAATTTGACAAACACCCTATC R: ATTATCCATTTTGCTTCTCCT	(CA)6	92–156	13	0.038	0.860	0.837	0.017
SapSSR_8	F: CACTAATCTCTGTGTGGGTGT R: GATGCGAGATTCTTTTTGATT	(GA)7	96–151	20	0.038	0.907	0.891	0.009
SapSSR_9	F: TTTCAGTTTCTGAAGAGTCCA R: GAGAGCCCATTACTCTCTAGG	(CT)6	93–144	30	0.245	0.953	0.941	0.010
SapSSR_10	F: TGTACGGATTGGAAGTCG R: CATAGGCCTGGTAGGTCAG	(CG)6	91–126	22	0.642	0.939	0.926	0.014
SapSSR_14	F: AAGAAACTGTTTTCAAACTCG R:AGAAGAAAGAGGTAGCAAAGC	(TG)8	95–141	12	0.000	0.893	0.873	0.033
SapSSR_18	F: TATGAACACACAACACCACAC R: ATCCATGCCTAAGGCTACTAT	(TC)7	97–132	13	0.019	0.910	0.893	0.010
SapSSR_21	F: AACAACGAGGAGAAGAAGAAG R: TCATACGTCGTCGTTTCTATT	(GA)8	92–128	31	0.208	0.953	0.941	0.006
SapSSR_33	F: AGAGCTAAATTTTCCTGCACT R: TTAAACCATCACGTTCAATTC	(TTG)6	131–185	26	0.736	0.959	0.948	0.010

Locus	Primer Sequences	Repeat Motif	Allele size range (bp)	Number of Alleles (k)	Observed Heterozygosity (H_{Obs})	Expected Heterozygosity (H_{Exp})	Polymorphic Information Content (PIC)	Probility of Identity (PI)
SapSSR_27	F: CGTCAATAGAGAGAGACTAAGGA R: CTGTTATTGGTTGCTTGAAGA	(TG)6	91–121	14	0.000	0.898	0.879	0.011
SapSSR_37	F: GTTGAGAGGCAAATTGAAGA R: AATGTTGCTTACGAGAAACTG	(CCT)5	102–165	25	0.283	0.947	0.935	0.012
SapSSR_54	F: AAGAGTATGAGAAGCGGAAGT R: TGATATGGTTCAAACAACTCC	(TGTT)3	122–186	23	0.132	0.935	0.922	0.006
SapSSR_55	F: CCATGCAGTGACCTTTTTA R: AAGAATGAGAATGAGGAGGAG	(CTGA)3	134–191	31	0.736	0.962	0.950	0.011
SapSSR_25	F: AGGAAAGAAGAGTGCCTAAAA R: TAATGCTCTTTTCATGAGGTG	(AG)11	94–133	12	0.000	0.883	0.863	0.035
SapSSR_35	F: GTGGCATATTGACTCTTATGG R: TAACAATGGGACGTTGAATAC	(CAT)5	98–149	16	0.038	0.891	0.872	0.034
SapSSR_36	F: TTTCGATTTTCTCATTACTGG R: GTCGTTTTGAGTTTGTGTGTT	(AAT)5	94–146	19	0.245	0.914	0.898	0.018
SapSSR_39	F: TTAAGAATCCCAAGCAAGAAT R: ATTGCACAATGTCTTTGGTC	(AGC)4	99–155	27	0.717	0.942	0.929	0.016
SapSSR_56	F: ATGGCTATAGCAGTTTGTCAG R: CAAATTTTTGGTCAATCTCAC	(TTAT)4	93–142	14	0.019	0.872	0.850	0.042
SapSSR_57	F: GGTCATGCTCTGGTCATTAT R: AAGCAAGAAAAGGAGCAAATA	(TTTA)3	94–151	11	0.000	0.878	0.857	0.043
SapSSR_58	F: TTGTTACCCACCTTTATTGAA R: CCACTTCTAATTCCTGACAAA	(TAAA)5	92–145	21	0.170	0.900	0.883	0.018
SapSSR_59	F: TTCCATACTGAATCATCACCT R: TGTAAAATTGGCAAACTGACT	(TTAT)5	97–159	28	0.906	0.955	0.944	0.022

Table 2
Range and mean values of PIC of different repeat types

Repeat type	Number of SSRs	Range	Mean PIC
Direpeats	18	0.837–0.959	0.911
Trirepeats	6	0.872–0.962	0.924
Tetrarepeats	6	0.850–0.950	0.901

Using the data generated from the 30 SSR markers, cluster analysis was done following Neighbor – joining tree method, which showed three clusters (Fig. 5). No clear separation was observed between the genotypes. In general, the hybrid varieties were grouped with the maternal parent. This clustering pattern was further assessed using the STRUCTURE V2.3.4 on the basis of assignment tests carried out. The Evanno method illustrated an ideal K = 3 (Fig. 6). This K = 3 authenticates the N-J analysis deduced with three clusters. The STRUCTURE analysis showed 30 genotypes shared ancestry among the population of 53 genotypes (56.6%; Fig. 7). Both the STRUCTURE analyses and N-J analysis showed admixed population.

