Bioinformation Volume 5

open access

www.bioinformation.net Issue 3

Hypothesis

Mining and characterization of EST derived microsatellites in *Curcuma longa* L.

Raj Kumar Joshi, Ananya Kuanar, Sujata Mohanty, Enketeswara Subudhi, Sanghamitra Nayak*

Centre of Biotechnology, School of Pharmaceutical Sciences, Siksha O Anusandhan University, Bhubaneswar-751003, India; Sanghamitra Nayak - Email: sanghamitran@yahoo.com; Phone : 09437061976 ; *Corresponding author

Received August 20, 2010; accepted August 26, 2010; published September 20, 2010

Abstract:
Turmeric (*Curcuma longa L.*) (Family: Zingiberaceae) is a perennial rhizomatous herbaceous plant often used as a spice since time immemorial. Turmeric plants are also widely known for its medicinal applications. Recently EST-derived SSRs (Simple sequence repeats) are a free by-product of the currently expanding EST (Expressed Sequence Tag) databases. SSRs have been widely applied as molecular markers in genetic studies. Development of high throughput method for detection of SSRs has given a new dimension in their use as molecular markers. A software tool SciRoKo was used to mine class I SSR in Curcuma EST database comprising 12953 sequences. A total of 568 non-redundant SSR loci were detected with an average of one SSR per 14.73 Kb of EST. Furthermore, trinucleotide was found to be the most abundant repeat type among 1–6-nucleotide repeat types. It accounted for 41.19% of the total, followed by the mononucleotide (20.07%) and hexanucleotide repeats (15.14%). Among all the repeat motifs, (A/T)n accounted for the highest proportion followed by (AGG)n. These detected SSRs can be greatly used for designing primers that can be used as markers for constructing saturated genetic maps and conducting comparative genomic studies in different Curcuma species.

Keywords: *Curcuma longa*, Expresses sequence tags, short sequence repeats, SciRoKo.

Background:
The genus Curcuma of the family Zingiberaceae constitutes 80 species all over Asia, South East Asia and Africa **[1]**. Turmeric, also known as the "golden spice" is one of the most important herbs in the tropical and sub-tropical countries. Turmeric rhizome is valued world over and has been in use from ancient time as a spice, food preservative, coloring agent, and in the traditional systems of medicine **[2]**. Its medicinal uses are indeed diverse, ranging from cosmetic face cream to the prevention of Alzheimer's disease. Turmeric is also qualified as the queen of natural Cox-2 inhibitors **[3]**. India is the world's largest producer, and exporter of turmeric followed by China, Indonesia, Bangladesh and Thailand **[4]**. The International Trade Centre, Geneva, has estimated an annual growth rate of 10% in the world demand for turmeric. Conventional crop improvement methods are not suitable in turmeric because it is not only completely sterile but also propagate exclusively by vegetative means. Characterization of *Curcuma longa* using molecular markers is very limited excepting a few sporadic reports on isozyme studies and genetic stability studies using RAPD **[5, 6].** Moreover, it is a well-known fact that the genotypic diversity of exclusively asexually reproducing plants like turmeric will be lost in the long course of evolution. Hence, the development of reliable and reproducible molecular markers in turmeric is highly essential to assess the genetic diversity for germplasm conservation and crop improvement

Microsatellites, or simple sequence repeats (SSRs), are stretches of DNA consisting of tandemly repeated short units of 1–6 base pairs in length. Compared with other molecular markers, simple sequence repeats (SSRs) are more advantageous because of their simplicity, high information, and co dominant nature and because they can be rapidly screened and analyzed by polymerase chain reaction (PCR) and gel electrophoresis. In addition, SSR loci are present not only in the non-coding regions of genes but are also widely distributed in the coding regions. Microsatellites are categorized into two groups- class I hypervariable markers with ≥20 repeats and class II potentially variable markers with ≤20 repeats. The standard method for development of genomic SSRs is highly time consuming and labor-intensive **[7]**. Recent advances in Curcuma genomic technologies have generated a large number of expressed sequence tags (ESTs) that has been made available in public database, thereby offering an opportunity to develop EST derived SSR markers by data mining. ESTs are short and single pass sequences read from mRNA (cDNA) **[8]** representing a snapshot of genes expressed in a given tissue and or at a given developmental stage. As of July 2010, GenBank had released 12593 EST sequences from Curcuma longa. In this context, the use of EST or cDNA-based SSRs has been reported for several species including grape **[9]**, sugarcane **[10]**, durum wheat **[11]** and rye **[12]**.

