Development and characterization of microsatellite markers, genetic diversity and population structure analysis in Sapota (*Manilkara zapota* (L.) P. Royen)

Pavithra N. Sathanandam · Prakash Patil · Ajitha Rekha · Iyamperumal Muthuvel · Amrutlal R. Patel · Ramesh B. Boggala · Adiveppa M. Shirol · Kundapura V. Ravishankar

Received: 18 June 2021 / Accepted: 20 April 2022 / Published online: 3 June 2022
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

Abstract *Manilkara zapota* (L.) P. Royen, a widely adapted 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, the study sequenced partial genomic DNA using next generation sequencing technology on Illumina HiSeq 2500 platform. 3.3 Gb data assembled into 6,396,224 contigs were analysed. From these contigs, 3591 simple sequence repeats were identified. Among these different types 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 randomly selected 30 primers were standardized and employed for amplification in 53 genotypes of Sapota. Six hundred and niny-two alleles were observed from 30 loci with a polymorphic information content that 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. The genetic diversity assessed by neighbour-joining and STRUCTURE assignment tests showed an admixed population with three groups. Analysis of molecular variance revealed a significant $F_{st}$ value of 0.69659 indicating high genetic differentiation among the 53 genotypes. Thus, the developed microsatellites will be advantageous in
assessing genetic diversity, developing linkage map, and also molecular characterization of genotypes.

**Keywords**  
Manilkara zapota · Illumina sequencing · Microsatellites · Polymorphic information content · Genetic diversity · Population structure

**Introduction**

Sapota (*Manilkara zapota* (L.) P. Royen) is a delectable fruit of tropical and sub-tropical regions, and belongs to the family *Sapotaceae*. It has spread to many tropical countries though it is native to Central America. In Asia, it was first introduced in Philippines by the Spanish, and later spread to other Asian countries (Meghala et al. 2005). In India, it was introduced during 1800’s from Mexico via Sri Lanka (Rekha et al. 2011), and is commonly known as “chikku”. The hard nature and wide range of adaptability to different conditions have made this tree popular in India. It is mainly meant for its fruit value. It is commercially grown in some countries 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-a-days, 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 programme.

Microsatellites, also known as Simple Sequence Repeats (SSRs), are segments of genomic sequences that include tandem repeats of small nucleotide motifs (1–6 bp) (Ellegren 2004). Because of their co-dominant character, multi-allelic nature, excellent repeatability, and high polymorphism information content, SSRs are widely employed in plant genetic studies. Initially, microsatellites were detected and isolated by microsatellite enrichment library method through selective hybridization. Another method is to discover microsatellite repeats in DNA databases such as EST sequences or genomic sequences. At present, Next Generation Sequencing (NGS), a robust revolutionizing technology, is used for microsatellites development (Ravishankar et al. 2015a, b; Araya et al. 2017). Using microsatellite enrichment library method, a small number of microsatellites have been isolated from Sapota and tested on a few genotypes (Arias et al 2020).

Hence, next generation sequencing technology was employed in the study to partially sequence Sapota genome which is promising approach to generate high throughput genome data at reduced cost and with less time. This would further help in the detection of thousands of microsatellite sites in the genome of target species. The NGS generated data were used to identify and standardize microsatellite markers in Sapota. Furthermore, the data were analysed for its genetic diversity and population structure of Sapota germplasm collected from different parts of India, which is available at ICAR—Indian Institute of Horticultural Research (IIHR), Bengaluru, India. The results of the analysis helped to understand the nature of diversity and genetic variation present in this introduced Sapota crop.

**Materials and methods**

**Plant material and DNA extraction**

Fifty-three genotypes of Sapota (supplementary material 1) were utilized in the study. Young leaf samples were collected from the Sapota germplasm collection which is maintained at ICAR—IIHR (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**

The total genomic DNA of ‘Cricket ball’ variety was used to perform paired end HiSeq Illumina
sequencing using 2500 platform following manufacturer’s protocol (QTLomics, Technologies Pvt 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 programme (Bolger et al. 2014). High-quality data were used for analysis i.e. bases having Phred score ≥ 40. Filtered reads were used to get optimal k-mer for assembly using Velvet Optimizer (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 1 kb were taken for SSR mining.

Microsatellites mining and primer designing

MISA software, a Perl-based SSRs motifs scrutinizing tool, 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 sequence, 75 primer pairs were chosen and tailed with M13 sequence ‘GTAAACGACGCTTCAGT’ 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 min followed by 35 cycles at 94° C for 45 s, an annealing temperature of 55 °C for 45 s, and 72 °C for one minute. The final extension reaction was proceeded at 72 °C for 7 min. For confirmation, the amplified products were separated on 2% agarose gel. The PCR samples were then 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/s Bioserve facility, Hyderabad.

Data analysis

The raw data generated through genotyping were analysed using GeneMarker V2.6.3 software (Softgenetics, LLC) for detecting the fragment size of the alleles. The generated fragment size data were used for genetic analysis using CERVUS 3.0 software (Kalinowski et al. 2007) for calculating the number of alleles (k), observed heterozygosity (Hobs), expected heterozygosity (Hexp) and polymorphic information content (PIC). The Probability of Identity (PI) was also calculated using IDENTITY 1.0 software (Wagner and Sefc 1999). The genetic relationship among 53 Sapota genotypes was analysed and examined by dendrogram analysis using neighbour-joining method with DARwin 6 software (Perrier and Jacquemoud-Collet 2006). For further confirmation, the genetic structure analysis was inferred with STRUCTURE V2.3.4 software (Kaeuffer et al. 2007) with a data set of 30 SSR markers. 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 were calculated using Evanno method, and graphical representations were deduced using the STRUCTURE HARVESTER programme (Earl and VonHoldt 2012; http://taylor0biology.ucla.edu/structureHarvester/) and STRUCTURE PLOT (Ramasamy et al. 2014; http://omicspeaks.com/strplot2/), respectively. AMOVA analysis was performed based on the population clusters obtained by STRUCTURE analysis using Arlequin V 3.5 software with 10,000 permutations (Excoffier et al. 2005). AMOVA derived genetic differentiation values (FST) 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 21,654,953 (98.31%) paired end reads were obtained finally. Assembly optimization was done with k-mer 31 as it had optimal readings for N50 which generated 6,396,224 contigs in the assembly (Supplementary file 2). The total length of assembled contigs was 839,591,847 bp. The longest contig had a length of 11,136 bp, and minimum length of the contig was 61 bp. The average GC content of the genome was 40.91%.

Microsatellites mining

Out of 16,416 sequences examined, 2591 sequences contained SSRs. Among 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 (0.08%). Monorepeats were found to be more predominant class compared to di, tri, tetra, penta and hexa repeats (Fig. 1). However, design primers using Batchprimer3 web-based software was done only for 1285 microsatellite regions (supplementary material 2). From that set, randomly selected 75 primers were 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%). 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%).

Validation of SSRs and genetic analysis

A set of 75 primers was chosen for experimental validation which included di, tri and tetra repeats. Out of these, 14 primers failed to amplify Sapota DNA. From the amplified set of primers, 30 were used for genotyping of 53 Sapota accessions. In the present study, the data on 30 polymorphic SSR markers were reported. 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 to 0.96 with an average of 0.9118. The mean PIC of loci with trirepeats were high 0.924 compared to that of di and tetra repeats (Table 1). Among the 30 loci, 14 markers showed PIC between 0.85 and 0.89, 13 markers with 0.90–0.95 range and 3 markers with PIC greater than 0.96. The PI (Probability of Identity) values ranged from 0.002 to 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.

Using the data generated from the 30 SSR markers, cluster analysis was done following Neighbour-joining tree method, which showed three clusters (Fig. 2). In general, the hybrid varieties were grouped

![Fig. 1 Distribution of repeat motifs in Sapota](image)
| Locus    | Primer sequences | Repeat Motif | Allele size range (bp) | Number of Alleles (k) | Observed Heterozygosity ($H_{\text{Obs}}$) | Expected Heterozygosity ($H_{\text{Exp}}$) | Polymorphic Information Content (PIC) | Probability of Identity (PI) |
|----------|------------------|--------------|------------------------|-----------------------|------------------------------------------|------------------------------------------|-------------------------------------------|---------------------------|
| SapSSR_1 | F: AACATT TATCAG GTGCCA ATA  
           | R: GCGAAC ACAAAA AGACAG TTA  | (CT)6                  | 91–153                | 18                                      | 0.094                                    | 0.887                                    | 0.868                     | 0.007                     |
| SapSSR_32 | F: TCAGAT GGGATT GGGATT CT  
             | R: TGAATC AAACTC AAGCGAG GT  | (AT)7                  | 174–252               | 34                                      | 0.113                                    | 0.962                                    | 0.951                     | 0.003                     |
| SapSSR_15 | F: CTGTTA AAGGGT CAGAGT CAG  
            | R: CAATAC AAAAGA CCAATT TGC  | (TC)9                  | 93–149                | 37                                      | 0.642                                    | 0.969                                    | 0.959                     | 0.008                     |
| SapSSR_2  | F: ATGTAT TGCCCT CTCACT TCC  
             | R: TGTCCT ACATCA TTAAGA AAC  | (AC)8                  | 90–155                | 15                                      | 0.170                                    | 0.909                                    | 0.892                     | 0.005                     |
| SapSSR_34 | F: GAACAG CCAGAT CGAGAAC  
              | R: CTGCAG CCGTCC GAACCT  | (CAG)5                  | 180–248               | 38                                      | 0.151                                    | 0.973                                    | 0.962                     | 0.002                     |
| SapSSR_16 | F: ATCTCC TTCTTT CCATGT AGC  
            | R: AAATAG TCTAAG TGGGGTT TCG  | (CT)7                  | 92–158                | 38                                      | 0.434                                    | 0.969                                    | 0.959                     | 0.006                     |
Table 1 (continued)

| 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) | Probability of Identity (PI) |
|---------|---------------------------|--------------|------------------------|-----------------------|-------------------------------------|-------------------------------------|--------------------------------------|-------------------------------|
| SapSSR_3 | F: TCCCTT TATGTG AACCTA TCA  
R: GCTGTA TCCAAA GAAAAA TGA | (TA)7         | 94–154                 | 13                    | 0.038                               | 0.905                               | 0.887                                | 0.009                          |
| SapSSR_4 | F: CCAGAC TCGCAA TCTAAT ATG  
R: TATAAA CCCCCT TCCCCCT CGT | (TG)6         | 95–150                 | 32                    | 0.509                               | 0.963                               | 0.952                                | 0.004                          |
| SapSSR_23 | F: AAGAAG ATGAAG CTAGGG AAA  
R: ATGCAG ACAGAA AAGAGT GAA | (AT)9         | 93–147                 | 31                    | 0.491                               | 0.963                               | 0.952                                | 0.006                          |
| SapSSR_26 | F: CAAATT GACAAA CACCCCT ATC  
R: TCACTT TACTAC TCAAGG TGGTCA | (AG)8         | 100–162                | 29                    | 0.925                               | 0.952                               | 0.941                                | 0.013                          |
| SapSSR_5  | F: CAAATT GACAAA CACCCCT ATC  
R: ATTATC CATTTT GCTTCT CCT | (CA)6         | 92–156                 | 13                    | 0.038                               | 0.860                               | 0.837                                | 0.017                          |
| SapSSR_8  | F: CACTAA TCTCTG TGTTGG TGT  
R: GATGCG AGATTC TTTTTT ATT | (GA)7         | 96–151                 | 20                    | 0.038                               | 0.907                               | 0.891                                | 0.009                          |
| 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) | Probability of Identity (PI) |
|-----------|------------------|--------------|------------------------|-----------------------|--------------------------------------|--------------------------------------|---------------------------------------|----------------------------|
| SapSSR_9  | F: TTTTCAG       | (CT)6        | 93–144                 | 30                    | 0.245                                | 0.953                                | 0.941                                 | 0.010                      |
|           | TTTCTG           |              |                        |                       |                                      |                                      |                                       |                            |
|           | AAGAGT           |              |                        |                       |                                      |                                      |                                       |                            |
|           | CCA              |              |                        |                       |                                      |                                      |                                       |                            |
|           | R: GAGAGC        |              |                        |                       |                                      |                                      |                                       |                            |
|           | CCATTA           |              |                        |                       |                                      |                                      |                                       |                            |
|           | CTCTCT           |              |                        |                       |                                      |                                      |                                       |                            |
|           | AGG              |              |                        |                       |                                      |                                      |                                       |                            |
| SapSSR_10 | F: TGTTACG       | (CG)6        | 91–126                 | 22                    | 0.642                                | 0.939                                | 0.926                                 | 0.014                      |
|           | GATTGG           |              |                        |                       |                                      |                                      |                                       |                            |
|           | AAGCTG           |              |                        |                       |                                      |                                      |                                       |                            |
|           | R: CATAGG        |              |                        |                       |                                      |                                      |                                       |                            |
|           | CCTGGT           |              |                        |                       |                                      |                                      |                                       |                            |
|           | AGGTCAG          |              |                        |                       |                                      |                                      |                                       |                            |
| SapSSR_14 | F: AAGAAAA       | (TG)8        | 95–141                 | 12                    | 0.000                                | 0.893                                | 0.873                                 | 0.033                      |
|           | CTGTIT          |              |                        |                       |                                      |                                      |                                       |                            |
|           | TCAAAAC          |              |                        |                       |                                      |                                      |                                       |                            |
|           | TCG              |              |                        |                       |                                      |                                      |                                       |                            |
|           | R: AGAAGA        |              |                        |                       |                                      |                                      |                                       |                            |
|           | AAGAGG           |              |                        |                       |                                      |                                      |                                       |                            |
|           | TAGCAA           |              |                        |                       |                                      |                                      |                                       |                            |