Analysis of Molecular Variance (AMOVA) showed only 6.33% of the variation was found among the groups, 63.33% of the variation was found among populations within groups and 30.34% of the variation within populations. The F_{st} value estimated from AMOVA was 0.6965 which showed high genetic differentiation (Table 3).

Table 3
AMOVA analysis of genetic variances within and among populations of *Manilkara zapota* (level of significance is based on 10,000 iterations)

Source of variation	d.f.	Sum of squares	Variance components	Percentage of variation
Among groups	2	101.126	0.90968	6.33
Among populations within groups	50	1127.600	9.09675	63.33
Within populations	53	231.00	4.35849	30.34
Total	105	1459.726	14.364	
Fixation index F_{st} = 0.69659				
P value = 0.00000				

Discussion

Sapota is one of the momentous fruit crop cultivated in India. It was an introduced crop by the Spaniards to Asia and in India it was found to be brought through SriLanka. Despite its cultivation throughout the world, very few studies on its genetic diversity using molecular markers.. So far molecular characterization in sapota has been done using dominant RAPD markers (Meghala et al. 2005; Jalawadi et al. 2014; Kumar et al. 2015).

In this study, we have developed microsatellite markers for sapota to characterize at molecular level and to have clear understanding of genetic diversity. SSR markers are widely used in species identification, genome mapping in crop breeding programs, forensics, phylogeography and population genetics due to their abundance availability in the genome, high polymorphism, easy reiteration and cost effectiveness (Ravishankar et al. 2011, 2015c). In the absence of sequenced genomes in non-model species like sapota, the enrichment of genomic libraries with microsatellite markers will be advantageous to develop molecular markers for genetic studies. However, so far there is no report on development of SSR markers in sapota. We have used NGS Illumina HiSeq 2500 platform for developing this informative and versatile DNA based microsatellite markers. At present, NGS has transformed the development of microsatellite markers quick, simple and cost effective with a high throughput data identifying a large number of loci in the genome (Ravishankar et al. 2015b, c; Unamba et al. 2015; Hodel et al. 2016).

The sapota cultivar Cricket Ball genomic DNA was sequenced which generated 3,326,257,143 bases from 22,028,193 reads and after assembly 6396224 contigs were obtained. The GC content of contigs was 40.91% which is within the range commonly observed for plant genomes (Smarda and Bures 2012). A total of 2591 sequences containing 3591 SSRs were identified. We noted mononucleotide repeats were more predominant for 59.1% of all the observed repeats followed by direpeats (28.6%). Others like tri, tetra, penta and hexarepeats accounted for less than 10% (Fig. 1). This finding was in accordance with other crops like mango (Ravishankar et al. 2015c), Garcinia gummi-gutta (Ravishankar et al. 2017), rice, sorghum, *Brachypodium, Arabidopsis, Populous* (Sonah et al. 2011) with predominant monorepeats. Dinucleotide repeats were also common in other crops like *Pouteria sapota* (Arias et al. 2015), Pomegranate (Ravishankar et al. 2015b), sour passion fruit (Araya et al. 2017), Manchurian walnut (Hu et al. 2016), American Cranberry (Zhu et al. 2012).

Among the different motif types, AT and AG dinucleotide repeat motifs, TCT and AAT trinucleotide repeat motifs and AAAT tetranucleotide repeat motif are higher in frequency (Fig. 3). Similar pattern was observed in many crops including *Pouteria sapota* (Arias et al. 2015). The higher frequency a particular repeat motif and its length in the plant genome might be due to selection pressure on that motif over the years during selection and evolution. The evolution of microsatellites in plant genome is not very well studied and also not understood properly. The most common explanation given is it may be due to mutational mechanism through replication slippage. The other likely causes are unequal crossing over, nucleotide substitution, and duplication events. However, they may not explain specific pattern of motif repeats in different species (Buschiazzo and Gemmell 2006; Sonah et al. 2011; Ravishankar et al. 2015c).

In this study, we report successful development of thirty polymorphic microsatellite markers with high PIC values more than 0.8. The mean PIC value was 0.912. The high number of polymorphic SSRs isolation may be due to the Illumina paired-end sequencing which provides an effective alternative to the expensive and time consuming conventional microsatellite enrichment library based method of genome wide SSRs isolation. According to Botstein et al. (1980), any locus with PIC more than 0.5 is highly polymorphic. The mean of observed heterozygosity was 0.291 and expected heterozygosity was 0.927. The number of alleles per locus ranged from 11 to 38 with a mean of 23. The PI values range from 0.0026 to 0.0370 with a mean of 0.0141. A high mean PIC value 0.912 and high mean alleles per locus 23 was observed which might be due to high heterozygosity in the species which also recorded a large number of alleles. We observed 17 (32%) SSR markers showing alleles more than 20 per locus indicating high heterozygosity and diversity in the genotypes used. The markers with low PI can be used as universal primer for sapota DNA fingerprinting.

The clustering pattern in N-J analysis showed three clusters and no clear separation was observed which deviates from the studies carried out in sapota germplasm in India using RAPD markers by Jalawadi (2014), Jalawadi et al. (2014), Kumar et al. (2014) stated that the clustering in sapota was based on size and shape of the fruit. Further analysis of the clustering pattern using the STRUCTURE program figured out ideal value of k = 3

which also showed admixed population with ancestry shared among 56.6% of the population. In *Pouteria sapota*, a study carried out by Arias et al. (2015) microsatellite markers showed a clustering pattern based on the geographical locations and the STRUCTURE analyses showed admixed population((Fig. 7). The hindrance in the clustering pattern in N-J analysis is probably due to the admixed population shown by STRUCTURE analyses.

Analysis of molecular variance revealed a significant F_{st} value of 0.69659 indicating high genetic differentiation among the 53 genotypes and 3 populations studied. There was a high differentiation among populations within groups and low differentiation among groups. The observed diversity among populations within groups indicates likely coexistence of different genotypes in the same region (Ravishankar et al. 2015a). In this study the genetic differentiation within the populations showed by the STRUCTURE analysis is similar with the AMOVA results.

Sapota is an introduced crop into India. In such crops the genetic differentiation is generally depends on the number cultivars introduced, degree of heterozygosity or the origin of the cultivars, which are unclear here. Hence, it is expected to be narrow and the genetic variation to be less. However, the results of this study shows that high genetic differentiation and diversity in the sapota population. This is in accordance with earlier studies by Jalawadi et al. (2014) and Kumar et al. (2014). Initially sapota might have been cultivated using seedlings, due to their high heterozygosity there are variations in the off-springs. Later it was selected and vegetatively propagated based on the preference of the region for fruit characteristics and yield. Therefore, there is a possibility of occurrence of wide diversity and a great extent of genetic variability in sapota might be originated due to seedling segregation and it was also possible a large number of seedlings or grafts of sapota was introduced to India from the place of origin. This is the first study in sapota where SSR or microsatellites were developed and genetic diversity of Indian collections was examined. The SSR markers developed would be helpful in developing linkage map, assessing genetic diversity and also molecular characterization of genotypes.

Declarations

Acknowledgments Authors acknowledge funding support from ICAR-AICRP fruits and ICAR-IIHR, Bengaluru, India.

Author contribution statement Project was conceptualized by KVR, PP. Sample Preparation, experiments and data analysis was carried out by NSP, KVR, NSP,IP, ARP,RB,AMS, AR and PP. KVR, PP, AR and NSP were involved in discussion of data and MS preparation

Ethics declaration

The authors declare that there is no conflict of interest

Availability of data and material

NGS data has been submitted to NCBI with accession ID SRP127995

References

1. Araya S, Martins AM, Junqueira NTV, et al (2017) Microsatellite marker development by partial sequencing of the sour passion fruit genome (*Passiflora edulis* Sims). BMC Genomics 18:549. doi: 10.1186/s12864-017-3881-5
2. Arias RS, Martínez-Castillo J, Sobolev VS, et al (2015) Development of a large set of microsatellite markers in zapote mamey (*Pouteria sapota* (Jacq.) H.E. Moore & Stearn) and their potential use in the study of the species. Molecules 20:11400–11417. doi: 10.3390/molecules200611400
3. Bhat ZA, Dhillon WS, Rashid R (2010) The role of Molecular Markers in Improvement of Fruit Crops. Not Sci Biol 2:22–30. doi: 10.15835/nsb.2.2.4222
4. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. doi: 10.1093/bioinformatics/btu170
5. Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–31
6. Buschiazzo E, Gemmell N (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. Bioessays 28:1040–1050. doi: 10.1002/bies.20470
7. Earl, D.A. and VonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour 4:359–361
8. Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5:435
9. Excoffier L, Laval G, Schneider S (2005) Arlequin: an integrated software package for population genetics data analysis. Evol Bioinforma 1:47–50. doi: 10.1111/j.1755-0998.2010.02847.x
10. Hodel RGJ, Gitzendanner MA, Germain-Aubrey CC, et al (2016) A New Resource for the Development of SSR Markers: Millions of Loci from a Thousand Plant Transcriptomes. Appl Plant Sci 4:1600024. doi: 10.3732/apps.1600024

11. Hu Z, Zhang T, Gao XX, et al (2016) De novo assembly and characterization of the leaf, bud, and fruit transcriptome from the vulnerable tree Juglans mandshurica for the development of 20 new microsatellite markers using Illumina sequencing. Mol Genet Genomics 291:849–862. doi: 10.1007/s00438-015-1147-y
12. Jalawadi S (2014) Molecular characterization of hybrids and landraces of sapota by. Asian J Hortic 9:72–75
13. Jalawadi S, Kanamadi VC, Basavarajappa HR, et al (2014) Molecular Characterization of Sapota Genotypes Using RAPD Markers. In: Acta Horticulturae. pp 61–68
14. Kaeuffer R, Reale D, Coltman DW, Pontier D (2007) Detecting population structure using STRUCTURE software: Effect of background linkage disequilibrium. Heredity (Edinb) 99:374–380. doi: 10.1038/sj.hdy.6801010
15. Kalinowski, Steven T, Taper ML, Marshall TC (2007) Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. Mol Ecol 16:1099–1106. doi: 10.1111/j.1365-294X.2007.03089.x
16. Kumar M, Saraswathy S, Ramesh Kumar S, et al (2015) Genetic Diversity Analysis in Sapota Cultivars as reavealed by RAPD markers. Environ Ecol 33:898–900
17. Meghala R, Ravishankar, K V, Anand L, Rekha A (2005) Genetic diversity of Indian sapota (*Manilkara zapota*) cultivars characterized by RAPD markers. Plant Genet Resour Newsl 43–46
18. Perrier X, Jacquemoud-Collet JP (2006) DARwin software. http:// darwin.cirad.fr/
19. Ramasamy RK, Ramasamy S, Bindroo BB, Naik VG (2014) STRUCTURE PLOT: A program for drawing elegant STRUCTURE bar plots in user friendly interface. Springerplus 3:1–3. doi: 10.1186/2193-1801-3-431