Keeping in view the above, the objectives of the research described in this paper were to assess the potential of existing public databases for the discovery of simple sequence repeats. We have mined updated EST tissue libraries of *Curcuma longa for* this analysis to find the SSR polymorphisms. SSR detecting software SciRoKo was used to identify the SSR polymorphisms. There are other SSR detecting softwares such as MISA **[7]**, SSRFinder **[13]**, SSRIT **[14]**, TRF **[15]**, TROLL **[16]**, Sputnik (http://espressosoftware.com/pages/sputnik. jsp), Modified Sputnik I **[17]** and Modified Sputnik II **[18]** but SciRoKo (SSR Classification and Investigation by Robert Kofler) **[19]** is the only software with user-friendly interface with a statistical analysis of genomic microsatellites and interpret results as html files.

Methodology:
EST database of NCBI contains 12953 *Curcuma longa* express sequence tag data. We have mined 12593 EST sequences consisting of two tissue libraries of rhizomes 6870 (DY395309-DY388440) and leaves 5723 (DY388439-DY382717). The EST sequences were screened against the UniVec database from NCBI (ftp://ftp.ncbi.nih.gov/pub/ UniVec/) for detecting vector and adapter sequences by using the program Cross_Match [Li et al 2006]; the following parameters were used: minmatch ≥13 and minscore ≥20. Furthermore, polyA/T tails and X characters were removed using the EST_trimmer.pl script (http://pgrc.ipk-gatersleben. de/misa/download/est_trimmer.pl) until no stretch of (A/T)5 or (X)1 was present in a window of 100bp at the 5′ or 3′ end, respectively. CAP3 program was used to assemble the EST sequence into contigs for creating a non-redundant dataset. The SSR detection tool SciRoKo version 1.0 **[19]** was used to detect EST-SSR loci. SciRoKo required inputs in fasta format.

Bioinformation Volume 5

open access

www.bioinformation.net Issue 3 **Hypothesis**

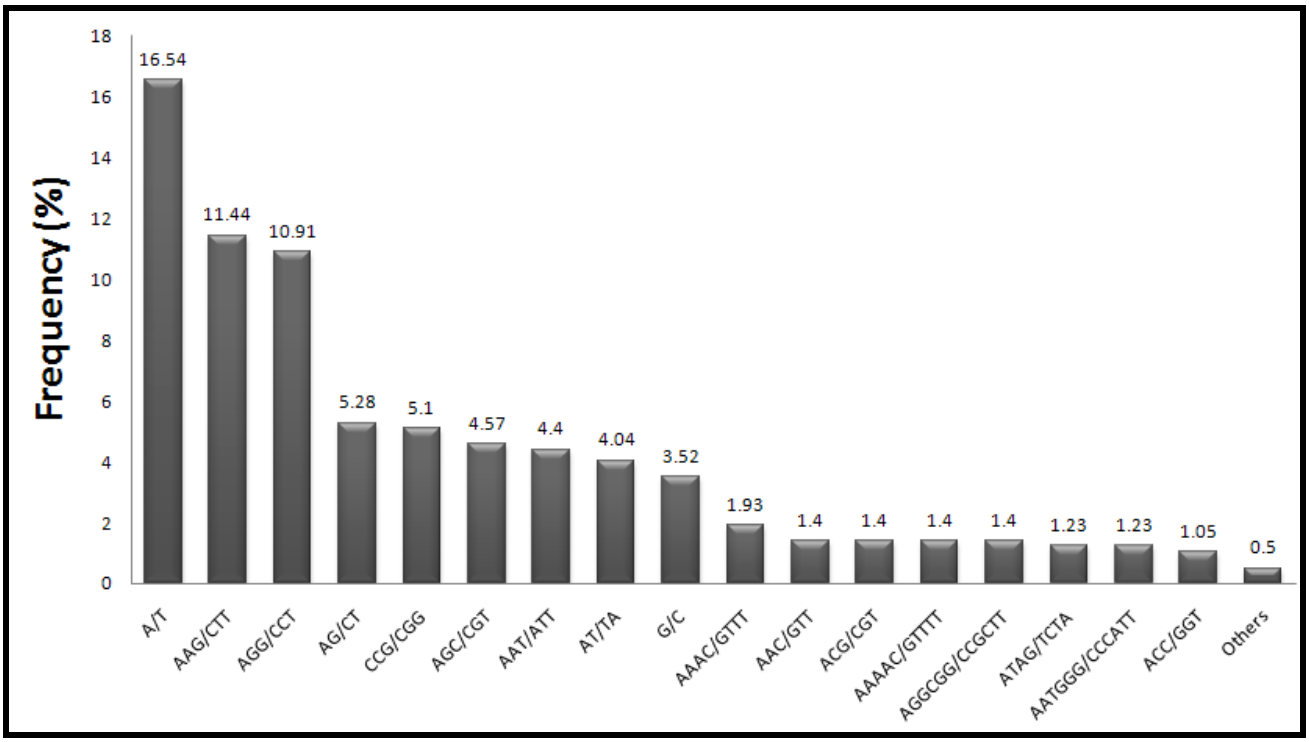

Figure 1: Distributions of EST-SSRs based on the motifs.

Discussion :

Large-scale sequencing of Expressed Sequence Tags and complete genomes offers information of use to plant breeding programs. With the completion of the first crop genome sequencing projects **[20]** the potential for plant breeding to be impacted by new technology has never been greater. A total of 12953 redundant EST sequences were retrieved from NCBI database representing about 8.4 Mb of *Curcuma longa* genome. During pre-processing, 38366Bp of empty vectors, low-quality sequences and Poly A/T tails were removed successfully. After sequence redundant analysis, 7139 unique sequences with combined length of 5.11 Mb were obtained and were used for mining of hyper variable class I microsatellites. Using the SciRoKo SSR mining program, while searching for SSRs with 1-6 nucleotide repeat motifs, 568 hypervariable SSR loci were observed **(Table 3 see supplementary material).** The frequency of SSR loci in turmeric EST was found to be one SSR in every 14.73 kb of EST sequence **(Table 1 see supplementary material)** that is higher as compared to earlier retrieved data of one SSR per 17.96 Kb **[21]**. Cardle et al **[22]** estimated the average distances between SSRs in sets of non-redundant ESTs in poplar (1/14.0 kb), cotton (1/20.0 kb), *Arabidopsis* (1/13.8 kb), maize (1/8.1 kb), rice (1/3.4 kb), tomato (1/11.1 kb) and soybean (1/7.4 kb). This clearly suggests that, with the increase in the transcript data of plants, SSR estimation in ESTs will become more precise and reliable.

The mined EST-SSRs were classified according to their structure into the simple motif type, with a single motif; and compound type, with more than two motifs. Among the 568 EST-SSRs, most (98.92%) consisted of simple repeats with no interruptions in the motif; whereas only six loci (1.05 %) were of the compound type **(table 2 see supplementary material)**. Among the 1−6 repeat types, the most abundant repeat type was the trinucleotide repeat type, which accounted for 41.19% of the total, followed by the mononucleotide (20.07%) and hexanucleotide types (15.14%). The dinucleotide, tetranucleotide and pentanucleotide types accounted for only 9.68%, 6.16% and 6.69% respectively. Many studies have suggested that the trinucleotide repeat is the main EST-SSR repeat type in most plants, followed by the dinucleotide and tetranucleotide repeat types **[22, 23]**. However, the most abundant motif in the trinucleotide repeat type differed among plants **[13]**. Kantety et al **[24]** showed that the (CCG)n repeat motif accounted for 32% and 49% of all repeat motifs in wheat and sorghum, respectively. Gupta et al. **[25]** found that the (AAG)n repeat was the most abundant motif in the trinucleotide repeat type. In a similar study Lu et al 2010, have found (AAG)n to be the most abundant repeat motif in *Gossypium barbadense*. Siju et al **[21]** also found (AAG)n to be the most abundant in turmeric accounting to 8.2%.

The dominance of trimeric SSRs observed in the present study could be attributed to the fact that the suppression of non-trimeric SSRs in the coding regions leads to frame shift mutations **[26]**. Moreover, we also found that the frequency of mononucleotide (20.07%) and hexanucleotide (15.14%) repeats was more as compared to other repeat motifs. This suggests that the functions of EST-SSRs derived from Curcuma longa may be different form other members of the Zingiberaceae family. In all the repeat motifs, most of the SSR repeat motifs derived from the ESTs were A/T (16.54%) followed by AAG/CTT (11.44%), AGG/CTT (10.91%) and AG/CT (5.28%) **(Figure. 1).** Rest of the repeat motifs accounted less than 5% contribution to the total SSR motifs. In the 1−6 repeat types, the most frequent repeat motifs were A/T, AG/CT, AAG/CTT, AAAC/GTTT, AAAAC/GTTTT, and AGGCGG/CCGCCT, which accounted for 78.9%, 54.54% 29.01%, 31.42%, 15.68% and 8.98% of all types, respectively. This frequency analysis of repeat motifs can be used as a potential source for designing repeat probes for effective targeting and isolation of microsatellite repeats from turmeric. Moreover, these probes can be used for designing informative primers that can be used for genetic diversity analysis and related studies.

Conclusion:

In total, we identified 568 non-redundant hypervariable microsatellites from EST data source of *Curcuma longa* using SSR identification tool SciRoKo. Development of SSR markers from EST-databases saves both cost and time, once a sufficient amount of EST sequences is available. These non-redundant SSR resources will not only be applied in studies of genetic variation and linkage mapping but also provide the foundation for an in-depth analysis of the characteristics of distribution of genes on chromosomes and for comparative genomic studies on different *Curcuma* species.

Acknowledgement:

We are grateful to Prof. Manoj Ranjan Nayak, President, Siksha O Anusandhan University for his able guidance and support.

References:

[1] KC Velayudhan et al., NBPGR (1999).

[2] PN Ravindran *et al.* Turmeric-the genus *Curcuma* CRC Press (2007).

ISSN 0973-2063 (online) 0973-8894 (print)

Bioinformation 5(3): 128-131 (2010)

[3] JA Duke. CRC handbook of medicinal plants (2003).
[4] Selvan *et al.* Indian Spices- production and utilization (2002).
[5] A Shamina *et al. Jour Hort Sci Biotechnol* **73**:479 (1998).
[6] S Shyamkumar, PhD thesis Calicut University (2008).
[7] T Thiel *et al. Theor Appl Genet*, **106**:411 (2003).
[8] MD Adams *et al. Science* **252**:1651 (1991).
[9] KD Scott *et al. Theor Appl Genet* **100**:723 (2000).
[10] G Cordeiro *et al. Plant Sci* **160**:1115 (2001).
[11] I Eujayl *et al. Theor Appl Genet*, **104**:399 (2002).
[12] B Hackauf *et al. Plant Breed*, **121**:17 (2002).
[13] L Gao *et al. Mol Breed* **12**:235 (2003).
[14] S Temnykh *et al. Genome Res* **11**:1441 (2001).
[15] G Benson *et al. Nucleic. Acid Res*, **27**:573 (1999).
[16] AT Castelo *et al. Bioinformatics* **18**:634 (2002).
[17] M Morgante *et al. Nat. Genet,* **30**:194 (2002).
[18] M La Rota *et al. BMC Genomics*, **6:**23 (2005).
[19] R Kofler *et al. Bioinformatics*, 23:1683 (2007).
[20] JYu *et al. Science* **296**:79 (2002) [PMID: 11935017]
[21] S Siju *et al. Mol Biotechnol* 44:140 (2010).
[22] L Cardle *et al. Genetics* **156**:847 (2000).
[23] R Kota *et al. Heriditas* **135**:145 (2001)
[24] RV Kantety *et al. Plant Mol Biol* **48**:501 (2003).
[25] PK Gupta *et al. Current Sci* **70**:45 (1996).
[26] D Metzger *et al. Genome Res* **10**:72 (2000).

Edited by P. Kangueane
Citation: Joshi *et al.* Bioinformation 5(3): 128-131 (2010)

Bioinformation Volume 5

open access

www.bioinformation.net Issue 3 **Hypothesis**

Supplementary materials:

Table 1: Summary of EST-derived microsatellies from the EST database of *Curcuma longa* L.

Parameters	Values
Total number of ESTs	12953
Total sequence analyzed	8366842bp
EST after vector and Poly A/T removal	8328476bp
Total gene sequences after assembly	7139 (5117624bp)
Total number of contigs	4035 (3125829bp)
Total number of singletons	3104 (1991795bp)
Total number of SSR loci located	568
Frequency of SSR loci in turmeric EST	1 per 14.730 kb

Table 2: Distribution of SSR motifs in *Curcuma longa*

Repeat motif type	Number	Frequency (%)	The most abundant motif	Frequency within their own repeat (%)
Mononucleotide	114	20.07	A/T	78.9
Dinucleotide	55	9.68	AG/CT	54.54
Trinucleotide	234	41.19	AAG/CTT	29.01
Tetranucleotide	35	6.16	AAAC/GTTT	31.42
Pentanucleotide	38	6.69	AAAAC/GTTTT	15.68
Hexanucleotide	86	15.14	AGGCGG/CCGCCT	8.98
Compound	6	1.05	/	/
Total	568	100	/	/

Table 3: Total number of detected SSR loci.

Motif	Number of loci	Motif	Number of loci
A/T	94	AAAATG/CATTTT	1
G/C	20	AAATCT/AGATTT	1
AT/TA	23	AACGCC/GGCGTT	2
AC/GT	2	AAGAGG/CCTCTT	4
AG/CT	30	AATGGG/CCCATT	7
AAC/GTT	8	ACCATC/GATGGT	1
AAG/CTT	65	ACCCCC/GGGGGT	2
AAT/ATT	25	ACCTCC/GGAGGT	2
ACC/GGT	6	AGATCG/CGATCT	3
ACG/CGT	8	AGCAGG/CCTGCT	2
AGC/CGT	26	AGGCGG/CCGCCT	8
AGG/CCT	62	AGGGCG/CGCCCT	1
ATC/GAT	5	ATCGGC/GCCGAT	1
CCG/CGG	29	AAAAAT/ATTTTT	3
AAAC/GTTT	11	AAAAAG/CTTTTT	3
AAAG/CTTT	4	AAAAAC/GTTTTT	3
AAAT/ATTT	5	ACCGCC/GGCGGT	3
AAGG/CCTT	1	AATAGG/CCTATT	1
AATT/TTAA	1	AATTCC/GGAATT	4
AGAT/ATCT	1	ACGGCG/CGCCGT	4
AGCC/GGCT	1	ATCACC/GGTGAT	4
ATAG/TCTA	7	ACGAGG/CCTCGT	3
AATC/TGAT	1	AAGGGG/CCCCTT	4
AGGG/CCCT	1	AAAAGG/CCTTTT	2
ATCG/CGAT	2	ACGGCG/CGCCGT	3
AAAAC/GTTTT	8	AGCCGG/CCGGCT	3
AAAAT/ATTTT	3	ATCAGC/GCTGAT	2
AAAGG/CCTTT	3	ATATCC/GGATAT	1
AACAG/CTGTT	2	AACGTC/GACGTT	1
AAGAG/CTCTT	4	AGCGGC/CGCTGC	2
AATGG/CCATT	1	ATCGTC/GACGAT	2
AAAAG/CTTTT	3	AAGGCG/CGCGTT	2
AATCC/CAATC	3	ACCAGC/GCTGGT	1
AGAGG/CCTCT	4	Compound SSRs	6
ACCGG/CCGGT	2	Total = 568	
ACAGC/GCTGT	4		
AGGGG/CCCCT	1		