Transcriptome Mining for EST-InDels and Development of EST-SSR Markers in Turmeric (Curcuma longa L.)

K.K. Sabu*, M.D. Shehenaz² and J. Amrutha³

¹Biotechnology and Bioinformatics Division, Jawaharlal Nehru Tropical Botanic Garden and Research Institute, Palode 695562, Thiruvananthapuram, Kerala, India
²Department of Computational Biology and Bioinformatics, University of Kerala, Thiruvananthapuram, 695581, Kerala, India
³Department of Bioinformatics, Union Christian College, Aluva 683102, Eranakulam, Kerala, India

Corresponding author: sabu@jntbgri.res.in (ORCID ID: 0000-0002-4312-1262)

Paper No. 705 Received: 05-04-2018 Accepted: 13-05-2018

ABSTRACT

Curcuma longa L. commonly known as turmeric is used as a culinary spice in India and many Asian countries. Turmeric has anti-fungal, anti-bacterial, anti-malarial and anti-cancer properties to mention a few. We have analyzed the transcriptome of C. longa retrieved from NCBI SRA database (SRR495223) for the development and validation of EST-SSR markers and identification of EST-InDels to be used for C. longa and related genera. A total of 337 primers were developed and 20 primers having the rating of 100% with the help of NetPrimer were selected and used for PCR validation in C. longa, Amomum subulatum Roxb. (large cardamom) and Elettaria cardamomum (L.) Maton (small cardamom). It was found that around 50% primers generated PCR products in both types of cardamom and 85% in C. longa. The developed primers worked with curcuma, large cardamom and small cardamom plants at varying levels. Besides, the transcriptome analysis detected high amount of deletion and 18 addition of bases which could be screened through the development of CAPS markers using the tool SNP2CAPS and it was found that 93 restriction enzyme can be used for screening these InDels.

Highlights

- Transcriptome sequence of Curcuma longa was used for the development of EST-SSR markers.
- The EST-SSR primers were validated in Curcuma longa and its related taxa such as Amomum subulatum and Elettaria cardamomum
- EST-InDels were also successfully identified in the transcriptome sequences of Curcuma longa.

Keywords: Curcuma longa, EST-SSR, EST-InDel, PCR, ClustalW, WebSat, NetPrimer, SNP2CAPS

Curcuma longa L. is a rhizomatous small plant from the ginger family (Zingiberaceae), commonly known as turmeric. For centuries in many parts of East Asian countries and India, turmeric has been used as an additive in daily cooking for colouring and flavoring, it is also being used as a household remedy for many curable conditions. Another most important use of turmeric is in making Kumkum. In alkaline pH turmeric turns red and it is widely used as Vermilion, also called as Kumkum which is an important component of Hindu religious ceremonies. The chemical constituents in turmeric comprises of a group of compounds known as curcuminoids, which includes curcumin (diferuloylmethane), demethoxycurcumin and bisdemethoxycurcumin. The curcumin represents 3.14% (on average) of powdered turmeric. It also includes some important volatile oils like termerone, zingiberene and atlantone with some common constituents like sugar, resins and proteins (Nagpal and Sood 2013).

Turmeric rhizomes are widely known for its bright yellow-orange characteristic color and offers as a less expensive yellow coloring in comparing with
saffron, thus it is also known as poor man’s saffron and turmeric gives a different taste in food than saffron. Turmeric provides a bright yellow dye that could be used as coloring for food, paints, textiles (fabric dye) and even in cosmetics. In some part of north India (Madhya Pradesh) turmeric is also used in making yellow paints for folk painting tradition. Turmeric rhizomes has also been used as traditional remedies by mixing with other plants to cure rage of conditions including headache, tonsillitis, wounds, snake bites, sprains, fractures (Ravindran 2007).

Curcumin is appreciated worldwide as an anti-malarial, anti-inflammatory and anti-tumor compound and it has also been known to modulate lipid metabolism, which has been implicated in obesity (Hamaguchi et al. 2010). Extracts from turmeric oil/oleoresin which contains curcuminoids and essential oils are used for flavoring and coloring (Haddad et al. 2011).

The genus Curcuma L. (Zingiberaceae) contains many taxa of economic, medicinal, ornamental and cultural importance. It has been considered that some species hybridize in the wild and the crosses could be naturalized (Skornickova and Sabu 2005; Skornickova et al. 2007a). It was detected six different chromosome counts (2n = 22, 42, 63, 70, 77 and 105) in Curcuma and the 2C-values varied from 1.66 pg in Curcuma vamana to 4.76 pg in Curcuma oligaantha, representing a 2.87-fold range (Skornickova et al. 2007b). This study also revealed intraspecific variations that exist in C.longa. These kinds of variations can be analyzed using molecular markers which can be developed through bioinformatics approaches. In fact, bioinformatics tools help in sequence analysis, gene finding, sequencing comparisons, transcriptome analysis and evolutionary genomics (Hogeweg 2011; Wong 2016). