TIU Transactions on Human sciences

Mining, analysis, and development of SSR-FDM marker in papaya using EST sequence

Aheria Dey [a] , Saurav Bhattacharya [b]

[a, b] *Techno India University, Kolkata, West Bengal, India, 700091*

Abstract

SSRmarker has proved valuable in plant genomic studies due to its polymorphic nature and high transferability rate across taxa. SSRmarkers developed from conserved coding genomic regions by utilizing ESTsequences in papaya are extremely useful as functional markers as they have better chances of association with protein-coding regions. In the present study, we have made utilized publicly available EST sequence data of papaya to mine andidentify and develop markers from annotated EST-SSR sequences. The papaya EST sequence (77528) retrieved from NCBI was processed and assembled using EGassembler. Using the MISA and KRAIT programmes, 144 non-redundant perfect SSRs were selected from assembled contig and singleton sequences. Mono/dinucleotide repeats (80.5%) were the most common among SSR motifs found in the papaya EST sequence, followed by a trinucleotide (12.5%), tetranucleotide (4.86%), and hexanucleotide repeat (2.08%). AG/AT, AAT, and AAG repetitions were discovered to be prevalent among the SSR repeat sequence types. Furthermore, the amino acids coded by each trinucleotide SSRs motif had the greatest distribution of Lysine, followed by Arginine, Asparagine, Isoleucine, Threonine, and Serine. Interpro study utilizing Blast2GO to annotate non-redundant EST-SSR sequences identified numerous significant functional domains such as AP2/ERF, 14-3-3 protein, calmodulin binding, WRKY transcription factor, and metal ion binding.The bulk of SSRs (67%) were found in the UTR region, with just 33% found in the coding region. Primer pairs for 128 SSR sequences in and around the ORF region were created using Primer3 and virtually tested for desired amplification with the FastPCR programme. The EST-SSR markers and characterized functional domains in our study will facilitate genetic diversity and marker-assisted breeding studies in papaya.

Keywords: FDM, EST-SSR, papaya, in silico.

1. Introduction

The papaya (*Carica papaya* L.) is a tropical and subtropical fruit crop with considerable nutritional and therapeutic value. [1–6].The genus is diploid (2n=18) with a relatively small genome size of 372Mb and represented by only one species[7][8,9]with a base composition of about 63.49% AT and 36.51% GC [10]. Papaya is a perennialcropthat produces monoecious, dioecious, and bisexual flowers[11].Molecular markers and PCR-based approaches for mapping different loci for papaya plant sex determination have previously been published[12–14]. Furthermore, genetic, and genomic investigations on papaya genotypes utilizing polymorphic markers[15–21], gene expression dies on fruit quality, biotic and abiotic stress [22–26], indels, SNPs, association studies and structural variations [27,28] have been reported. Several studies on papaya have revealed significant genetic level variability in morphological (phenotypic) features as well as tolerance to biotic and abiotic stress[29–32]. Previously, most papaya breeding programmes relied on molecular markers (RAPD, ISSR, and AFLP) produced from genomic loci that were mostly connected with non-coding portions of the genome, rendering them inefficient in functional trait-related investigations. Studies that have contributed to the establishment of linkage maps and the use of sequencing methods and SSR loci, particularly in papaya, are limited[9,33,34]. Early studies on the origins of papaya and the genetic relationship among cultivated and wild relatives relied primarily on isozyme and dominant markers [35–38]. Studiesutilizingthe hypervariable microsatellite regions for the generation of SSR markers for cultivar fingerprinting, diversification analysis and sex determination have been described in papaya [30,32,39–41]. Recently SNPs and indels linked with ripening-related genes through whole-genome sequencing were also reported in papaya[24,27]. Keeping up with the expansion in accessible genetic and genomic resources in papaya, particularly the availability of genomic and transcriptomic sequences, comparable research utilizing omics-based techniques for functional evaluation of papaya genotypes and identifying important features associated with fruit ripening and stress tolerance must be developed.

Microsatellites are short stretches of repetitive DNA of about 1-10 nucleotides, constituting a major portion of junk regions within the DNA. These microsatellite regions are often referred to as SSRand they are often

associated with genetic instabilities and thus have various evolutionary contributions [42,43]. Single-locus codominant markers, like SSRs (microsatellites), have been shown to be more robust, polymorphic, reproducible, and have high cross-taxon transferability rates, permitting high-throughput DNA typing when compared to traditional multi-locus markers (RAPD, ISSR, and AFLP) [44]. These SSR markers can be found in both protein-coding and non-coding sections of the genome and exhibit extensive genome coverage [43].Traditional SSR marker or genomic SSR marker development faces several constraints, including the need for sequence information, distribution in both transcribed and non-transcribed genomic regions, a time and labor-intensive method of development, and sensitive detection and analysis methods, which have made de novo SSR marker development a difficult task[45]. Moreover, in several reportsamplified loci were found to be species-specific and less useful in inter-taxon or larger groups[46]. On the contrary, genic SSRs or EST-SSRs which are derived from expressed cDNAs offer an important low-cost alternative by exploiting publicly available genomic resources and sequence data for searching, identification and characterization of deposited SSR sequences from the transcribed genomic regions [47–49]. EST-derivedmarkers have better transferability across taxa than genomic SSR markers, stronger association, and physical linkage with expressed genes as they are designed from coding regions and represent functional domainmarkers (FDM) that are critical locations for effective marker-assisted selection. [49,50]. Such expressedsequences (ESTs) derived from the entire expressed cDNA pool represent robust functionallyannotated marker sequences with predicted protein domain signatures. Hence these DNA markers derived from functionally defined and validated sequences have better chances ofassociation with polymorphic traits so that can be successfully employed in molecular breedingapproaches [51–53].Functional markers also sometimes referred to as Functional Domain Markers (FDM) are a more modern concept contrary to random genome markers as they tend to functionally characterize an allele sequence from which the polymorphic sequence is identified. It is crucial that the gene of interest for which markers are developed be annotated with its function for exploring the true potential of this technique. Hence, SSR FDM is superior to other molecular markers in various aspects like polymorphism, association with coding genes, sequence targeted marker development, and does not require prior mapping [54,55]. It is also useful for genetic mapping studies, for studying genetic diversities, polymorphism detection, and interspecific breeding in multiple plant species[54,56].As a result, the current work aims to rapidly explore the available EST resources for papaya and design and verify functional markers using in silico techniques.

2. Materials and Methods

2.1. Retrieval, Processing and Assembly of EST sequence

The available ESTs were trimmed at 5' end or 3' end for any poly-A or poly T stretches and further cleaned for vector and adaptor sequence contamination, low-complexity filtering and other contamination. The cleaned quality sequences were subjected to contig assembling with default parameters with standalone processing using 6 CPUs for data analysis. All deposited raw EST sequences of *Carica papaya* were retrieved from National Centre for Biotechnology Information (NCBI) and downloaded in Fasta format. A total of 77,528EST sequences were downloaded representing expressed cDNA data from different plant tissues (leaves, roots) grown under variable growth conditions.All steps of EST pre-processing, clustering, and assembling were done using the online web server EGassembler (https://www.genome.jp/tools/egassembler/)[57]. Both the contig and singleton outputs were combined to form non-redundant sequence data. The final non-redundant contiguous sequence (contig) output file was downloaded and saved in FASTA format and further utilized for SSR mining.

2.2. SSR mining and identification of functional domain marker

For SSR identification, assembled contig sequences were only used and mined for SSR-containing regions using KRAIT Program[58]and cross-checked using the MISA tool (MIcroSAtellite identification tool;http://pgrc.ipkgatersleben.de/misa/misa.html)[59] with default search settings. SSR containing contig sequences were further analyzed for FDMsusing the InterProScan tool in the Blast2GO program [60].InterProScan provides the platform to analyze functional domains with the help of member databases, such as BlastProDom, FPrintScan, HMMPIR, HMMPfam, HMMSmart, HMMTigr, ProfileScan, HAMAP, PatternScan, SuperFamily, SignalPHMM, TMHMM, HMMPanther, and Gene3D. Additionally, EST-SSR sequences were searched for significant matches against a non-redundant protein database using BLASTx(http://blast.ncbi.nlm.nih.gov/Blast.cgi) and further mapped to obtain associated GO hits. GO hits from Interpro scan and BLASTx were further pooled to obtain functional GO annotations and GO terms depicting three descriptors i.e., biological process (BP), cellular component (CC) and molecular function (MF) were assigned to them. All the analyses were performed using the Blast2GO program. SSR contigs were annotated and mapped against the KEGG database to retrieve the enzyme commission (EC) IDs.

2.3. SSR Primer designing, ORF prediction and *in silico* amplification

Primers were designed for SSR containing contig sequences using Primer3 software with default parameters: optimum primer size = 20.0 (range of 18–27), optimum annealing temperature = 60.0 (range of 57.0–63.0), GC content of 20–80 %.Open reading frames (ORFs) were predicted for all SSR-containing sequences using the ORF Finder available at NCBI using standard genetic code. The relative position of the SSR motifi.e., whether the SSR was present within the ORF, in the 5' or 3'-untranslated region (UTR) was also recorded. The screened primers were then crosschecked for *in silico* amplification by FastPCR[61]version 6.5.

1). There were 2 (AAC/TTG)n, 7 (AAG/TTC)n, 1 (ACC/TGG)n, 1 (ACG/TGC)n, 1 (AGC/TCG)n, 4 (AGG/TCC)n, 2 (ATC/TAG)n trinucleotide repeat motif types detected. The type and number of tetra-nucleotide repeats were – (AAAC/TTTG)n, (AAAG/TTTC)n, (AAGG/TTCC)n, (AATC/TTAG)n, (AATT/TTAA)n, (ATGC/TACG)n, (AATG/TTAC)n of each type. Lastly thehexanucleotide repeats were (AAAAAG/TTTTTC)n, (AAA ACC/TTTTGG)n, (AAA GCC/TTTCGG)n of each type. Of the trinucleotides, SSRs the amino acids coded by them showed the maximum distribution of Lysine, followed by Arginine, Asparagine and Isoleucine, Threonine and Serine (Fig 2).

Table 1. Summary of motif types detected from 144 perfect SSR from assembled EST sequence.

Type	Counts	Length (bp)	Percent (%)	Average Length (bp)	Relative Abundance (loci/Mb)	Relative Density (bp/Mb)
Mono	86	1185	59.72	13.78	65.45	901.77
Di	30	456	20.83	15.2	22.83	347.01
Tri	18	294	12.5	16.33	13.7	223.73
Tetra	7	116	4.86	16.57	5.33	88.27
Hexa	3	78	2.08	26.0	2.28	59.36

3. Results

3.1.Processing Assembly and distribution of SSR

A totalof 77,528 *C. papaya* ESTs were collected from NCBI's dbEST database (https://www.ncbi.nlm.nih.gov/) representing different plant parts (leaf, stem, and roots) grown under different biotic and abiotic conditions. Following assembly, a non-redundant group of ESTs was assembled that included contigs and singletons, hereafter referred to as "assembled EST sequences." Following the elimination of duplicated, junk, and other undesirable sequences, 232 contig sequences and 1629 singleton sequences were recovered. MISA and the Krait tool detected 16 perfect SSRs out of 232 contigs, while 128 perfect SSRs were discovered out of 1629 singletons.By analysing the sequences, 144 perfect SSRs were found, of which the maximum wasmono/dinucleotide repeats (cumulative) (116 / 80.55%), followed by trinucleotide repeats (18 / 12.5%), tetranucleotide repeats (7 / 4.86%) and hexanucleotide repeats (3 / 2.08%). There were 69 (A/T)n, 17 (C/G)n, 5 (AC/TG)n, 17 (AG/TC)n, 8 (AT/TA)n mono/dinucleotide repeats motif types (Fig

3.2. Protein annotation and Functional domain marker (FDM) analysis of SSR-ESTs

A total of 144 SSR-containing sequences were analysed for FDMs, of which 58 FDM markers from singleton and 16 FDM markers from contig sequences were detected. The SSR-containing sequences were subjected to Blastx through NCBI Blast and Blast2GO, and out of 144 SSR containing sequences, 66 (50 out of 128 singleton SSR sequences and all 16 contig SSR sequence) showed that the SSRs lie within or near protein-coding genes. Each BlastX translated proteins were mapped to their Interpro IDs databases identified from databases such as SignalPHMM, TMHMM, HMMPanther, Pfam, SMART, Panther and Gene3Dand annotatedbased on the GO annotation categories (Biological process, Molecular process, andCellular function).Among the important functional domains identified were AP2/ERF and B3 domain-containing transcription

factor, 14-3-3-like protein GF14 kappa isoform, EARLY RESPONSIVE TO DEHYDRATION 15, S-adenosylmethionine synthase, calmodulin-binding, ADP ribosylation GTPase activator, NAC domain, metal ion binding, DNA binding, WRKY transcription factor, Hydrophobic seed protein domain,malate synthetase domain,GATA transcription factor, zinc finger domain. GO terms were assigned to EST-SSRS with significant matches. The function of 144 SSR-containing sequences was annotated against the non-redundant (nr) protein database performed using the Blast2GO module. Annotations were recorded for a total of 94 (65%) sequences including contig and singleton sequences. The molecular function (MF) refers to the product function of the gene at the molecular level and includes the catalytic and binding activities of a gene. The molecular processes indicate that there was a pretty much even

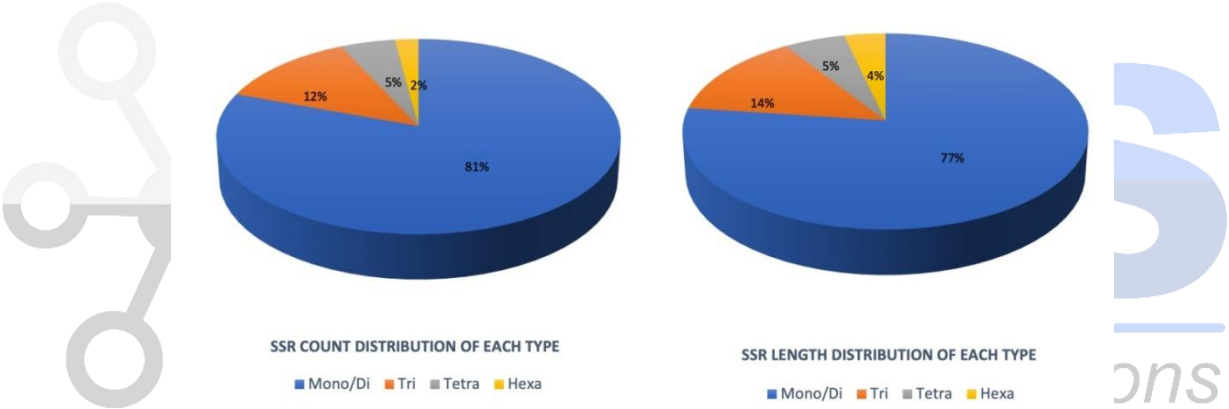

Fig 1. Frequency distribution of length and count of SSR motif types in assembled EST The sequence of papaya.

distribution of protein functions with SSR sequences, of which the most significant proteins with molecular functions were – Oxidoreductaseactivity(7%)(GO:0005506, GO:0016491), ATP Binding(7%) (GO:0004478), Transcription factors, peptidase activityGO:0008233) (GO:0004176), transferase activity(methionine adenosyl 3%, gamma-glutamyl 2%), nucleic acidbinding(DNA 8% RNA 2%) (GO:0043565) (GO:0003677), ion binding activity(Zinc 5%, iron 2%, heme 2%) (GO:0046872), bindingactivity(calmodulin 2%, FAD 2%,NAD 2%, protein 3%, ribosome 3%) (GO:0005516) and various otherenzymatic and structural functions (Fig 3).

A biological process (BP) is a series of events accomplished by one or more ordered assemblies of molecular function. Analysis of the biological Processindicates that there was a predominance of the analysed proteins in the proteolysis process (14%) (GO:0006508), followed by lipid biosynthesis and transcription regulation (11% each). The other significant processes include stress response (8%), metabolic processes (cellular nitrogen 5%, lipid 5%), biosynthetic process (5%), S-adenosylmethionine biosynthesis (5%), catabolic process (5%), phosphorylation (5%) and other minor processes (Fig 3).

The final GO category is the cellular component (CC) that describes subcellular structures and macromolecular complexes. GO-CC terms may thus be used to annotate cellular locations of gene products. The most abundant cellular process from annotated SSR-ESTs in our results is shown to be associated predominantly with nuclear functions (40%), followed by cytoplasmic function (25%) and the rest scattered in the cell wall, plasma membrane, mitochondria, ribosome, lysosome, and extracellular space.Additionally, several singleton EST sequences were found to be associated with unknown/uncharacterized domains/proteins, such as SSR55 (IPR005518) (Fig 3).

3.3. Prediction of ORF and in-silico PCR in C. papaya SSR-EST

We attempted to analyse the presence of ORF and its distribution in the SSR containing EST sequences using the ORF Finder tool. The distribution of SSR revealed that about 67% of them were located in the 5 and 3' UTR. Additionally, 34% of the SSRs were located downstream of the ORF, and about 33% had SSRs upstream of the ORF region. The rest 33% SSRs were located within the coding sequence (Fig 4).

Primer pairs were designed for SSR containing contig sequences using Primer3 software with default parameters: optimum primer size = 20.0 (range of 18–27), optimum annealing temperature = 60.0 (range of 57.0–63.0), GC content of 20–80 %. Out of 144, SSRs detected, it was possible to design primers for 128 sequences (89%). The representative 50 primer sequences for selected SSR contigs were given in Table 3.

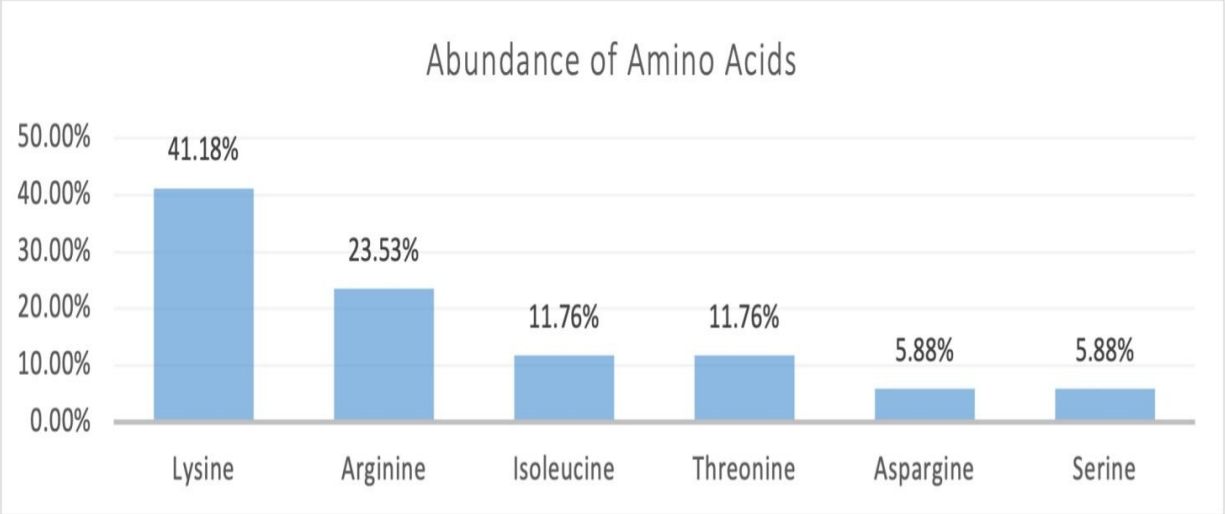

Fig 2. Amino acid distribution of translated trinucleotide motifs detected in assembled EST sequence.

4. Discussion

EST sequences provide valuable tools and information to develop polymorphic functional markers,particularly in crops like papaya with limited reports of genic markers. Out of the total 1861 assembled sequences, 144 SSR-containingsequences were identified, revealingthat 7.73% of the sequences contain SSR markers. Most of the SSR repeats weremono and dinucleotide repeats, which is about 81% of the total SSRs detected as similarly observed in previous reports in papaya[62]. This data is also comparable tothe findings observed in other plants such as Prunus 8.32% [63], Ocimum 7.79% [64], lesser than Euphorbia 11.77% [65], but higher than Arabidopsis 2.4% [66], rice 4.7%, wheat 3.2%, and maize 1.5% [67].

