In Silico Identification and Characterization of Genic-SSRs in Jackfruit (Artocarpus heterophyllus)

Devendra Kumar Singh, M. Anwar Mallick and Binay K. Singh

Research Journal of Agricultural Sciences
An International Journal

P- ISSN: 0976-1675
E- ISSN: 2249-4538

Volume: 13
Issue: 06

Res. Jr. of Agril. Sci. (2022) 13: 1831–1834

CARAS

Res. Jr. of Agril. Sci. (Nov-Dec 2022)
13(6): 1831–1834

www.rjas.org

ISSN: 0976-1675 (P)
ISSN: 2249-4538 (E)

Full Length Research Article

In Silico Identification and Characterization of Genic-SSRs in Jackfruit (*Artocarpus heterophyllus*)

Devendra Kumar Singh*[1,2], M. Anwar Mallick[2] and Binay K. Singh[3]

Received: 17 Sep 2022 | Revised accepted: 17 Nov 2022 | Published online: 12 Dec 2022

ABSTRACT

Jackfruit (*Artocarpus heterophyllus*), an economically and nutritionally important fruit, contains many healthy ingredients required to achieve nutritional security for the rapidly growing population. A draft genome sequence of *A. heterophyllus* is available in the public domain with minimal characterization. In the present study, we identified 19,934 genic-SSRs using the genomic location of the genes known in the reference genome. The detailed analysis of genic-SRRs showed that out of 19,934 genic-SSRs, 3,510 and 16,424 were located in the coding (CDS) and intronic part of the genes, respectively. It was found that trinucleotide repeat was the most predominant class (89%) in CDS-derived SSRs, while dinucleotide repeats (56.4%) were dominant in intronic SSRs. 491 and 321 distinct repeat motifs were present in intronic- and CDS-derived SSRs. GAA/TTC was the most abundant trinucleotide repeat motif with a frequency of 8.72% in CDS-derived SSRs, whereas AT/AT was the most abundant dinucleotide repeat motif with a frequency of 18.66 % in intronic SSRs.

Key words: Genic SSRs, Jackfruit, Intron, CDS, Genome sequence

Artocarpus heterophyllus (2n=4X=56), commonly known as Jackfruit or Jack, is a widespread and economically significant tree belonging to the Moraceae family [1]. It is native to the rainforests of the Western Ghats of India. Besides India, Jackfruit is also found in tropical and sub-tropical countries like Sri Lanka, Bangladesh, Burma, Philippines, Indonesia, Thailand, Malaysia, and Brazil [2-3]. *Artocarpus heterophyllus* is a multifaceted tree because every part serves a different purpose [4]. However, the scientific community has paid little attention to this multi-utility tree species. There have been few publications on the development of molecular markers in *A. heterophyllus* [5]. As a result, developing a diverse range of molecular markers for *A. heterophyllus* is essential for genetic improvement and long-term conservation. DNA-based molecular markers are an important and adaptable tool in plant breeding, taxonomy, physiology, and genetic engineering [6-7]. SSRs-based molecular markers are widely used to characterize a large set of germplasm at a minimal cost within a short time [8]. The SSR markers are distributed across the genome and show co-dominant inheritance. SSR markers are often used to study population genetics, phylogenetic relationships, and genetic diversity [9-11]. These markers have also been widely used for constructing linkage maps and marker-assisted selection [12-13].

This study aims to identify and characterize genic-SSR markers in *Artocarpus heterophyllus*. We identified genome-wide genic-SSRs using gene location information available in the reference genome. Further, they were categorized into CDS and intronic SSRs based on the position of repeat motifs in the gene. The SSRs identified in the present study provide potentially critical molecular markers for marker-based studies in *Artocarpus heterophyllus*.

MATERIALS AND METHODS

Identification of genic SSRs and primer designing

Reference genome sequence (FASTA) and gene annotation (GFF) of *A. heterophyllus* was downloaded from 'Online Resource for Community Annotation of Eukaryotes' (https://bioinformatics.psb.ugent.be/orcae/aocc/overview/Arthe). Kraitv1.3.3 tool [14] was employed to detect genome-wide SSRs in *Artocarpus heterophyllus*. Only the SSR loci with 2-6 nucleotides simple sequences repeated at least four times were selected. Mononucleotide repeats were filtered out in the study. We selected only gene-based SSRs for further studies using the gene annotation information. BatchPrimer3 v1.0 software (an integral part of Krait v1.3.3) was used to design the primers for the mined genic-SSRs. The following criteria were applied for designing the primers: primer length = 18–27 bases (optimal of 20 bases), GC content= 30–80% with primer GC clamp 2, annealing temperature =58–65 °C (optimal 60 °C), and the

* Devendra Kumar Singh
✉ devndri16@gmail.com

[1, 3] ICAR – Indian Institute of Agricultural Biotechnology, Ranchi - 834 003, Jharkhand, India

[2] University Department of Biotechnology, Vinoba Bhave University, Hazaribagh - 825 001, Jharkhand, India

product size = 100–300 bp. The detailed flow chart for identifying gene-based SSRs in *Artocarpus heterophyllus* showed in (Fig 1).

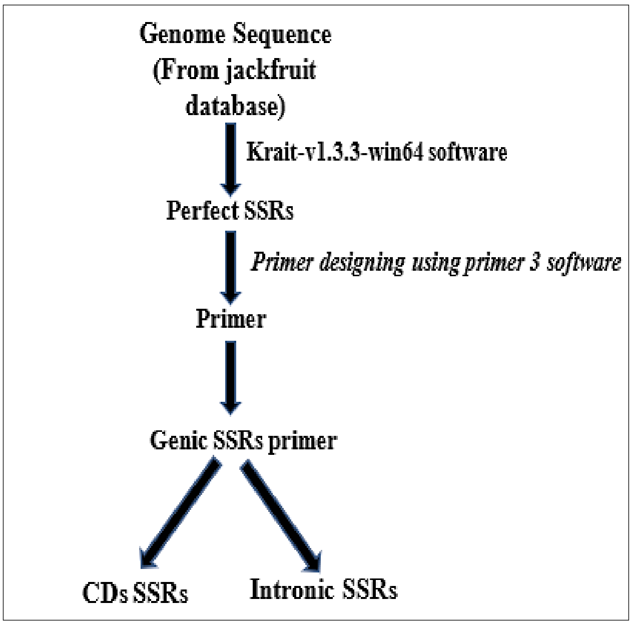
Genome Sequence (From jackfruit database)
Krait-v1.3.3-win64 software
Perfect SSRs
Primer designing using primer 3 software
Primer
Genic SSRs primer
CDs SSRs
Intronic SSRs

Fig 1 Pipeline used to identify the perfect CDS and intron-derived genic SSRs

Characterization of genic- SSRs

The identified genic-SSRs were further categorized into CDS-based and intronic SSRs depending on their position in the gene. Both the SSR types were characterized for repeat units, the number of reiterations, etc.

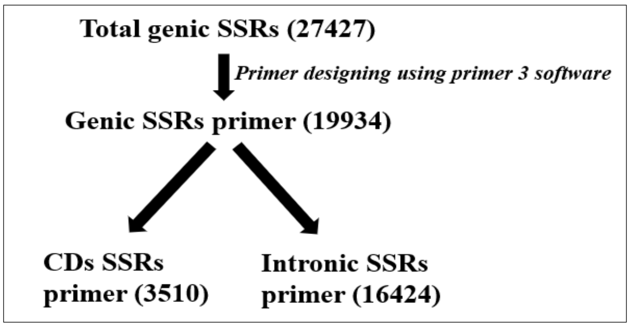
Total genic SSRs (27427)
Primer designing using primer 3 software
Genic SSRs primer (19934)
CDs SSRs primer (3510)
Intronic SSRs primer (16424)

Fig 2 Flow diagram indicating the summary of the numbers of SSRs identified and primers designed

RESULTS AND DISCUSSION

Mining of genic-SSRs and primer designing

A total of 27,427 gene-based SSRs were identified in the *A. heterophyllus* genome using Krait v1.3.3. However, PCR primers could be designed only for 19,934 SSR loci. The flanking sequences for the remaining 7,493 SSR loci were too short, or the nature of the sequence did not fulfil the criteria for primer design using Krait v1.3.3 software. Of the 19,934 SSR loci for which primers could be designed, 3,510 (17.60%) were located in the CDS, whereas 16,424 (82.40%) SSRs were found in the intronic part of the gene (Fig 2).

Analysis of repeat motif distribution revealed that trinucleotide repeat (89%) was the predominant class, followed by hexanucleotide (7.49%) in CDS. In comparison, dinucleotide repeats (56.4%) were the predominant class in intronic SSRs, followed by trinucleotide (20.03%) and tetranucleotide (15.96%) (Fig 3).

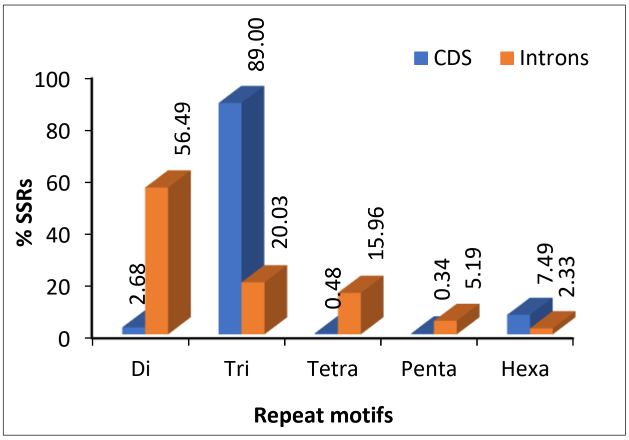

Fig 3 Distribution frequency of repeat motifs in the CDS and intronic SSRs

These are evident in many species wherein the exons (unlike other genomic regions) contain rare dinucleotide and tetranucleotide SSRs but have many more triplet and hexanucleotide SSRs [15-18]. More than 92% of the predicted SSRs within coding sequences have repeat-unit sizes that are a multiple of three [19]. All human chromosomes, except the Y chromosome, have a frequency of triplet repeats roughly two times higher in exonic than in intronic and intergenic regions [20]. Similar results are also reported in plants, animals, and microbes.

Table 1 Distribution characteristics of Intronic SSR motifs in this study

SSR motif	Number	Percentage (%)	No. of reiterations	Average length (bp)
Dinucleotide	9,278	56.49	12	19.30
Trinucleotide	3,290	20.03	57	19.07
Tetranucleotide	2,621	15.96	123	18.12
Pentanucleotide	853	5.19	139	21.54
Hexanucleotide	382	2.33	160	25.79
Total	16424	100	491	20.76

Table 2 Distribution characteristics of CDS SSR motifs in this study

SSR motif	Number	Percentage (%)	No. of reiterations	Average length (bp)
Dinucleotide	94	2.68	10	17.81
Trinucleotide	3124	89.00	60	17.47
Tetranucleotide	17	0.48	14	16.71
Pentanucleotide	12	0.34	10	20.83
Hexanucleotide	263	7.49	227	25.80
Total	3510	100	321	19.72

Frequency distribution of SSR repeat motifs

We calculated the frequency distribution of SSR repeat motifs for both intronic and CDS-derived SSRs. It was found that a total of 491 and 321 distinct types of repeat motifs were present in intronic- (Table 1) and CDS-derived SSRs (Table 2).

GAA/TTC was the most abundant trinucleotide repeat motif, with a frequency of 8.72% in CDS-derived SSRs (Table 3), whereas AT/AT was the most abundant dinucleotide repeat motif, with a frequency of 18.66% in intronic SSRs (Table 4). Similar results are reported in the EST-SSRs developed in flax (*Linum usitatissimum* L.), wherein GAA/TTC was the most abundant motif with a frequency of 10.2% [21]. Similarly, cassava intronic SSRs have AT/AT, the most abundant motif, with a frequency of 22% [22].

Table 3 frequency distribution of the ten most abundant repeat motifs in the CDS-derived SSRs in *Artocarpus heterophyllus*

Repeat motifs	No. of reiteration of the motif															Total	Percentage
	4	5	6	7	8	9	10	11	12	13	14	15	16	21	26		
GAA/TTC	-	165	82	33	9	6	4	2	2	2	1	-	-	-	-	306	8.72
AGA/TCT	-	142	60	28	18	8	4	3	3	1	-	-	-	-	-	267	7.61
AAG/CTT	-	124	61	28	9	2	3	2	1	1	-	-	1	-	-	232	6.61
GGA/TCC	-	88	48	17	13	3	6	2	-	-	-	-	-	-	-	177	5.04
CAA/TTG	-	89	33	18	9	2	4	2	-	1	-	-	-	-	-	158	4.50
TCA/TGA	-	91	41	10	6	4	1	2	-	-	-	-	-	-	-	155	4.42
CTC/GAG	-	77	36	13	11	7	-	-	-	-	-	-	-	-	-	144	4.10
ATC/GAT	-	77	30	14	3	2	1	-	1	-	-	-	-	-	-	128	3.65
CCG/CGG	-	85	25	8	7	-	-	-	-	-	-	-	-	-	-	125	3.56
CCA/TGG	-	68	24	15	6	3	2	2	1	-	-	-	-	-	-	121	3.45
Other motifs	225	831	301	170	82	32	19	14	10	6	3	1	1	1	1	1697	48.35
Total	225	1837	741	354	173	69	44	29	18	11	4	1	2	1	1	3510	

Table 4 Frequency distribution of the ten most abundant repeat motifs in the intronic SSRs in *A. heterophyllus*

Repeat motifs	No. of reiteration of the motif																												Total	%
	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	29	35	40	66		
AT/AT	-	-	-	601	635	509	390	267	180	145	101	84	44	34	30	14	14	3	3	-	7	2	-	-	-	-	1	1	3065	18.66
TA/TA	-	-	-	354	355	296	214	169	110	68	40	35	16	17	8	8	1	1	1	2	-	-	-	1	-	-	-	-	1696	10.33
AG / CT	-	-	-	440	287	223	161	147	114	67	53	39	31	16	8	8	9	3	3	1	2	-	2	3	1	-	-	-	1618	9.85
GA / TC	-	-	-	296	241	186	132	103	88	79	41	37	19	14	9	12	8	4	1	1	1	1	-	1	-	1	-	-	1275	7.76
AAT / ATT	-	322	227	132	77	44	20	23	6	7	3	1	1	1	-	-	-	1	-	-	-	-	-	-	-	-	-	-	865	5.27
AC / GT	-	-	-	242	176	146	92	71	38	20	15	8	9	3	4	2	2	2	1	1	-	-	-	-	-	1	-	-	833	5.07
CA / TG	-	-	-	242	187	141	81	53	37	16	15	5	4	2	2	-	-	-	-	-	-	-	-	-	-	-	-	-	785	4.78
AAAT / ATTT	419	158	27	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	609	3.71
TAA / TTA	-	228	133	80	36	33	19	9	6	2	3	2	-	-	1	-	-	1	-	-	-	-	-	-	-	-	-	-	553	3.37
TAAA / TTTA	262	86	12	3	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	364	2.22
Other motifs	1961	1446	684	308	164	84	42	31	27	7	2	2	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	4761	28.99
Total	2642	2240	1083	2703	2159	1662	1151	873	606	411	273	213	125	88	62	45	34	15	9	5	10	3	2	5	1	2	1	1	16424	