|           | AGC              |              |                        |                       |                                      |                                      |                                       |                            |
| SapSSR_18 | F: TATGAA        | (TC)7        | 97–132                 | 13                    | 0.019                                | 0.910                                | 0.893                                 | 0.010                      |
|           | CACACA           |              |                        |                       |                                      |                                      |                                       |                            |
|           | ACACCA           |              |                        |                       |                                      |                                      |                                       |                            |
|           | CAC              |              |                        |                       |                                      |                                      |                                       |                            |
|           | R: ATCCAT        |              |                        |                       |                                      |                                      |                                       |                            |
|           | GCCTAA           |              |                        |                       |                                      |                                      |                                       |                            |
|           | GGCTAC           |              |                        |                       |                                      |                                      |                                       |                            |
|           | TAT              |              |                        |                       |                                      |                                      |                                       |                            |
| SapSSR_21 | F: AACAAC        | (GA)8        | 92–128                 | 31                    | 0.208                                | 0.953                                | 0.941                                 | 0.006                      |
|           | GAGGAG           |              |                        |                       |                                      |                                      |                                       |                            |
|           | AAGAAG           |              |                        |                       |                                      |                                      |                                       |                            |
|           | AAG              |              |                        |                       |                                      |                                      |                                       |                            |
|           | R: TCATAC        |              |                        |                       |                                      |                                      |                                       |                            |
|           | GTGTCG           |              |                        |                       |                                      |                                      |                                       |                            |
|           | GTTTCCTT         |              |                        |                       |                                      |                                      |                                       |                            |
|           | ATT              |              |                        |                       |                                      |                                      |                                       |                            |
| SapSSR_33 | F: AGAGCT        | (TTG)6       | 131–185                | 26                    | 0.736                                | 0.959                                | 0.948                                 | 0.010                      |
|           | AAATTT           |              |                        |                       |                                      |                                      |                                       |                            |
|           | TCTGTC           |              |                        |                       |                                      |                                      |                                       |                            |
|           | ACT              |              |                        |                       |                                      |                                      |                                       |                            |
|           | R: TTAACAC       |              |                        |                       |                                      |                                      |                                       |                            |
|           | CATCAC           |              |                        |                       |                                      |                                      |                                       |                            |
|           | GTTCAA           |              |                        |                       |                                      |                                      |                                       |                            |
|           | TTC              |              |                        |                       |                                      |                                      |                                       |                            |
| Locus     | Primer sequences | Repeat Motif | Allele size range (bp) | Number of Alleles (k) | Observed Heterozygosity ($H_{\text{Obs}}$) | Expected Heterozygosity ($H_{\text{Exp}}$) | Polymorphic Information Content (PIC) | Probability of Identity (PI) |
|-----------|------------------|--------------|------------------------|-----------------------|-------------------------------------------|-------------------------------------------|-------------------------------------|-----------------------------|
| SapSSR_27 | F: CGTCAA TTAGAGA GAGACT AAGGA R: CTGTTA TTGGTT GCTTGA AGA | (TG)6 | 91–121 | 14 | 0.000 | 0.898 | 0.879 | 0.011 |
| SapSSR_37 | F: GTTGAG AGGCAA ATTGAA GA R: AATGTT GCTTAC GAGAAA CTG | (CCT)5 | 102–165 | 25 | 0.283 | 0.947 | 0.935 | 0.012 |
| SapSSR_54 | F: AAGAGT ATGAGA AGCGGA AGT R: TGATAT GGTTCA AACAC TCC | (TGGT)3 | 122–186 | 23 | 0.132 | 0.935 | 0.922 | 0.006 |
| SapSSR_55 | F: CCATGC AGTGAC CTTTTTA R: AAGAAT GAGAAAT GAGGAG GAG | (CTGA)3 | 134–191 | 31 | 0.736 | 0.962 | 0.950 | 0.011 |
| SapSSR_25 | F: AGGAAA GAAGAG TGCCCTA AAA R: TAATGC TCTTTT CATGAG GTG | (AG)11 | 94–133 | 12 | 0.000 | 0.883 | 0.863 | 0.035 |
| SapSSR_35 | F: GTGGCA TATTGA CTCTTA TGG R: TAACAA TGGGAC GTTTGA TAC | (CAT)5 | 98–149 | 16 | 0.038 | 0.891 | 0.872 | 0.034 |
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. 3a). 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. 3). Both the STRU
Fig. 2 Neighbour-joining analysis of 53 genotypes of Sapota using DARwin 6.0 software.
CTURE 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 2).