The relative density of SSRs in papaya was found to be about 1.62 SSRs per kb which is highly frequent. This data is also in congruence to densities observed in previous EST analysis in papaya (1SSR per 0.7kb)[68,69], Prunus [63], Solanaceae chloroplast genome (1.26 SSRs per kb) [70], while relatively higher than Citrus (0.51 SSRs per kb) [71], Barley (0.13 SSRs per kb), Maize (0.13 SSRs per kb), Wheat (0.16 SSRs per kb), Rice (0.26 SSRs per kb) [72], Euphorbia (0.18 SSRs per kb) [65] and Mentha (0.29 SSRs per kb)[73]. Our results showed that mononucleotide and dinucleotide repeats were the most abundant corresponding to 81% followed by trinucleotide repeats at 12%.These findings are consistent with previous reports on papaya [73] and Mentha (0.29 SSRs per kb) [73]., Our results showed that mononucleotide and dinucleotide repeats at 12%. These findings are consistent with previous reportson papaya [73], andMentha, Euphorbia, Prunus, and Ocimum. In mononucleotide, A/T repeats are most abundant and in dinucleotide AG/CT repeats are most frequent and were similarly reported inprevious studies in papaya [69] and plants such as Euphorbia [65], Mentha [73], and cereals like rice, barley and wheat [67]. It was also noticed that GC/CG repeats were absent in papaya as observed in Euphorbia, Mentha and various crop species. Among tri-nucleotide repeats, AAG/CTT was the most significant and supported by previous findings[68,69] (Fig 5).

Interpro scan analysis assigned 286 functionaldomains mapped to 144 SSR contigs.Interpro signature domains identified included a wide array of functionally important protein domains such as Zinc fingerdomains, S-adenosylmethionine synthetase domains, kinase and phosphatase domains,FAD/NAD/GTPase/NAC binding domains, 14-3-3 domain, F-box domain, Gamma-glutamylacyltransferase domain, WRKY domain.The GO annotations revealed 61 unique GO terms of which 40 were Molecular functions, 13Biological processes and 8 Cellular functions that revealed a significant functionally diverse mRNA pool in the expressed portion of the papaya genome. Such annotated data from contig sequences showed their matching to expressed proteins/domains potentially involved in various processes such as cell signalling, phytohormone signalling, fruit ripening, stress metabolism, and photosynthesis. Most enriched GO terms for molecular function (GO:0016491 and GO:0043565), biological process (GO:0006508, GO:0034641, GO:0003677), and cellular component (GO:0005634) indicated its association with fruit

development and associated signalling pathways in papaya.

Our data also revealed a predominance of SSR sequences in the UTR regions and relatively less dispersed in the ORF region of papaya transcriptome. The biasedness of SSR location to UTRs oftenindicates sequence conservation and evolutionary significance [74]. Being highly polymorphic/variable, this might appear significant to avoid large-scale changes in the protein coding regions. Additionally, significant number of SSR sequences have been identified in the papaya EST sequence ORF region in our analysis that were further utilized to design primers targeting SSR regions. Primer pairs for 128 SSR containing contig sequences were successfully designed and amplified using the Fast PCR tool. In addition to the basic quality check such as GC content, secondary structure, and 3'-complementarity, primer pairs were screened based on the result of the ORF-finder, and primers were selected only for those contigs having SSR motifs present at or close to the ORF region (SSR-FDM)tohave a better chance of association and segregation of the amplified SSR loci with the protein-coding regions.The transferability rate among the taxon and the degree of theirpolymorphic nature determines their suitability to be used as potential markers to be used in diversity, phylogenetic and molecular studies.EST-SSR has been utilized in several plants for genetic diversity analysis, linkage mapping, and varietal identification. Though microsatellite markers have previously been reported in papaya, the majority of them were derived from genomic DNA.EST- SSR markers reported in our study could be utilized for cultivar identification, sex determination among papaya genotypes, and marker-assisted breeding. The availability of genome sequence and huge transcriptomic data in papaya has paved the path for omics approaches to derive functional markers with better transferability rate and association with the trait of interest.

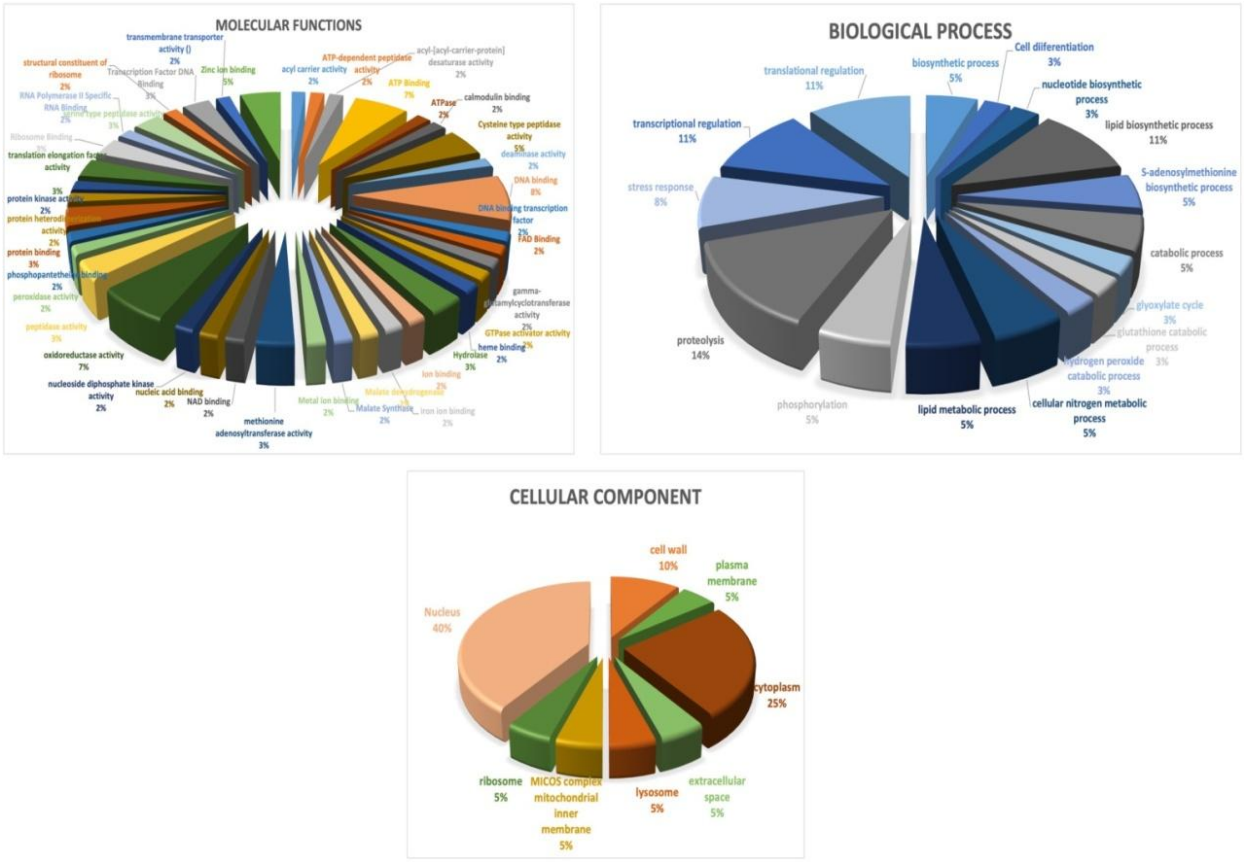

Fig 3. Graphical representation of GO annotated SSR containing EST sequences in papaya and its functional categorization into molecular function, biological process and cellular component.

Table 2: The abundance of repeat sequence across different SSR motif types in assembled EST sequence of papaya.

Motif	Type	Counts	Length (bp)	Percent (%)	Average Length (bp)	Relative Abundance (loci/Mb)	Relative Density (bp/Mb)
A	1	69	952	47.92	13.8	52.51	724.46
C	1	17	233	11.81	13.71	12.94	177.31
AC	2	5	72	3.47	14.4	3.8	54.79
AG	2	17	260	11.81	15.29	12.94	197.86
AT	2	8	124	5.56	15.5	6.09	94.36
AAC	3	2	30	1.39	15.0	1.52	22.83
AAG	3	7	117	4.86	16.71	5.33	89.04
ACC	3	1	15	0.69	15.0	0.76	11.41
ACG	3	1	18	0.69	18.0	0.76	13.7
AGC	3	1	15	0.69	15.0	0.76	11.41
AGG	3	4	66	2.78	16.5	3.04	50.23
ATC	3	2	33	1.39	16.5	1.52	25.11
AAAC	4	1	16	0.69	16.0	0.76	12.18
AAAG	4	1	16	0.69	16.0	0.76	12.18
AAGG	4	1	16	0.69	16.0	0.76	12.18
AATC	4	1	16	0.69	16.0	0.76	12.18
AATG	4	1	16	0.69	16.0	0.76	12.18
AATT	4	1	16	0.69	16.0	0.76	12.18
ATGC	4	1	20	0.69	20.0	0.76	15.22
AAAAAG	6	1	24	0.69	24.0	0.76	18.26
AAAACC	6	1	24	0.69	24.0	0.76	18.26
AAAGCC	6	1	30	0.69	30.0	0.76	22.83

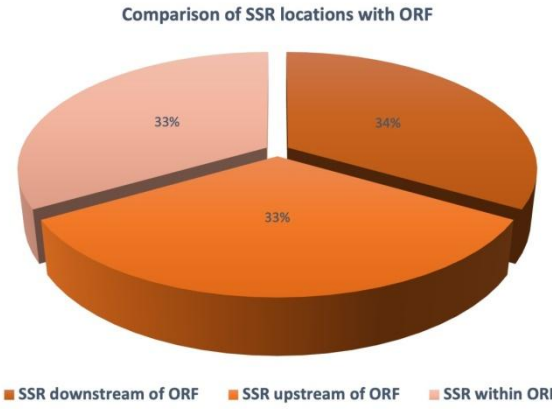

Fig 4. Distribution of location of SSR sequence in the coding and non-coding region of papaya EST.

Table 3. List of SSR primers with amplified product size derived from annotated EST-SSR sequence of papaya

Sequence	FORWARD PRIMER	REVERSE PRIMER	Amplicon Length (b.p)
Contig 29	GGACACTGTTGCTGTTGGAA	GCGATGGTTGATTCTGGTTT	303
Contig 230	GCTGCACTTGCTCTCTTCCT	GCATGGAGTCACTGGAGGTT	216
Contig 79	TGGTTCGAGGAGCTTCTAACAT	TGAACATGGGCTCCTACGAT	434
Contig 10	GCCTTTTGTGCTCGTCTAACTTT	AGCTGATGTATCAATCCGTTCTC	459
Contig 33	GTGGCAATAGCAGGTTCTTGAA	GGGCTAACTGCTTCCATTGTC	360
Contig 38	GTTCATACAAGAAGGGCCTCAA	GATCACCACCTACGAGGGAA	560
Contig 167	AATTTCCAGCAGATCGACGTT	CCGGGATTTCTTAGGAAACAAGT	641
Contig 174	GAGGCTGAAATCGCATGAGA	ATCTCCGGGGCAACTTGTAA	616
Contig 229	AACGAGAGAAGAAATCAGGGTTT	GGTCCTTGGCTAGCTGAATTT	489
Contig 23	CTGCAGGTATGGCTGTATTTGAT	TTCATTCATAAACAGTGGCAAGC	683
Contig 39	TTAACACTCCACTCCTCCAAA	GACGCTAAGATCTGCTCCAA	462
Contig 49	CCCCGCAGTAGAAAACTATCA	CAGCATCCTCAGAGGTCTCTT	482
Contig 128	CTAAAAGTGACACACACACACAC	CGTGGCTCAAAATGGTATTCACT	500
Contig 160	GTTTTTGCACGAGGAAACGTATA	CGATCGTTCTTGTCCAGTCATA	312
>gi|159518724|emb|am904489.1|	TCATATGGTGGATGCGAAGA	GGTTTTGGAGCAACGAACTG	240
>gi|159518425|emb|am903941.1|	CCCTTTCATCTCCATCTCCA	CTTTTCATCAGCGTCAGCAG	213
>gi|159518222|emb|am903738.1|	TGCAGAGAAAAGGGAAGCAC	CGGCAAGGGAAGAGTAAACA	
>gi|159518116|emb|am904288.1|	TCCCCATCTCTCCTCTTCCT	TCCAAACAGCAGAACACAGC	371
>gi|159518062|emb|am904234.1|	TTCCCTCAGATACTCGTTGCT	CTTTCCAGTAACCGGCATCA	544
>gi|159517953|emb|am904125.1|	TCTCTCTCTCTCTCTCTACGG	CCTTAGTGTTTCCATTCTCAGTC	362
>gi|159517897|emb|am904069.1|	CTCTTTGCAGCGTTCTGTCT	GAGGAACGTTTTGACCAGTCTT	465
>gi|159517855|emb|am904027.1|	CACTTGAGGCTTCACCCATT	GAAAACTCCCCTTGGCTTCT	217
>gi|159517729|emb|am903464.1|	CTTGTGGCATGCATGATAAACT	GAGAAGTTCAGCGAATCCCTC	479
>gi|159517679|emb|am903667.1|	TACCTCTTCTTTATTCTCGCTCC	ACCAGTAACTGAAGCATCAGGT	315
>gi|159517655|emb|am903643.1|	GCCGTGTCTTCTATCAAGGA	GCCTTACTTCATTAGCGAGATCT	538
>gi|159517521|emb|am903548.1|	CGTCGACAGACAATCCAAGA	TTTTCCCTTCCAAAGCTCCT	274
>gi|186882785|gb|ex292624.1|	CCGTTCACAGTCCCATCTTT	ACACCCTCCCAAACCATACA	366
>gi|186882659|gb|ex235860.1|	GCTGCTTTTGCGAGTTATCC	ATCAACCGCGCAGATATTTT	322
>gi|186882598|gb|ex255894.1|	CAGTCTCGCTCTCCTTCCAT	CAATCAAACTCGGGGTCAAT	297
>gi|186881775|gb|ex264273.1|	ACCGGAGGCCATTGTAAAGA	CCCGAGGGACAGCTGTTAAT	258
>gi|186881706|gb|ex249127.1|	CCCAGGAGTGTGTGGACTTTATA	CCATAAGTGTACAAGCGATACG	124
>gi|186881700|gb|ex303276.1|	CACACTCTTATTCCAAGCTTGC	GTTCAGTGTTTGACCCCTCAATA	325
>gi|186880913|gb|ex280246.1|	TGGATTTTCTGAGTTGCTTCGT	AGCATCCTGGGTGAGTTCAC	366
>gi|186880871|gb|ex247422.1|	TAAACGACTCTCGCCATCCT	GGCGTATTTGTTGGATCACC	203
>gi|159518769|emb|am904534.1|	AAAAGTGACACACACACACACA	GGAACCAAGGTAGATCATGCT	257
>gi|159518705|emb|am904470.1|	TTTTCAGCACACGAGACAGC	GTCGGGGAAGGATTACGATT	202
>gi|159518325|emb|am903841.1|	TCCTCTTCACTAGATTCCCAGAA	ACACAGCATTGCTCTGAACC	341
>gi|159518291|emb|am903807.1|	TCTCTCTCTCTCTCTGAAACGC	GCCAAGTAGCGGTGATAATCT	437
>gi|159518129|emb|am904301.1|	CCTTGTGGATATCTGGACAAAGC	GCACCATGCCTATTTGACTTCA	244

