PROCEEDINGS OF THE II[nd] INTERNATIONAL SYMPOSIUM ON PAPAYA

Convener

N. Kumar

Madurai, India

December 9-12, 2008

ISHS Section Tropical and Subtropical Fruits
ISHS Working Group on Papaya

Acta Horticulturae 851
January 2010

Mining of Expressed Sequence Tag (EST) Libraries and Core Nucleotide Sequences for Simple Sequence Repeats (SSR) in Papaya

V. Arumugam[1], A. Riju[2] and V. Arunachalam[3,4]
[1]Horticultural College & Research Institute, Tamil Nadu Agricultural Unjversity, Coimbatore - 641003, Tamil Nadu, India
[2]Aikkal, Kanul, Kannur, Kerala -670564, India
[3]Central Plantation Crops Research Institute, Indian Council of Agricultural Research, Kudlu P.O. Kasaragod - 671124, Kerala, India
[4]Genetic Transformation Laboratory, International Crops Research Institute for Semi-Arid Tropics (ICRISAT), Patancheru, Hyderabad - 502324 Andhra Pradesh, India

Keywords: *Carica*, fruit, in silico, molecular marker, transcriptome

Abstract

Papaya, an economically important fruit plant, is polygamous in nature. *Carica papaya* is a native of tropical America and a member of the family Caricaceae. Expressed Sequence Tags (ESTs) are fragments of gene transcripts that provide researchers a quick and inexpensive route for discovering new genes. Available EST resources (1283) and core nucleotide (303) sequences were mined to develop Sequences for Simple Sequence Repeat (SSR) maker and its abundance in the papaya genome. Computational tools MISA and ETRA were employed to find the microsatellite repeats or SSRs ranging from monomer to decamer from assembled ESTs, singlets as well as core nucleotide sequences. A total of 1301 and 1134 SSR were reported in EST and core nucleotide respectively. Class I type SSR were found in papaya genome at a frequency of 1 SSR/8.7 kbp and Class II SSRs were very highly frequent at 1 SSR/552 bp. Pentanucleotide repeats and mononucleotide repeats were found to be abundant in EST sequences and monomer repeats and dinulceotide repeats are abundant in core nucleotide sequences. A database was generated http://www.riju.mybioscience.com/papaya.

INTRODUCTION

Papaya plant has a diploid genome of 372 Mb size (Arumuganathan and Earle, 1991) distributed in nine pairs of chromosomes. Expressed Sequence Tags or ESTs are a cheap and quicker substitute to whole genome sequencing. They are providing researchers a means to study the partial sequencing of expressed genes and it is necessary for transcriptome profiling and gene discovery. Microsatellites or Simple Sequence Repeats (SSR) are stretches of DNA containing tandem repeats of mono, di, tri, tetra and above nucleotide units ubiquitously distributed throughout the genome. Molecular genetic markers developed from ESTs being genic markers and are highly useful in developing linkage maps and marker assisted breeding programs and population genetics (Ellis and Burke, 2007). Papaya genome is sequenced recently which opens up new scope for computational biology tools (Ming et al., 2008). They reported that striking amplifications in gene number within particular functional groups suggest roles in the evolution of tree-like habit, deposition and remobilization of starch reserves, attraction of seed dispersal agents, and adaptation to tropical day lengths. Hence new molecular markers developed from EST resources will be helpful in saturating the linkage maps and identifying marker-trait associations.

MATERIALS AND METHODS

Sets (1283) and core nucleotide (303) sequences of papaya were retrieved from NCBI of fruit monocarp and peel tissues (Embank accession numbers CF569397 - CF569400, CF588412, CO373888 - CO373912, DT527739 - DT527752, EL784267 - EL784289, AM903458 - AM904540, AM930710 - AM930842. These EST sequences were processed to minimize the sequencing errors and avoid redundant sequences and

Proc. II[nd] IS on Papaya
Eds.: N. Kumar et al.
Acta Hort. 851, ISHS 2010

grouped using prep software. We obtained 154 consigns and 703 singlet by this process. We used the in silico methods, MISA (Thiele et al., 2003), and ETRA (visual C++ Karaka et al., 2005) to locate simple sequence repeats from consigns, singletons and core nucleotide sequences. We identified two type of Sirs Class I (≥20 bop), or hyper variable markers and Class II (≤19 bop). We employed primer 3 (Whitehead Institute, Cambridge, MA, USA) to design left as well as right flanking sequence of the detected microsatellites. Putative information of singletons and conges were detected by compared with the non-redundant protein using BLASTX program of NCBI (Latched et al., 1997).

RESULTS AND DISCUSSION

We have used powerful SSR finding tools namely MISA (able to detect Sirs monomer to hexane including compound repeats) and ETRA (Dimmer to above defamer locator) to detect SSR from consigns, singlet and core nucleotide sequences. We found a total of 1194 SSRs, in which there are 18 compound type repeats were included with length 20 bp or more. SSRs located were categorized according to their size Class I (≥20 bp) and Class II (≤19 bp). Among the 1194 detected SSR, 74 represents Class I and remaining 1120 belongs to Class II. The frequency of eSSR were calculated 1 SSR at every 566 bp (total base pair examined is 676262). Class I type SSR were found in papaya genome at a frequency of 1 SSR/8.7 kbp and Class II SSRs were very highly frequent at 1 SSR/552 bp. In Class I type dinucleotide (20%) repeats were found to be more abundant followed by trinucleotide (14%), monomer (9%), and hexamer (9%) and above decamer repeats (9%). While considering the Class II type of repeats pentamer (55.10%) and hexamer (27%) found to be more abundant than other repeats. Overall pentamer repeats are seen abundant (51.8%) followed by hexamer (26.2%), monomer (6.8%), heptamer (4.9%) and dimer (3%) repeats (Table 1). Among the core nucleotide sequences, a total of 1134 SSRs (1026 perfect and 108 compound repeats) were found. Of these dimeric repeats are common and abundant in Class I SSRs and monomeric were highly frequent in Class II SSRs. SSR motifs are grouped into unique classes based on the property of DNA base complementarity (Jurka and Pethiygoda, 1995).