20. Rastegar S (2015) Physical , Biochemical and Mineral Evaluation of Sapota Fruits During Growth , Development and Ripening. Agric Commun 3:14–19
21. Ravishankar KV, Vasudeva R, Hemanth B, et al (2017) Isolation and characterization of microsatellite markers in *Garcinia gummi-gutta* by next-generation sequencing and cross-species amplification. J Genet 96:213–218. doi: 10.1007/s12041-017-0756-0
22. Ravishankar K V., Bommisetty P, Bajpai A, et al (2015a) Genetic diversity and population structure analysis of mango (*Mangifera indica*) cultivars assessed by microsatellite markers. Trees 29:775–783. doi: 10.1007/s00468-015-1155-x
23. Ravishankar K V., Chaturvedi K, Puttaraju N, et al (2015b) Mining and characterization of SSRs from pomegranate (*Punica granatum* L.) by pyrosequencing. Plant Breed 134:247–254. doi: 10.1111/pbr.12238
24. Ravishankar K V., Dinesh MR, Nischita P, Sandya BS (2015c) Development and characterization of microsatellite markers in mango (*Mangifera indica*) using next-generation sequencing technology and their transferability across species. Mol Breed 35:93. doi: 10.1007/s11032-015-0289-2
25. Ravishankar K V., Mani BH-R, Anand L, Dinesh MR (2011) Development of new microsatellite markers from Mango (*Mangifera indica*) and cross-species amplification. Am J Bot 98:e96–e99. doi: 10.3732/ajb.1000263
26. Ravishankar K V, Anand L, Dinesh MR (2000) Assessment of genetic relatedness among mango cultivars of India using RAPD markers. J Hortic Sci Biotechnol 75:198–201. doi: 10.1080/14620316.2000.11511223
27. Rekha A, Dinesh MR, Venugopalan R, Murthy BNS (2011) Genetic correlation and cluster analysis in sapota (*Manilkara zapota*). J Hortic Sci 6:101–104
28. Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. Nat Biotechnol 18:233–234. doi: 10.1038/72708
29. Siddiqui MW, Longkumer M, Ahmad MS, et al (2014) Postharvest biology and technology of sapota: a concise review. Acta Physiol Plant 36:3115–3122. doi: 10.1007/s11738-014-1696-4
30. Smarda P, Bures P (2012) The Variation of Base Composition in Plant Genomes. In: Plant Genome Diversity Volume 1. pp 209–235
31. Sonah H, Deshmukh RK, Sharma A, et al (2011) Genome-wide distribution and organization of Microsatellites in plants: An insight into marker development in Brachypodium. PLoS One 6:e21298. doi: 10.1371/journal.pone.0021298
32. Unamba CIN, Nag A, Sharma RK (2015) Next Generation Sequencing Technologies: The Doorway to the Unexplored Genomics of Non-Model Plants. Front Plant Sci 6:1074. doi: 10.3389/fpls.2015.01074
33. Wagner HW, Sefc KM (1999) IDENTITY 1.0. Centre for Applied Genetics. University of Agricultural Sciences, Vienna, Austria
34. You FM, Huo N, Gu YQ, et al (2008) BatchPrimer3: A high throughput web application for PCR and sequencing primer design. BMC Bioinformatics 9:1–13. doi: 10.1186/1471-2105-9-253
35. Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829. doi: 10.1101/gr.074492.107
36. Zhu H, Senalik D, McCown BH, et al (2012) Mining and validation of pyrosequenced simple sequence repeats (SSRs) from American cranberry (*Vaccinium macrocarpon* Ait.). Theor Appl Genet 124:87–96. doi: 10.1007/s00122-011-1689-2