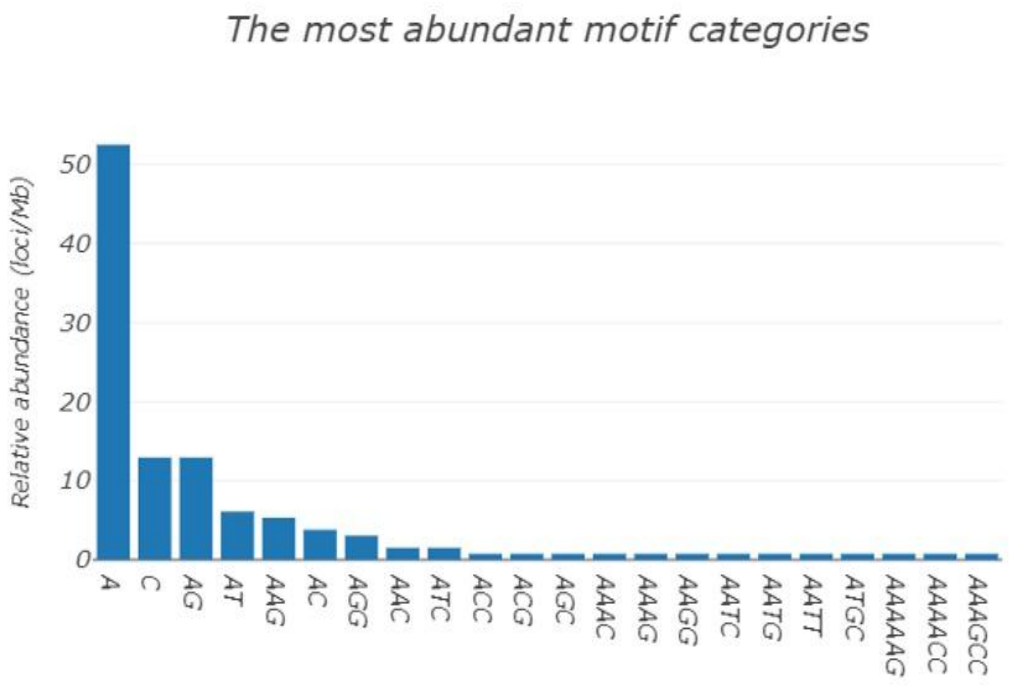

Fig 5. Figure showing relative abundance of the mined SSRs motif types from assembled EST sequence.

5. References

[1] De Oliveira, J. G., & Vitória, A. P. (2011). Papaya: Nutritional and pharmacological characterization, and quality loss due to physiological disorders. An overview. *Food Research International*, *44*(5), 1306-1313.

[2] Santana, L. F., Inada, A. C., Espirito Santo, B. L. S. D., Filiú, W. F., Pott, A., Alves, F. M., ... & Hiane, P. A. (2019). Nutraceutical potential of Carica papaya in metabolic syndrome. *Nutrients*, *11*(7), 1608.

[3] Nakamura, Yoshimasa, et al. "Papaya seed represents a rich source of biologically active isothiocyanate." *Journal of agricultural and food chemistry* 55.11 (2007): 4407-4413.

[4] Srivastava, A. K., & Singh, V. K. (2016). Carica papaya-A herbal medicine. *International Journal of Research Studies in Biosciences*, *4*(11), 19-25.

[5] Dotto, J. M., & Abihudi, S. A. (2021). Nutraceutical value of Carica papaya: A review. *Scientific African*, *13*, e00933.

[6] Vij, T., & Prashar, Y. (2015). A review on medicinal properties of Carica papaya Linn. *Asian Pacific Journal of Tropical Disease*, *5*(1), 1-6.

[7] Badillo, V. M. (2000). Carica L. vs. Vasconcella St. Hil.(Caricaceae) con la rehabilitación de este último. *Ernstia*, *10*(2), 74-79.

[8] Araújo, F. S., Carvalho, C. R., & Clarindo, W. R. (2010). Genome size, base composition and karyotype of Carica papaya L. *The Nucleus*, *53*, 25-31.

[9] Ming, R., & Moore, P. H. (Eds.). (2013). *Genetics and genomics of papaya* (Vol. 10). Springer Science & Business Media.

[10] Araújo, F. S., Carvalho, C. R., & Clarindo, W. R. (2010). Genome size, base composition and karyotype of Carica papaya L. *The Nucleus*, *53*, 25-31.

[11] Jiménez, V. M., Mora-Newcomer, E., & Gutiérrez-Soto, M. V. (2014). Biology of the papaya plant. *Genetics and genomics of papaya*, 17-33.

[12] Lee, C. Y., Lin, H. J., Viswanath, K. K., Lin, C. P., Chang, B. C. H., Chiu, P. H., ... & Chen, F. C. (2018). The development of functional mapping by three sex-related loci on the third whorl of different sex types of Carica papaya L. *PLoS One*, *13*(3), e0194605.

[13] Deputy JC, Ming R, Ma H, Liu Z, Fitch MMM, Wang M, et al. Molecular markers for sex determination in papaya (Carica papaya L.). Theoretical and Applied Genetics 2003 106:1 2002;106:107–11. https://doi.org/10.1007/S00122-002-0995-0.

[14] Ming, R., Yu, Q., & Moore, P. H. (2007, June). Sex determination in papaya. In *Seminars in cell & developmental biology* (Vol. 18, No. 3, pp. 401-408). Academic Press.

[15] Kim, M. S., Moore, P. H., Zee, F., Fitch, M. M., Steiger, D. L., Manshardt, R. M., ... & Ming, R. (2002). Genetic diversity of Carica papaya as revealed by AFLP markers. *Genome*, *45*(3), 503-512.

[16] Ming, R., & Moore, P. H. (Eds.). (2013). *Genetics and genomics of papaya* (Vol. 10). Springer Science & Business Media.

[17] Vidal, N. M., Grazziotin, A. L., Ramos, H. C. C., Pereira, M. G., & Venancio, T. M. (2014). Development of a gene-centered SSR atlas as a resource for papaya (Carica papaya) marker-assisted selection and population genetic studies. *PloS one*, *9*(11), e112654.

