HIBISCUS

ASIAN JOURNAL OF PLANT BIOLOGY

Website: http://journal.hibiscuspublisher.com/index.php/AJPB

Asian Journal of Plant Biology (AJPB)

EISSN: 2289-5868

Mining of single nucleotide polymorphism (SNP) and simple sequence repeats (SSRs) from EST tropical fruits

Z.A. Rabiatul Adawiah*, A.R. Shahril Firdaus, A. Norzihan and A.B. Umi Kalsom

Biotechnology Research Centre, MARDI Headquarters, Serdang, P.O. Box 12301, 50774 Kuala Lumpur, Malaysia

*Corresponding author: rabiatul@mardi.gov.my
Rabiatul Adawiah
Biotechnology Research Centre,
MARDI Headquarters, Serdang,
P.O. Box 12301, 50774 Kuala Lumpur, Malaysia

HISTORY

Received: 5th December 2014
Received In Revised Form: 24th December 2014
Accepted:27th December 2014

KEYWORDS

tropical fruits
simple sequence repeat (SSR)
expressed sequence tag (EST)
in silico
molecular marker

ABSTRACT

The advancement in genomics technology has produced vast amount of expressed sequence tags (ESTs) sequence from tropical fruits. These resources have increased the public availability of ESTs sequence from year after year. Therefore, this effort permits mining of single nucleotide polymorphism (SNP) and simple sequence repeat (SSR) from EST tropical fruits. SNP and SSR are types of molecular marker which commonly used in modern genetic analysis for wide application such as diversity analysis, linkage analysis and association study. In this study, a small scale EST sequences from tropical fruits (pineapple, mango, coconut and banana) were retrieved from dbEST database (www.ncbi.nlm.nih.gov/dbEST/) as of March 2013. Various bioinformatics tools were applied for rapid discovery of SNP and SSR marker from EST sequences. We analyzed 31,920 unigenes (contigs and singletons) representing a total of 77,418 ESTs from four tropical fruits for their potential use in developing SNP and SSR markers. A total of 13,709 EST-SNP were discovered while a total of 4853 EST-SSR were discovered from these four tropical fruits. The most abundant EST-SSR repeat is from trinucleotide (15,957 repeats) followed by dinucleotide (13,797 repeats) and tetranucleotide (973 repeats). Here, 1738 primers from SNP while 2033 primers from SSR were passed through the setting criteria and were selected for validation using genotyping platform. This study not only serves as a resource for marker development in tropical fruits but can provide a better insight into the selection of candidate genes of interest.

INTRODUCTION

The great demand on tropical fruits is constantly increasing when peoples are aware of the tropical fruits benefits. The major tropical fruits such as papaya, pineapple, mango, banana, star fruit and coconut are rich in vitamins, mineral and other phytonutrients that beneficial to human health. Therefore regardless of the demands, an effort has been made which used molecular marker technology to increase the production and quality traits of interest in tropical fruits.

Molecular markers are pieces of DNA that are known to be located near genes and inherited traits of interest [5]. There are many types of molecular marker such as amplified fragment length polymorphism (AFLP), random fragment polymorphic DNA, (RFLP), single sequence repeat (SSR), single nucleotide polymorphism (SNP) and random amplified polymorphic DNA (RAPD). In modern agriculture, microsatellite or known as SSR and SNP are among the preferred markers which commonly used in modern genetic analysis.

SSRs are tandem repeats of 2-6 nucleotides, multiallelic, co-dominant and reveal high level of polymorphism [9]. SNP is a single base change at DNA sequence with alternative of two possible nucleotides at a define position [20]. SNP is as a powerful marker due to their highly abundance and variants in organism [19]; [11]; [5]. SNP is less polymorphic but amenable to high throughput automation using current genotyping technology platform [14]. Whilst SSR marker is polymorphic because of its biallelic nature [14].

The set of SNP and SSR markers are useful in the construction of genetic map, to determine association between genotype and phenotype and for genetic diversity study [4]; [8]; [16]. Yet, although some knowledge has recently been acquired in the fields of molecular marker on tropical fruits, little is known about the application of SNP and SSR in tropical fruits [1]. The development of SNP and SSR markers in tropical fruits are not well established as compared to others fruit such as citrus[18]; [2] and rosacea[12]. SNPs have been developed only for a few tropical fruit; for example resistance to papaya ringspot virus (PRSV) in Carica papaya [6].To date, conventional markers such as SCAR, RAPD and RFLP are widely used for fingerprinting and linkage map in pineapple [3] and mango[13]. However, the development of these conventional markers is labor intensive and time consuming. Hence, there is a need to develop SNP and SSR markers in tropical fruits which can be achieved through the utilization of public domain expressed sequence tags (ESTs).