**Discussion**

Sapota was an introduced crop by the Spaniards to Asia, and it was found to be brought to India through
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 DNA markers like RAPD (Meghala et al. 2005; Jalawadi et al. 2014; Kumar et al. 2015).

In present study, microsatellite markers for Sapota were developed 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 programmes, 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 only one report on development of SSR markers in Sapota using microsatellite library enrichment method (Arias et al. 2020), which identified a 47 SSR markers and number of alleles reported in this study was low, as they used only 4 genotypes. The distribution of different types of repeat motifs cannot assessed using microsatellite enrichment library method. NGS Illumina HiSeq 2500 platform was utilized for developing this informative and versatile DNA based microsatellite markers. At present, NGS has transformed the development of microsatellite markers in to a quick, simple and cost effective method 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 6,396,224 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. It is found that mononucleotide repeats were more predominant, accounting for 59.1% of all the observed repeats followed by dinucleotides (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 different motif types, AT and AG dinucleotide repeat motifs, TCT and AAT trinucleotide repeat motifs and AAAT tetranucleotide repeat motif are higher in frequency. 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 neither well studied nor understood properly. The reason could be due to mutational mechanism through replication slippage. The other causes might be 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).

Present study reports 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 Illumina paired-end sequencing which provides an effective alternative

### Table 2 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
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 average number of alleles per locus was 23, with a range of 11 to 38. With a mean value of 0.0141, the PI values vary from 0.0026 to 0.0370. The higher PIC and relatively high alleles per locus indicate the presence of high heterozygosity in this species. It is 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 primers for Sapota DNA fingerprinting.

The clustering pattern in N-J analysis showed three clusters (Fig. 2), cluster 1 has most of the genotypes from Gujarat state, Eastern region of India, Cluster 3 includes genotypes mainly from southern states Tamil Nadu, Telangana, and Andhra Pradesh., Cluster 2 has admixture of genotypes from different regions. Generally, there is an affiliation towards geographical region where they originated. However, there is no clear separation was observed. Our results deviate 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 programme 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. 3). 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 Fst 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. 2015c). In this study the genetic differentiation within the populations showed by the STRUCTURE analysis is similar with the AMOVA results.

Sapota is a crop that has been introduced to India. The variability of such introduced crops is determined by the number of cultivars/genotypes introduced and the degree of heterozygosity, both of which are unknown in this case. As a result, it is projected to be narrow, with little genetic variation. However, the results show 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 nature there are high variations in the off-springs. However, low observed heterozygosity in the open pollinated crop, might be due to small number of introduction, and Sapota is grown in small packets predominately as a single variety, which causes a type of inbreeding depression due to lack of diverse pollen sources. It was later selected and vegetatively propagated based on the preferences of the region for fruit quality and productivity. As a result of seedling segregation, considerable diversity and a high degree of genetic variability in Sapota could have emerged. It is also possible that a significant number of seedlings or grafts of Sapota were imported to India from the place of its origin. This is the first study to employ microsatellites to evaluate the genetic diversity of Indian Sapota collections. The SSR markers developed would aid in the development of a linkage map, the assessment of genetic diversity, and molecular characterization.

Acknowledgements Authors acknowledge funding support from ICAR-AICRP on Fruits and ICAR-IIHR, Bengaluru, India.

Author’s contribution Project was conceptualized by KVR, PP. Sample Preparation, experiments and data analysis were 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.

Funding This study was funded by ICAR-AICRP on Fruits and ICAR-IIHR, Bengaluru.

Availability of data and material NGS data have been submitted to NCBI with accession ID SRP127995.

Declarations
Conflict of interest  The authors declare that there is no conflict of interest.

References

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. https://doi.org/10.1186/s12864-017-3881-5

Arias RS, Martinez-Castillo J, Sobolev VS et al (2015) Development of a large set of microsatellite markers in zapote maneye (*Pouteria sapota* (Jacq.) H.E. Moore & Stearn) and their potential use in the study of the species. Molecules 20:11400–11417. https://doi.org/10.3390/molecules200611400

Arias RS, Ballard LL, Duke MV et al (2020) Development of nuclear microsatellite markers to facilitate germplasm conservation and population genetics studies of five groups of tropical perennial plants with edible fruits and shoots: rambutan (*Nepheleium lappaceum* L.), sapodilla (*Manilkara zapota* (L.) P. Royen), lychee (*Litchi chinensis* Sonn.), mangosteen (*Garcinia mangostana* Linn. and *Garcinia cochinchinensis* (Lour.) Choisy) and bamboo (*Bambusa vulgaris* Schrad. ex JC Wendl and Guadua angustifolia Kunth). Genetic Res Crop Evol 67:1715–1731

Bhat ZA, Dhillon WS, Rashid R (2010) The role of molecular markers in improvement of fruit crops. Not Sci Biol 2:22–30. https://doi.org/10.15835/nsb.2.2.4222

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170

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 32:314–331

Buschiazzo E, Gemmell N (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. BioEssays 28:1040–1050. https://doi.org/10.1002/bies.20470

Earl DA, VonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Res 4:359–361

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5:435

Excoffier L, Laval G, Schneider S (2005) Arlequin: an integrated software package for population genetics data analysis. Evol Bioinforma 1:47–50. https://doi.org/10.1111/j.1755-0998.2010.02847.x