[18] Oliveira, G. A. F., Dantas, J. L. L., & Oliveira, E. J. (2015). Informativeness of minisatellite and microsatellite markers for genetic analysis in papaya. *Genetica*, *143*(5), 613-631.

[19] De Jesus, O. N., de Freitas, J. P. X., Dantas, J. L. L., & de Oliveira, E. J. (2013). Use of morpho-agronomic traits and DNA profiling for classification of genetic diversity in papaya. *Genetics and Molecular Research*, *12*(4), 6646-6663.

[20] Sengupta, S., Das, B., Prasad, M., Acharyya, P., & Ghose, T. K. (2013). A comparative survey of genetic diversity among a set of Caricaceae accessions using microsatellite markers. *SpringerPlus*, *2*, 1-10.

[21] Deputy, J., Ming, R., Ma, H., Liu, Z., Fitch, M., Wang, M., ... & Stiles, J. L. (2002). Molecular markers for sex determination in papaya (Carica papaya L.). *Theoretical and applied genetics*, *106*, 107-111.

[22] Paull, Robert E., et al. "Fruit development, ripening and quality related genes in the papaya genome." *Tropical Plant Biology* 1 (2008): 246-277.

[23] Fabi, João Paulo, et al. "Analysis of ripening-related gene expression in papaya using an Arabidopsis-based microarray." *BMC plant biology* 12 (2012): 1-19.

[24] Fabi, João Paulo, et al. "Analysis of ripening-related gene expression in papaya using an Arabidopsis-based microarray." *BMC plant biology* 12 (2012): 1-19.

[25] Gamboa-Tuz, S. D., Pereira-Santana, A., Zamora-Briseño, J. A., Castano, E., Espadas-Gil, F., Ayala-Sumuano, J. T., ... & Rodríguez-Zapata, L. C. (2018). Transcriptomics and co-expression networks reveal tissue-specific responses and regulatory hubs under mild and severe drought in papaya (Carica papaya L.). *Scientific Reports*, *8*(1), 14539.

[26] Estrella-Maldonado, H., Ramírez, A. G., Ortiz, G. F., Peraza-Echeverria, S., Martinez-de La Vega, O., Gongora-Castillo, E., & Santamaria, J. M. (2021). Transcriptomic analysis reveals key transcription factors associated to drought tolerance in a wild papaya (Carica papaya) genotype. *Plos one*, *16*(1), e0245855.

[27] Bohry, D., Ramos, H. C. C., Dos Santos, P. H. D., Boechat, M. S. B., Arêdes, F. A. S., Pirovani, A. A. V., & Pereira, M. G. (2021). Discovery of SNPs and InDels in papaya genotypes and its potential for marker assisted selection of fruit quality traits. *Scientific Reports*, *11*(1), 292.

[28] Liao, Z., Zhang, X., Zhang, S., Lin, Z., Zhang, X., & Ming, R. (2021). Structural variations in papaya genomes. *BMC genomics*, *22*(1), 1-13.

[29] Saran, P. L., Choudhary, R., Solanki, I. S., Patil, P., & Kumar, S. (2015). Genetic variability and relationship studies in new Indian papaya (Carica papaya L.) germplasm using morphological and molecular markers. *Turkish Journal of Agriculture and Forestry*, *39*(2), 310-321.

[30] de Oliveira, E. J., dos Santos Silva, A., de Carvalho, F. M., Dos Santos, L. F., Costa, J. L., de Oliveira Amorim, V. B., & Dantas, J. L. L. (2010). Polymorphic microsatellite marker set for Carica papaya L. and its use in molecular-assisted selection. *Euphytica*, *173*, 279-287.

[31] Coppens D'Eeckenbrugge, G., Restrepo, M. T., Jiménez, D., & Mora, E. (2005, November). Morphological and isozyme characterization of common papaya in Costa Rica. In *I International Symposium on Papaya 740* (pp. 109-120).

[32] Chávez-Pesqueira, M., & Núnez-Farfán, J. (2016). Genetic diversity and structure of wild populations of Carica papaya in Northern Mesoamerica inferred by nuclear microsatellites and chloroplast markers. *Annals of Botany*, *118*(7), 1293-1306.

[33] Nantawan, U., Kanchana-Udomkan, C., Bar, I., & Ford, R. (2019). Linkage mapping and quantitative trait loci analysis of sweetness and other fruit quality traits in papaya. *BMC plant biology*, *19*, 1-11.

[34] Chen, C., Yu, Q., Hou, S., Li, Y., Eustice, M., Skelton, R. L., ... & Ming, R. (2007). Construction of a sequence-tagged high-density genetic map of papaya for comparative structural and evolutionary genomics in brassicales. *Genetics*, *177*(4), 2481-2491.

[35] Lemos, E. G. M., Silva, C. L. S. P., & Zaidan, H. A. (2002). Identification of sex in Carica papaya L. using RAPD markers. *Euphytica*, *127*, 179-184..

[36] Jobin-Decor, M. P., Graham, G. C., Henry, R. J., & Drew, R. A. (1997). RAPD and isozyme analysis of genetic relationships between Carica papaya and wild relatives. *Genetic Resources and Crop Evolution*, *44*, 471-477.

[37] Sharon, D., Hillel, J., Vainstein, A., & Lavi, U. (1992). Application of DNA fingerprints for identification and genetic analysis of Carica papaya and other Carica species. *Euphytica*, *62*, 119-126..

[38] Stiles, J. I., Lemme, C., Sondur, S., Morshidi, M. B., & Manshardt, R. (1993). Using randomly amplified polymorphic DNA for evaluating genetic relationships among papaya cultivars. *Theoretical and Applied Genetics*, *85*, 697-701.

[39] De Oliveira, E. J., Amorim, V. B. O., Matos, E. L. S., Costa, J. L., da Silva Castellen, M., Pádua, J. G., & Dantas, J. L. L. (2010). Polymorphism of microsatellite markers in papaya (Carica papaya L.). *Plant Molecular Biology Reporter*, *28*, 519-530.

[40] Sengupta, S., Das, B., Prasad, M., Acharyya, P., & Ghose, T. K. (2013). A comparative survey of genetic diversity among a set of Caricaceae accessions using microsatellite markers. *SpringerPlus*, *2*, 1-10.

[41] Chiu, C. T., Wang, C. W., Chen, F. C., Chin, S. W., Liu, C. C., Lee, M. J., ... & Lee, C. Y. (2015). Sexual genetic and simple sequence repeat (SSR) analysis for molecular marker development on the all hermaphrodite papaya. *Genetics and Molecular Research*, *14*(1), 2502-2511.

[42] Gemayel, R., Cho, J., Boeynaems, S., & Verstrepen, K. J. (2012). Beyond junk-variable tandem repeats as facilitators of rapid evolution of regulatory and coding sequences. *Genes*, *3*(3), 461-480.

[43] Vieira, M. L. C., Santini, L., Diniz, A. L., & Munhoz, C. D. F. (2016). Microsatellite markers: what they mean and why they are so useful. *Genetics and molecular biology*, *39*, 312-328.

[44] Mishra, A., Singh, P. K., Bhandawat, A., Sharma, V., Sharma, V., Singh, P., ... & Sharma, H. (2022). Analysis of SSR and SNP markers. In *Bioinformatics* (pp. 131-144). Academic Press.