ESTs are sequence proportions of complementary DNA copies of mRNA that represent part of the transcribed portion of the genome in given condition [16]. ESTs are short sequences comprises about 200-800 nucleotide bases in length [16].A large amount of ESTs data in tropical fruits have been generated and deposited in public database (Table 1). This huge amount of ESTs sequence is now possible to mine SNP and SSR through utilization of bioinformatics tools.

Currently, mining *in silico* SNP and SSR markers from ESTs sequence tropical fruits are among the efficient approach. This computational approach was proven to be time efficient, cost effective and labour extensive [20]. This is due to the selection of the most reliable and robust marker before validation step through experimental approach. In addition, numerous bioinformatics software is available to mine SNP and SSR from ESTs sequence.

The development of SNP and SSR markers consist of two parts i) discovery of SNP and SSR and ii) validation of SNP and SSR. The SNP and SSR discovery can be achieved through sequencing and bioinformatics approach whilst the validation of SNP and SSR can be conducted using high throughput genotyping platform such as Sequenom Mass Array and ABI 373XL.

The SNP and SSR mining or known as *in silico* method refers to the polymorphism screening on sequences from different individuals using dedicated software [20]. This method determines either true polymorphisms or sequencing errors by establishing a likelihood algorithm of a particular locus for being polymorphic [20]. However, these putative SNP and SSR markers could not be fully utilized until these markers are experimentally validated.

In this paper, we present the computational approach to mine SNP and SSR markers in tropical fruits from publically available EST sequences. We have selected pineapple, coconut, mango and banana as our crop of interest. These selected tropical fruits are among the commercial tropical fruits with high diversity in Malaysia. The putative EST-SNP and EST-SSR markers from selected tropical fruits can be utilized for fingerprinting and association study. This information on putative EST-SNP and EST-SSR markers will enrich the numbers of SSR and SNP markers pools that enabled the development of more tropical fruits marker.

MATERIALS AND METHODOLOGY

Data source and pre processing data

A total of 77, 418 EST sequences from tropical fruits (pineapple, coconut, mango and banana) were downloaded from dBEST (http://www.ncbi.nlm.nih.gov/dbEST/). Vector contamination was removed from the sequences using SeqTrim tools (https://github.com/bastodian/SeqTrim). The clean EST sequences were then trimmed using est_trimmer.pl. The remained sequences were assembled using cap3 (http://seq.cs.iastate.edu/cap3.html) with selected parameters –p 90, -o 50. The assembly produced contigs and singletones for SSR and SNP marker discovery.

EST-SSR markers discovery

SSRs were detected using MIcroSAtellite (MISA) (http://pgrc.ipk-gatersleben.de/misa/), a microsatellite identification tool. We had set two criterias for selection of polymorphic SSRs; i. to select SSR with represents more than 2 alleles and ii. to select contigs and singleton containing less than 1 SSR marker. These criteria were performed to avoid redundance of SSR markers when designing the primer. A custom made Perl script is written to select qualified SSRs by following these criteria. A bioinformatic tool, cdbfasta (http://souceforge.net/projects/cdbfasta/) was performed to extract contigs and singleton with single polymorphic EST-SSR marker. Each species were assembled and annotated separately. This large scale analysis was running in Linux environment (Figure 1).

EST-SNP marker discovery

A customized bioinformatic pipeline developed by Michealmore Lab, Genome Centre, UC Davis California (http:// http://cgpdb.ucdavis.edu/SNP_Discovery_CDS/) was used to mine EST-SNP from EST sequences of selected tropical fruits. This EST-SNP discovery pipeline was developed based on CAP3 assembly and customized Python script. The contigs from CAP3 assembly were extracted for SNP mining using customized Phython script.

Stringent quality criteria were set during this process. We exclude SNP presents in multiple contigs which is more than 4 contigs to avoid redundancy. SNPs are also excluded if more than one SNP presence in 60bp downstream and upstream of SNP position. The following criteria were performed to select reliable and robust EST-SNP marker.

Table 1: Summary of EST tropical fruits in NCBI dbEST.

Bil	Crops	EST as of 20th August 2013	SNP	SSR
1.	*Carica papaya*	77, 528	None found	712
2.	*Musa spp. (banana)*	45, 714	None found	550
3.	*Mangifera indica (mango)*	1689	None found	462
4.	*Ananas (pineapple)*	5978	None found	289
5.	*Cocos nucifera (coconut)*	8142	None found	0
6.	*C. maxima*	73	None found	17
7.	*Dimocarpus longan*	196	None found	0
8.	*Psidium sp. (guava)*	12	None found	24
9.	*Litchi chinensis /lychee*	35	None found	27
10.	*Durio spp .*	3	None found	7
11.	*Garcinia mangostana*	165	None found	0
12	*Averrhoa carambola/belimbing*	1	None found	0
13.	*Nephelium lappaceum*	0	None found	7

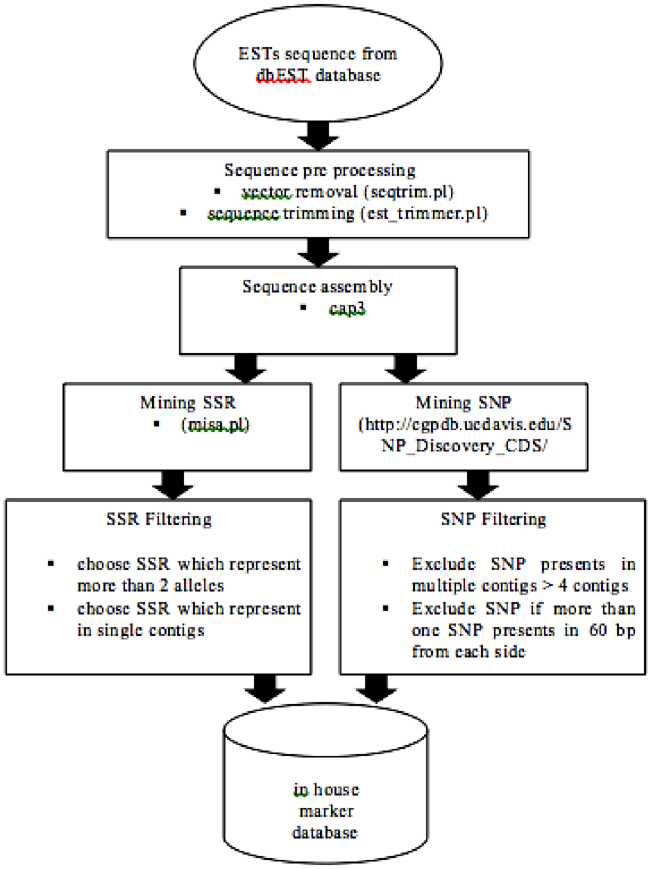
ESTs sequence from dbEST database
Sequence pre processing
▪ vector removal (seqtrim.pl)
▪ sequence trimming (est_trimmer.pl)
Sequence assembly
▪ cap3
Mining SSR
▪ (misa.pl)
Mining SNP
(http://cgpdb.ucdavis.edu/SNP_Discovery_CDS/
SSR Filtering
▪ choose SSR which represent more than 2 alleles
▪ choose SSR which represent in single contigs
SNP Filtering
▪ Exclude SNP presents in multiple contigs > 4 contigs
▪ Exclude SNP if more than one SNP presents in 60 bp from each side
in house marker database

Figure 1: Bioinformatics pipeline for SNP and SSR mining .

RESULTS AND DISCUSSIONS

Data pre processing and assembly of pineapple, mango, coconut and banana

Of the 77, 418 EST sequences, 31,920 unigenes (contigs and singletones) sequences were successfully pre processed and assembled. Among four tropical fruits, banana showed the highest number of unigenes and followed by mango. This was expected because ofhigh numbersof EST sequences in banana and mango as compared to number of EST sequences from pineapple and coconut.

EST-SNP & EST-SSR discovery of pineapple, mango, coconut and banana

Table 2 summarized number of SSR and SNP discovered in four selected tropical fruits. The analysis showed that the number of SNP markers are higher than SSR markers. This is due to the nature where SNPs are more abundant in genome as compared to SSR [15].

The highest SNP markers were observed in banana with a total of 11,322 SNP markers discovered. The results indicated the increased number of banana EST sequences in dbEST database. Mango showed the high frequency of SNP markers after banana with a total of 1755 SNP markers were discovered. However, among the four tropical fruits, mango showed the lowest frequency of SSR markers whilst banana has the highest frequency SSR markers with a total of 3523 SSR markeres.

The frequency of SNP and SSR markers in coconut are considered low after comparing to frequency of pineapple SNP markers. Even though the assembly of EST pineapple was lower than coconut assembly but the number of SSR and SNP markers were found higher than coconut. Low frequency of markers is due to less number of singleton in coconut EST sequences.

Repeat Types of EST-SSR markers

The repeat motifs of SSR markers are one of the important elements to choose the reliable marker to be validated in genotyping platform. Table 3 lists the distribution of repeat types in tropical fruits SSR. The predominance repeats type was dinucleotide with a total of 1350 occurrences followed by trinucleotide with a total of 1311 occurrences. Pentanucleotide is the least abundance repeat type with a total of 14 occurrences. However, tetranucleotide and pentanucleotide repeat types are not found in mango.

In this study, four motifs (AG, CT, AT and GA) were found to be the highest motif in dinucleotide repeat of pineapple, mango, coconut and banana (Figure 2). The high level of occurrencein GA/CT motifs could be due to the high level of occurrence of the translated amino acid products of the motifs [13]. Among the dinucleotide repeat in banana, GA motif is the highest motif and it has been observed 262 times (Figure 2). The least common repeat type is identified in mango where (CT)4. Meanwhile, TCG, AGA, AAG and CTC are the highest motif in the trinucleotide repeat. In banana, CTC motif is the most abundant motif with 59 frequency (Figure 3). The difference in polymorphism frequency observed among the different motifs may be due to the nature of motif [7].

Table 2: Summary of EST-SSR markers and EST-SNP markers.

	pineapple	mango	coconut	banana
Number of ESTs	5978	17 584	8142	45, 714
Number of contigs	725	1269	1636	5579
Number of singletons	2822	3268	1916	14, 705
Number of EST after sequence pre processing & assembly	3547	4537	3552	20, 284
Total number SSRs discovered	792	169	369	3523

Table 3: Distributions of different repeat type.

Repeat Type	pineapple	mango	coconut	banana	total
dinucleotide	330	20	56	944	1350
trinucleotide	236	75	96	904	1311
tetranucleotide	15	0	2	32	49
pentanucleotide	7	0	1	6	14
hexanucleotide	15	1	1	11	28

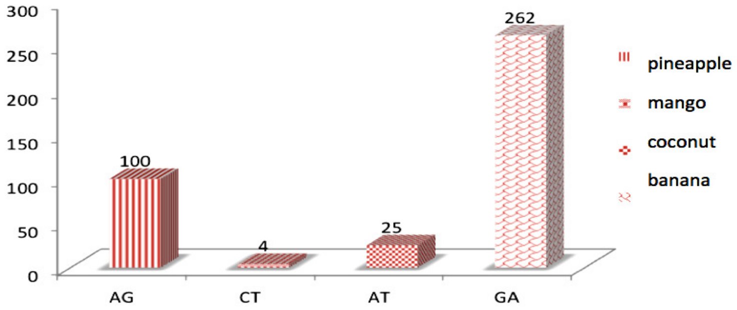

Figure 2: Distributions of the highest dinucleotide SSRs motif in pineapple, mango, coconut and banana

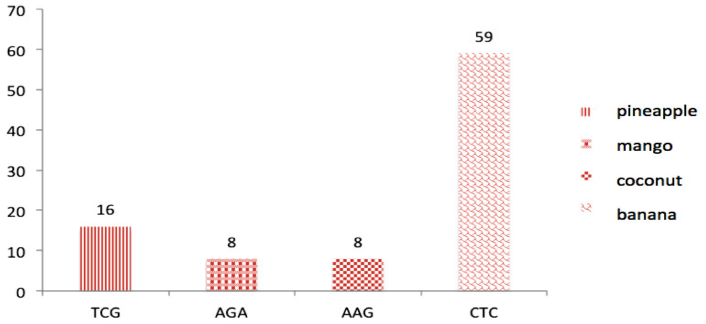

Figure 3: Distributions of the highest trinucleotide SSRs motif in pineapple, mango, coconut and banana

We managed to discover the total number of 4853 EST- SSR and 1738 EST-SNP in these selected tropical fruits. We have developed a pipeline for SNP and SSR mining from ESTs sequence using computational approach. This study is greatly improved the efficiency of SNP and SSR mining in the current scenario where the need of efficient and informative biological markers are required to explore the beneficial of tropical fruits.Putative EST-SNP and EST-SSR markers mined from these four tropical fruits are expected to be validated using genotyping platform such as Sequenom and ABI 3730XL.

Even though this article describes SNP and SSR discovery from ESTs sequencing, there is a growing interest in the mining of SSR and SNP from next generation sequencing (NGS) data. These data will increasingly be applied for the discovery and characterization of polymorphic SNP and SSR in wide range of species.

CONCLUSION

In summary, this preliminary study can be used as a platform for marker discovery and validation in tropical fruits by utilizing EST sequences using computational approach. The potential SNP and SSR markers may provide a molecular basis to be applied for genetic diversity and association study.

ACKNOWLEDGEMENT

This author would like to thank Dr Sanimah Simoh for reviewing this paper.

REFERENCES

[1] Arias RS, Borrone JW, Tondo CL, Irish BM, Schnell RJ. Genomics of tropical fruit tree crops. In: Schnell RJ and Priyadarshan PM, Genomics of Tree Crops: Springer Link. 2012; 210-239.

[2] Barkley NA, Roose ML, Krueger RR, Federici CT. Assessing genetic diversity and population structure in a citrus germplasm collection utilizing simple sequence repeat markers (SSRs). Theoretical and applied genetics. 2006; 112(8):1519-1531.

[3] Carlier JD, d'Eeckenbrugge GC, Leitão G, Pineapple. In: Genome mapping and molecular breeding in plants, fruits and nuts, (Kole, C.). 2007 (4); 331-342.

[4] Collard BCY, Mackill DJ. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philosophical Transactions of the Royal Society B: Biological Sciences. 2008; 363(1491): 557-572

[5] Duran C, Edwards D, Batley J. Molecular marker discovery and genetic map visualization. Bioinformatics: Tools and Applications, Springer Science. 2009; 165-189.

[6] Dillon S, Ramage C, Ashmore S, Drew RA. Development of a codominant CAPS marker linked to PRSV-P resistance in highland papaya. Theor Appl Genet. 2006; 113:1159–1169.

[7] Duran C, Singhania R, Raman H, Batley J, Edwards D. Predicting polymorphic EST-SSRs in silico. Molecular Ecology Resources. 2013; 13: 538-545

[8] Ganal MW, Altmann T, Röder MS. SNP identification in crop plants. Current Opinion in Plant Biology.2013; 12(2): 211-217.

[8] Hamilton JP, Hansey CN, Whitty BR., Stoffel K, Massa AN, Deynze AV, Jong WD, Douches DS, Buell CR. Single nucleotide polymorphism discovery in elite north american potato germplasm. BMC Genomics. 2011; 12(302):11.

[9] Ibitoye DO, Akin-Idowu PE. Marker-assisted-selection(MAS): A fast track to increase genetic gain in horticultural crop breeding. African Journal Biotechnology. 2011; 9(52):8889-8895.

[10] Imelfort M, Duran C, Batley J, Edwards D. Discovering genetic polymorphisms in next-generation sequencing data. Plant Biotechnology Journal. 2009; 7(4):312-317.

[11] Jena KK, Mackill DJ. Molecular markers and their use in marker-assisted selection in rice. Crop Science. 2008; 48(4):1266-1276.

[12] Jung S, Staton M, Lee T, Blenda A, Svancara R, Abbott A, Main D. GDR (Genome Database for Rosaceae): integrated web-database for Rosaceae genomics and genetics data. Nucleic acids research. 2008; 36: 1034-1040

[13] Kantety RV, La Rota M, Matthews DE, Sorrells ME. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol Biol. 2002; 48:501-510.

[14] Kashkush K, Jinggui F, Tomer E, Hillel J, Lavi U. Cultivar identification and genetic map of mango (*Mangifera indica*). Euphytica. 2001: 122:129–136.

[15] Kumpatla SP, Buyyarapu R., Abdurakhmonov IY, Mammadov JA. (2012). Genomics-Assisted Plant Breeding in the 21st Century: Technological Advances and Progress Plant Breeding. I. Y. Abdurakhmonov. INTECH. 2012; 352.

[16] Nagaraj SH, Gasser RB. A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief Bionformatics. 2007; 8(1): 6-21.

[17] Nielsen R., Paul JS, Albrechtsen A, Yun SS. Genotype and SNP calling from next generation sequencing data. Nature Reviews Genetics. 2011; 12: 443-451.

[18] Novelli VM., Takita MA, Machado MA. Identification and analysis of single nucleotide polymorphisms (SNPs) in citrus. Euphytica. 2004; 138:227–237.

[19] Rafalski A. Applications of single nucleotide polymorphisms in crop genetics. Current Opinion in Plant Biology. 2002; 5(2): 94-100

[20] Useche FJ, Gao G, Hanafey M, Rafalski A. High throughput identification, database storage and analysis of SNPs in EST Sequences. Genome Informatics. 2001; 12:194-203.