Figures

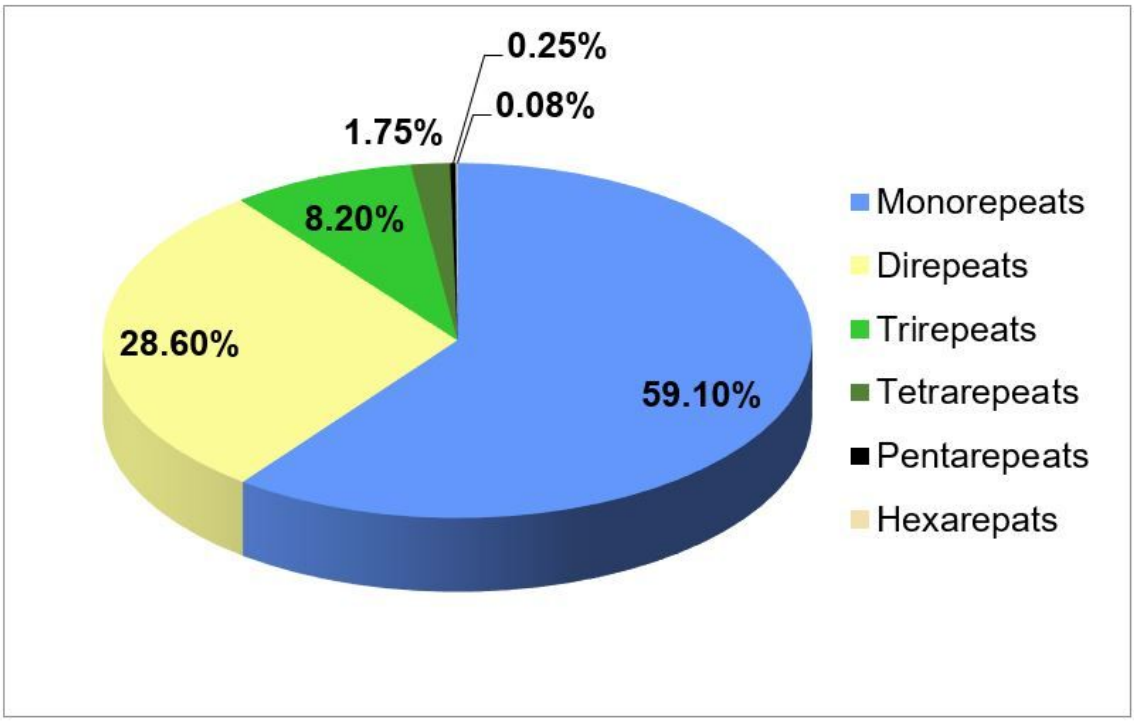

Figure 1

Distribution of repeat motifs in Sapota

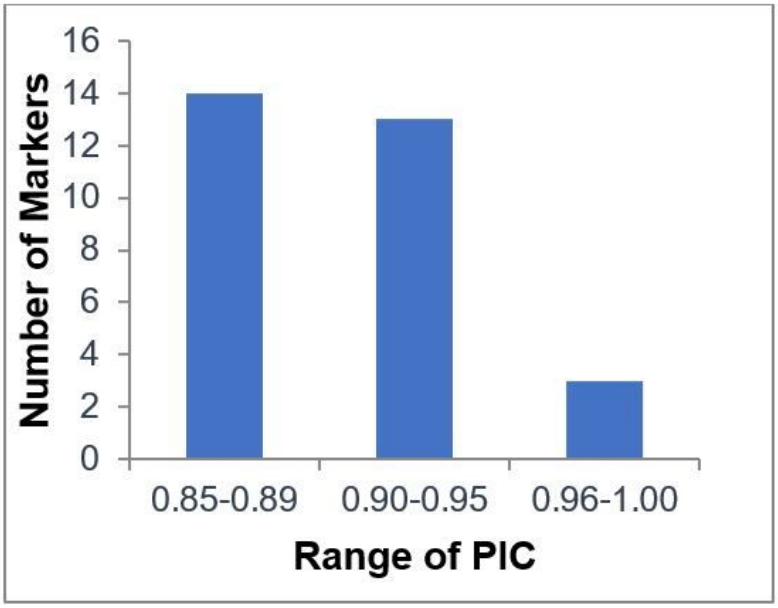

Figure 2

Frequency distribution of 30 SSR markers in Sapota with respect to range of PIC

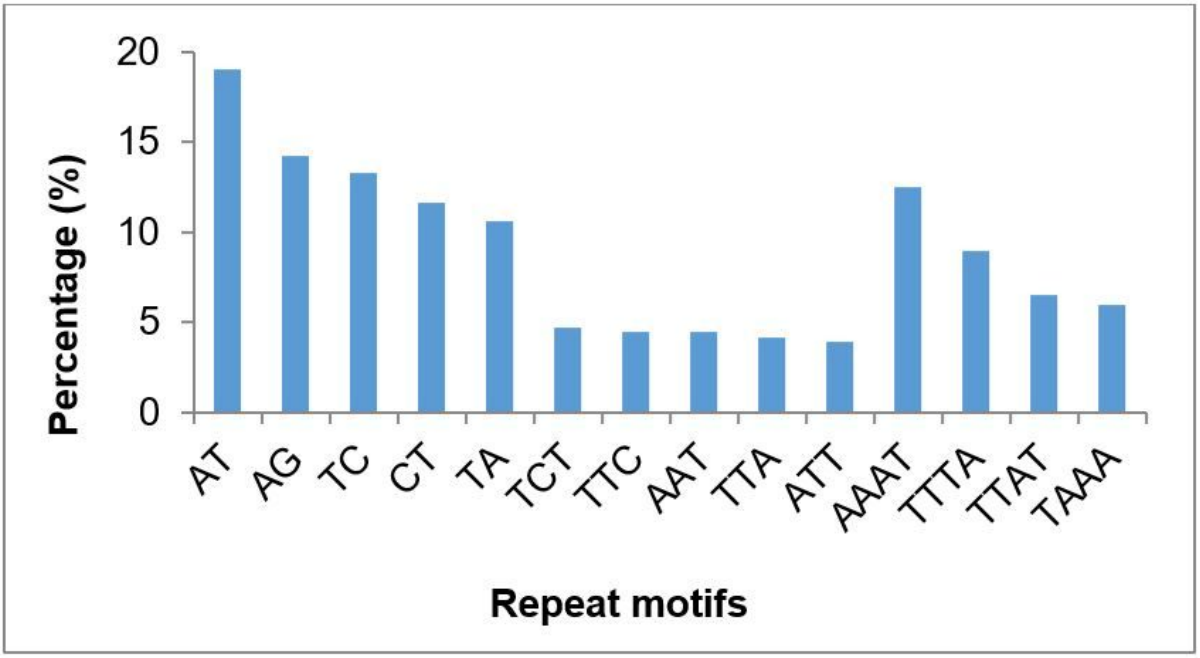

Figure 3

Frequent AT rich repeat motifs distributed in each class

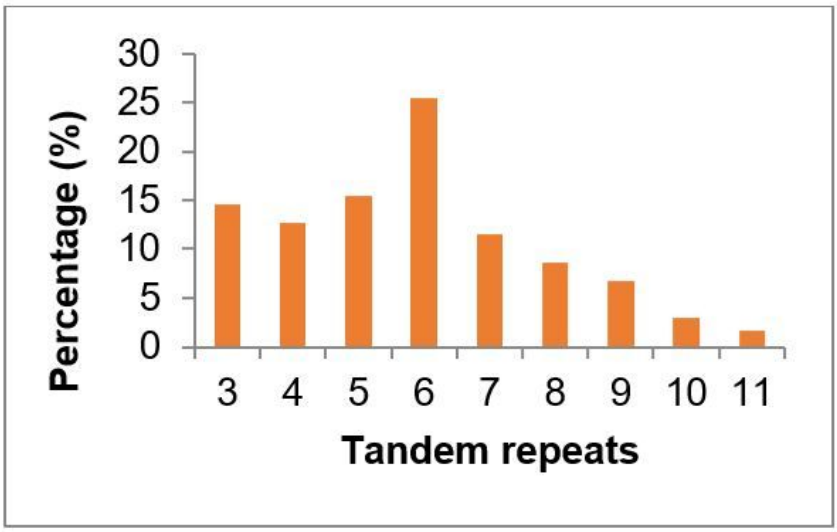