SSRs in coding regions are often biased toward a certain nucleotide composition. The A/T repeats occur more frequently than G/C repeats [23]. In *Arabidopsis thaliana* and cereals, exons, and ESTs indicate a higher frequency for GA/CT repeats than for AT repeats [24-25]. Plant genomes contain fewer AC/GT repetitions than animal genomes. This pattern might be explained by the higher frequency of some amino acids in plants compared to animals. AGC is the most prevalent trinucleotide repeat in the animal kingdom. The most common triplet motif in dicot plants is the AAG (28.3% - 42.1%). The most frequent trinucleotide repeats in cereal species, however, is CCG, which in the case of wheat is 32%, sorghum 49%, and sorghum 39.3% [26-27]. Monocot genomes are distinguished by their high frequency of CCG repeats, which may be related to their higher GC content [25]. In monocot species, the AAT motifs are the least frequent (1%). This may be accounted for because TAA-based variations encode stop codons that immediately impact eukaryotic protein production [28]. Depending on the type of encoded amino acid, distinct codon repeats frequencies also vary greatly. Different kinds of proteins contain higher concentrations of particular amino acid repetitions. Ser repeats are highly related to membrane transporter proteins, while acidic and polar amino acid repeats are significantly associated with transcription factors and protein kinases.

CONCLUSION

Using the chromosomal locations of the known genes in the reference genome of the jackfruit, we discovered 19,934 genic-SSRs. 3,510 and 16,424 of the 19,934 genic-SSRs were found in the intronic and coding (CDS) regions, respectively, of the genes. In CDS-derived SSRs, trinucleotide repeats dominated (89%), while dinucleotide repeats (56.4%) dominated intronic SSRs. SSRs produced from intronic and CDS regions had 491 and 321 unique repeat motifs. In CDS-derived SSRs, GAA/TTC was the most prevalent trinucleotide repeat, with a frequency of 8.72%. In contrast, AT/AT was the most prevalent dinucleotide repeat motif in intronic SSRs, with a frequency of 18.66%. The study's genic-SSRs add to the already-developed genomic resources and would be very helpful for the marker-based studies for the species.

Conflict of interest

The authors of this manuscript have no competing interests.

Acknowledgement

We sincerely acknowledge the Director, ICAR - Indian Institute of Agricultural Biotechnology, Ranchi - 834 003, Jharkhand, India, for providing financial support and the facilities to carry out this research work.

LITERATURE CITED

1. Darlington CD, Wylie A P. 1956. *Chromosome Atlas of Flowering Plants*. Second Edition. The Macmillan Co. 60 Fifth Ave. New York 11: 519.
2. Wangchu L, Singh D, Mitra SK. 2013. Studies on the diversity and selection of superior types in Jackfruit (*Artocarpus heterophyllus* Lam.). *Genetic Resources and Crop Evolution* 60(5): 1749-1762.
3. Vazhacharickal P, Mathew J, Kuriakose A, Abraham B, Mathew R, Albin A, Thomson D, Thomas R, Varghese N, Jose S. 2015. Chemistry and medicinal properties of jackfruit (*A. heterophyllus*): a review on the current status of knowledge. *International Journal of Innovative Research and Review* 3(2): 83-95.
4. Burkill IH, Birtwistle W. 1966. A Dictionary of the Economic Products of the Malaya Peninsula. [2D ed.] ed. Kuala Lumpur Malaysia: *Published on behalf of the governments of Malaysia and Singapore by the Ministry of Agriculture and cooperatives.* pp 2444.
5. Kavya K, Shyamalamma S, Gayatri S. 2019. Morphological and molecular genetic diversity analysis using SSR markers in Jackfruit (Artocarpus heterophyllus Lam.) genotypes for pulp colour. *Indian Jr. Agric. Research* 53(1): 8-16.
6. Bernardo R. 2003. Parental selection, number of breeding populations, and size of each population in inbred development. *Theor. Appl. Genetics* 107: 1252-1256.
7. Kesawat M, Das B. 2009. Review on Molecular marker: Its application in crop improvement. *Journal of Crop Science and Biotechnology* 12: 168-178.
8. Zhu H, Song P, Koo DH. 2016. Genome wide characterization of simple sequence repeats in watermelon genome and their application in comparative mapping and genetic diversity analysis. *BMC Genomics* 17: 557.
9. Li G, Hongjin W, Tao L, Jianbo Li, Shixiao La, Ennian Y, Zujun Y. 2016. New molecular markers and cytogenetic probes enable chromosome identification of wheat-Thinopyrum intermedium introgression lines for improving protein and gluten contents. *Planta* 244(4): 865-876.
10. Jia Q, Zhu J, Wang J. 2016. Genetic mapping and molecular marker development for the gene Pre2 controlling purple grains in barley. *Euphytica* 208(2): 215-223.
11. Li GY, McVetty PBE, Quiros CF. 2013. SRAP molecular marker technology in plant science. *In*: (Eds) S. B. Andersen. *Plant Breeding from Laboratories to Fields*. pp 23-43.
12. Fraser LG, Tsang GK, Datson PM. 2009. A gene-rich linkage map in the dioecious species *Actinidia chinensis* (kiwifruit) reveals putative X/Y sex-determining chromosomes. *BMC Genomics* 10: 102.
13. Rajaram V, Nepolean T, Senthilvel S. 2013. Pearl millet [*Pennisetum glaucum* (L.) R. Br.] consensus linkage map constructed using four RIL mapping populations and newly developed EST-SSRs. *BMC Genomics* 14: 159.
14. 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: 681-683.
15. Field D, Wills C. 1996. Long, polymorphic microsatellites in simple organisms. *Proc. R. Soc. London Ser. B* 263: 209-251.
16. Edwards YJ, Elgar G, Clark MS, Bishop MJ. 1998. The identification and characterization of microsatellites in the compact genome of the Japanese pufferfish, Fugu rubripes: perspectives in functional and comparative genomic analyses. *Jr. Mol. Biology* 278: 843-854.
17. Metzgar D, Bytof J, Wills C. 2000. Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Research* 10: 72-80.
18. Young ET, Sloan JS, Van Riper K. 2000. Trinucleotide repeats are clustered in regulatory genes in *Saccharomyces cerevisiae*. *Genetics* 154: 1053-1068.
19. Wren JD, Forgacs E, Fondon JW 3rd, Pertsemlidis A, Cheng SY, Gallardo T, Williams RS, Shohet RV, Minna JD, Garner HR. 2000. Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. *Am. Jr. Hum. Genetics* 67: 345-356.
20. Subramanian S, Mishra RK, Singh L. 2003. Genomewide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biology* 4: R13.
21. Cloutier S, Niu Z, Dalta R, Duguid S. 2009. Development and analysis of EST-SSRs for flax (*Linum usitatissimum* L.). *Theor. Appl. Genetics* 119(1): 53-63. doi: 10.1007/s00122-009-1016-3.
22. Vásquez A, López C. 2014. In silico genome comparison and distribution analysis of simple sequences repeats in cassava. *Int. Jr. Genomics* 461-471. doi: 10.1155/2014/471461
23. Olivero M, Ruggiero T, Coltella N, Maffe' A, Calogero R, Medico E, Di Renzo MF. 2003. Amplification of repeat-containing transcribed sequences (ARTS): a transcriptome fingerprinting strategy to detect functionally relevant microsatellite mutations in cancer. *Nucleic Acids Research* 31.
24. Kantety RV, La Rota M, Matthews DE, Sorrells ME. 2002. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Mol. Biology* 48: 501-510.
25. Morgante M, Hanafey M, Powell W. 2002. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genetics* 30:194-200.
26. Varshney RK, Thiel T, Stein N, Langridge P, Graner A. 2002. In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cell. Mol. Biol. Letters* 7: 537-546.
27. Thiel T, Michalek W, Varshney RK, Graner A. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genetics* 106: 411-422.
28. Cordeiro GM, Casu R, McIntyre CL, Manners JM, Henry RJ. 2001. Microsatellite markers from sugarcane (*Saccharum spp*) ESTs across transferable to erianthus and sorghum. *Plant Science* 160:1115-1123.