[45] Iniguez-Luy, F. L., Voort, A. V., & Osborn, T. C. (2008). Development of a set of public SSR markers derived from genomic sequence of a rapid cycling Brassica oleracea L. genotype. *Theoretical and applied genetics*, *117*, 977-985.

[46] Ellis, J. R., & Burke, J. M. (2007). EST-SSRs as a resource for population genetic analyses. *Heredity*, *99*(2), 125-132.

[47] Hu, J., Wang, L., & Li, J. (2011). Comparison of genomic SSR and EST-SSR markers for estimating genetic diversity in cucumber. *Biologia Plantarum*, *55*, 577-580.

[48] Ellis, J. R., & Burke, J. M. (2007). EST-SSRs as a resource for population genetic analyses. *Heredity*, *99*(2), 125-132.

[49] Varshney, R. K., Graner, A., & Sorrells, M. E. (2005). Genic microsatellite markers in plants: features and applications. *TRENDS in Biotechnology*, *23*(1), 48-55.

[50] Saha, M. C., Mian, M. R., Eujayl, I., Zwonitzer, J. C., Wang, L., & May, G. D. (2004). Tall fescue EST-SSR markers with transferability across several grass species. *Theoretical and Applied Genetics*, *109*, 783-791.

[51] Fan, M., Gao, Y., Gao, Y., Wu, Z., Liu, H., & Zhang, Q. (2019). Characterization and development of EST-SSR markers from transcriptome sequences of chrysanthemum (Chrysanthemum× morifolium Ramat.). *HortScience*, *54*(5), 772-778.

[52] Diola, V., Barbosa, M. H. P., Veiga, C. F. M., & Fernandes, E. C. (2014). Molecular markers EST-SSRs for genotype-phenotype association in sugarcane. *Sugar Tech*, *16*, 241-249.

[53] Yu, J. K., La Rota, M., Kantety, R. V., & Sorrells, M. E. (2004). EST derived SSR markers for comparative mapping in wheat and rice. *Molecular Genetics and Genomics*, *271*, 742-751.

[54] Andersen, J. R., & Lübberstedt, T. (2003). Functional markers in plants. *Trends in plant science*, *8*(11), 554-560.

[55] Yu, J. K., Paik, H., Choi, J. P., Han, J. H., Choe, J. K., & Hur, C. G. (2010). Functional domain marker (FDM): an in silico demonstration in Solanaceae using simple sequence repeats (SSRs). *Plant molecular biology reporter*, *28*, 352-356.

[56] Marconi, T. G., Costa, E. A., Miranda, H. R., Mancini, M. C., Cardoso-Silva, C. B., Oliveira, K. M., ... & Souza, A. P. (2011). Functional markers for gene mapping and genetic diversity studies in sugarcane. *BMC Research Notes*, *4*(1), 1-9.

[57] Masoudi-Nejad, A., Tonomura, K., Kawashima, S., Moriya, Y., Suzuki, M., Itoh, M., ... & Goto, S. (2006). EGassembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments. *Nucleic acids research*, *34*(suppl_2), W459-W462.

[58] Du, L., Zhang, C., Liu, Q., Zhang, X., & Yue, B. (2018). Krait: an ultrafast tool for genome-wide survey of microsatellites and primer design. *Bioinformatics*, *34*(4), 681-683.

[59] Beier, S., Thiel, T., Münch, T., Scholz, U., & Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. *Bioinformatics*, *33*(16), 2583-2585.

[60] Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*(18), 3674-3676.

[61] Kalendar, R., Khassenov, B., Ramankulov, Y., Samuilova, O., & Ivanov, K. I. (2017). FastPCR: An in silico tool for fast primer and probe design and advanced sequence analysis. *Genomics*, *109*(3-4), 312-319.

[62] Priyanka, P., Kumar, D., Yadav, A., Yadav, K., & Dwivedi, U. N. (2017). Analysis of Simple Sequence Repeats Information from Floral Expressed Sequence Tags Resources of Papaya (Carica papaya L.). *American Journal of Plant Sciences*, *8*(09), 2315.

[63] Sorkheh, K., Prudencio, A. S., Ghebinejad, A., Dehkordi, M. K., Erogul, D., Rubio, M., & Martínez-Gómez, P. (2016). In silico search, characterization and validation of new EST-SSR markers in the genus Prunus. *BMC Research Notes*, *9*(1), 1-11.

[64] Gupta, S., Shukla, R., Roy, S., Sen, N., & Sharma, A. (2010). In'Silico'SSR and FDM analysis through EST sequences in'Ocimum basilicum'. *Plant Omics*, *3*(4), 121-128.

[65] Sen, S., Dehury, B., Sahu, J., Rathi, S., & Yadav, R. N. S. (2018). Mining and comparative survey of EST–SSR markers among members of Euphorbiaceae family. *Molecular biology reports*, *45*, 453-468.

[66] Cardle, L., Ramsay, L., Milbourne, D., Macaulay, M., Marshall, D., & Waugh, R. (2000). Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics*, *156*(2), 847-854.

[67] Kantety, R. V., La Rota, M., Matthews, D. E., & Sorrells, M. E. (2002). Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant molecular biology*, *48*, 501-510..

[68] Arumugam, V., Riju, A., & Arunachalam, V. (2008, December). Mining of Expressed Sequence Tag (EST) Libraries and Core Nucleotide Sequences for Simple Sequence Repeats (SSR) in Papaya. In *II International Symposium on Papaya 851* (pp. 197-200).

[69] Wang, J., Chen, C., Na, J. K., Yu, Q., Hou, S., Paull, R. E., ... & Ming, R. (2008). Genome-wide comparative analyses of microsatellites in papaya. *Tropical Plant Biology*, *1*, 278-292.

[70] Tambarussi, E. V., Melotto-Passarin, D. M., Gonzalez, S. G., Brigati, J. B., de Jesus, F. A., Barbosa, A. L., ... & Carrer, H. (2009). In silico analysis of simple sequence repeats from chloroplast genomes of Solanaceae species. *Crop Breeding and Applied Biotechnology*, *9*(4).

[71] Palmieri, D. A., Novelli, V. M., Bastianel, M., Cristofani-Yaly, M., Astúa-Monge, G., Carlos, E. F., ... & Machado, M. A. (2007). Frequency and distribution of microsatellites from ESTs of citrus. *Genetics and Molecular Biology*, *30*, 1009-1018.

[72] Varshney, R. K., Thiel, T., Stein, N., Langridge, P., & Graner, A. (2002). In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cellular and Molecular Biology Letters*, *7*(2A), 537-546.

[73] Kumar, B., Kumar, U., & Yadav, H. K. (2015). Identification of EST–SSRs and molecular diversity analysis in Mentha piperita. *The Crop Journal*, *3*(4), 335-342.

[74] Zhang, L., Zuo, K., Zhang, F., Cao, Y., Wang, J., Zhang, Y., ... & Tang, K. (2006). Conservation of noncoding microsatellites in plants: implication for gene regulation. *Bmc Genomics*, *7*, 1-14.