Figure 4

Distribution of tandem repeat types

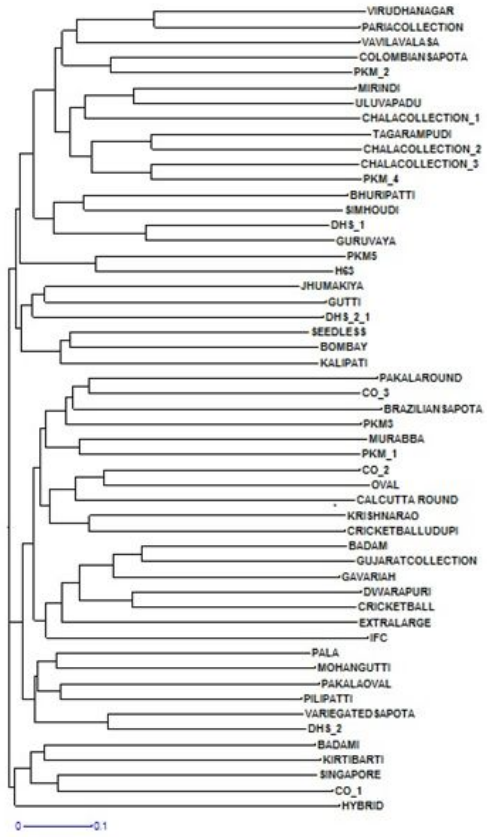

Figure 5

Neighbor-joining analysis of 53 genotypes of Sapota using DARwin 6.0 software.

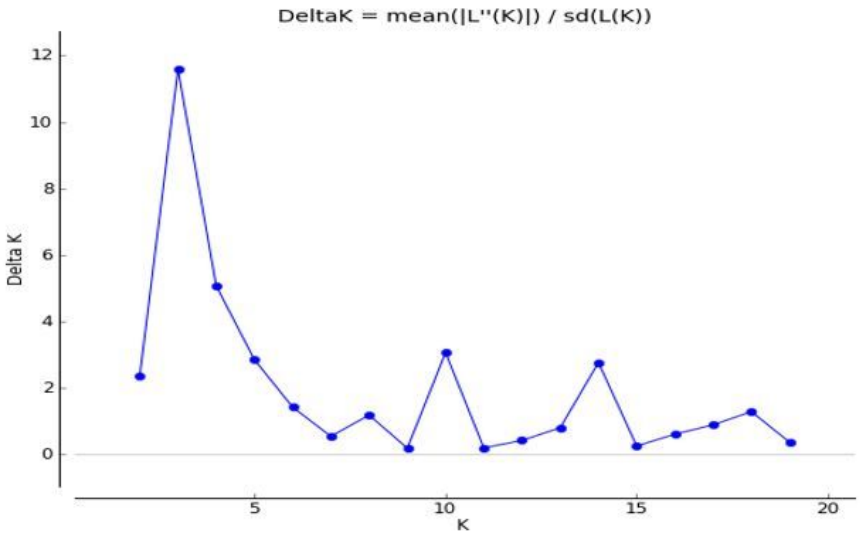

Figure 6

Graph of delta k values to determine ideal number of groups present in sapota using 30 SSR loci and the Evanno method implemented in STRUCTURE HARVESTER program