We have also found higher order repeats such as above decamer, 12 bp repeats to 30 bp repeats in papaya ESTs. In core nucleotides 1134 SSR sites were observed among that 221 belong to Class I and 913 belongs to Class II. Overall mononucleotide (52%), dinulceotide (30%) were found to be abundant than trinucleotide (7%) tetranucleotide (0.8%) and pentanucleotide (0.53%) repeats. Among the mononucleotide, 'A/T' was accounted 93.3%. Dinculeotide repeat class 'AG/GA/CT/TC' and 'AT/TA' were frequent than the other class. Frequency of SSRs in ESTs is reported as 1/11.8 Kb in rice; 1/23.8 Kb in soybean, 1/17.24 Kb in wheat and 1/28.32 Kb in maize (Gao et al., 2003).

Among the dimerism AG/GA/TC/CT type was more common than other groups in papaya ESTs. Among the trinucleotide repeats AAG/AGA/GAA/CTT/TTC/TCT class were dominant than other trinucleotide classes in papaya. The former class was found to be most frequent in dicot plants (Kumpatla and Mukhopadhyay, 2005). We have also located seven higher-level SSR with size 12 to 30 bp. These rare repeats can be potential markers in detecting polymorphism and mapping studies. Trimeric motifs are abundant in ESTs of many plant genomes than other types (Gao et al., 2003) whereas, dimerics were abundant in 38 dicot plants (Kumpatla and Mukhopadhyay, 2005). Our study shows that pentameric and hexameric repeats are dominant in the individual ESTs and monomeric patterns are prevalent in the core nucleotides of papaya. The assembled ESTs were translated using BLASTX option and the putative gene information like zinc finger (AN1-like) family protein, etc. were obtained. Present study gives additional microsatellites from ESTs for use in papaya genome analysis and details are available as public domain database http:/www.riju.mybioscience.com/papaya/.

Literature Cited

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, I., Zhang, Z., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein

database search programs. Nucleic Acids Research 25:3389-3402.
Arumuganathan, K. and Earle, E. D. 1991. Nuclear DNA content of some important plant species. Plant Molecular Biology Reporter 9:208-218.
Ellis, J.R. and Burke, J.M. 2007. EST-SSRs as a resource for population genetic analyses. Heredity 99:125-132.
Eustice, M., Yu, Q., Lai, C.W., Hou, S., Thimmapuram, J., Liu, L., Alam, M., Moore, P.H., Presting, G.G. and Ming, R. 2008. Development and application of microsatellite markers for genomic analysis of papaya. Tree Genetics and Genomics 4:333-341.
Gao, L.F., Tang, J.F., Li, H.W. and Jia, J.Z. 2003. Analysis of microsatellites in major crops assessed by computational and experimental approaches. Mol. Breed. 12:245-261.
Green, P. 1999. Phred, Phrap, Consed. http://www.phrap.org/phredphrapconsed.html [Online] http://www.ncbi.nlm.nih.gov/dbEST/ retrieved on 23. 03. 2008.
Jurka, J. and Pethiygoda, C. 1995. Simple repetitive DNA sequences from primates: compilation and analysis. Journal of Molecular Evolution. 40:120-126.
Karaca, M., Bilgen, M., Naci Onus, A., Gul Ince, A. and Elmasulu, S.Y. 2005. Exact Tandem Repeats Analyzer (E-TRA): A new program for DNA sequence mining. Journal of Genetics 84(1):49-54.
Kumpatla, S.P. and Mukhopadhyay, S. 2005. Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. Genome 48:985-998.
Ming, R., Hou, H., Feng, Y. and Yu, Q. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452:991-996.
Thiel, T., Michalek, V. and Graner, A. 2003. Exploiting EST data-bases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theoretical and Applied Genetics 106:411-422.

Tables

Table 1. Distribution of different types of SSRs observed in in silico analysis of papaya ESTs and core nucleotide.

SSR type	EST				Core nucleotide			
	Class I	Class II	Total	Percentage (%)	Class I	Class II	Total	Percentage (%)
Monomer	7	74	81	6.78	4	584	588	51.85
Dimer	15	25	40	3.35	75	268	343	30.24
Trimer	10	18	28	2.34	19	61	80	7.05
Tetramer	4	0	4	0.34	9	0	9	0.79
Pentamer	1	617	618	51.76	6	0	6	0.53
Hexamer	7	306	313	26.21				
Heptamer	1	57	58	4.86				
Octamer	0	14	14	1.17				
Nonamer	1	9	10	0.84				
Decamer	3	0	3	0.25				
Compound	18	0	18	1.51	108	0	108	9.52
Above decamer	7	0	7	0.59				
Total	74	1120	1194		221	913	1134	