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. https://doi.org/10.3732/apps.1600024

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. https://doi.org/10.1007/s00438-015-1147-y

Jalawadi S (2014) Molecular characterization of hybrids and landraces of sapota by. Asian J Hortic 9:72–75

Jalawadi S, Kanamadi VC, Basavarajappa HR, et al (2014) Molecular characterization of sapota Genotypes Using RAPD Markers. In: Acta Horticulturae. pp 61–68

Kaeuffer R, Reale D, Colman DW, Pontier D (2007) Detecting population structure using STRUCTURE software: effect of background linkage disequilibrium. Heredity (edinnb) 99:374–380. https://doi.org/10.1038/sj.hdy.6801010

Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. Mol Ecol 16:1099–1106. https://doi.org/10.1111/j.1365-294X.2007.03089.x

Kumar M, Saraswathy S, Ramesh Kumar S et al (2015) Genetic diversity analysis in Sapota Cultivars as revealed by RAPD markers. Environ Ecol 33:898–900

Meghala R, Ravishankar KV, Anand L, Rekha A (2005) Genetic diversity of Indian sapota (*Manilkara zapota*) cultivars characterized by RAPD markers. Plant Genet Resour News 1:43–46

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

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. https://doi.org/10.1186/2193-1801-3-431

Rastegar S (2015) Physical, Biochemical and Mineral Evaluation of Sapota Fruits During Growth, Development and Ripening. Agric Commun 3:14–19

Ravishankar KV, Anand L, Dinesh MR (2000) Assessment of genetic relatedness among mango cultivars of India using RAPD markers. J Hortic Sci Biotechnol 75:198–201. https://doi.org/10.1080/14620316.2000.11511223

Ravishankar KV, 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. https://doi.org/10.3732/ajb.1000263

Ravishankar KV, Chaturvedi K, Putteraju N et al (2015a) Mining and characterization of SSRs from pomegranate (*Punica granatum* L.) by pyrosequencing. Plant Breed 134:247–254. https://doi.org/10.1111/pbr.12238

Ravishankar KV, Dinesh MR, Nischita P, Sandya BS (2015b) Development and characterization of microsatellite markers in mango (*Mangifera indica*) using next-generation sequencing technology and their transferability across species. Mol Breed 35:93. https://doi.org/10.1007/s11032-015-0289-2

Ravishankar KV, Bommisetty P, Bajpai A et al (2015c) Genetic diversity and population structure analysis of mango (*Mangifera indica*) cultivars assessed by microsatellite markers. Trees 29:775–783. https://doi.org/10.1007/s00468-015-1155-x

Ravishankar KV, Vasudeva R, Hemanth B et al (2017) Isolation and characterization of microsatellite markers in *Garcinia guinmi-gutta* by next-generation sequencing and...
cross-species amplification. J Genet 96:213–218. https://doi.org/10.1007/s12041-017-0756-0
Rekha A, Dinesh MR, Venugopalan R, Murthy BNS (2011) Genetic correlation and cluster analysis in sapota (Manilkara zapota). J Hortic Sci 6:101–104
Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. Nat Biotechnol 18:233–234. https://doi.org/10.1038/72708
Siddiqui MW, Longkumer M, Ahmad MS et al (2014) Postharvest biology and technology of sapota: a concise review. Acta Physiol Plant 36:3115–3122. https://doi.org/10.1007/s11738-014-1696-4
Smarda P, Bures P (2012) The variation of base composition in plant genomes. Plant Genome Diversity 1:209–235
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. https://doi.org/10.1371/journal.pone.0021298
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. https://doi.org/10.3389/fpls.2015.01074

Wagner HW, Sefc KM (1999) IDENTITY 1.0. Centre for Applied Genetics. University of Agricultural Sciences, Vienna, Austria
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. https://doi.org/10.1186/1471-2105-9-253
Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829. https://doi.org/10.1101/gr.074492.107
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. https://doi.org/10.1007/s00122-011-1689-2

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.