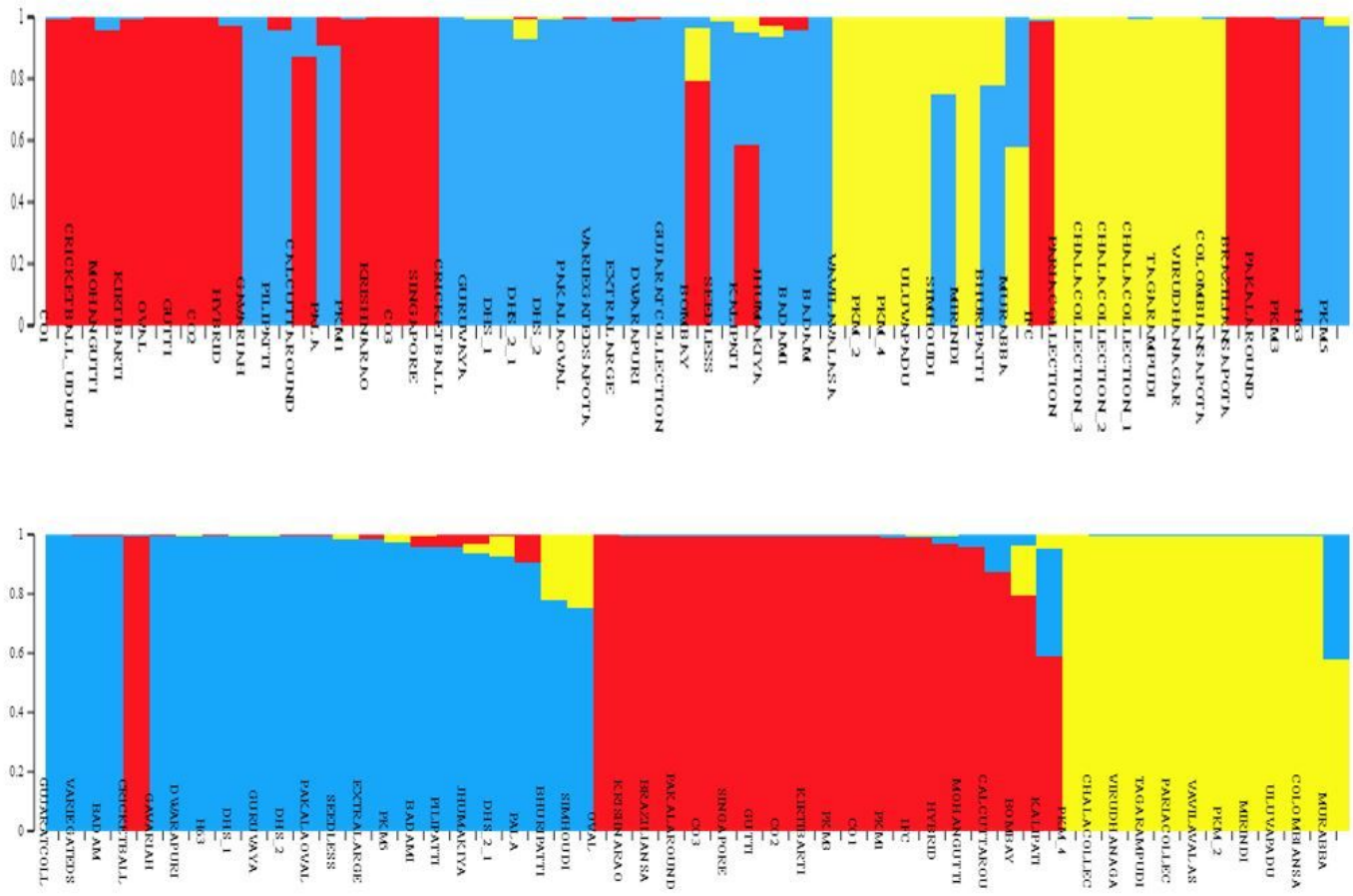

Figure 7

Proportion of estimated ancestry coefficients of sapota populations using 30 SSR loci and STRUCTURE program with k=3, depicted using STRUCTURE PLOT

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryfile1.docx
- supplementaryfile2.xlsx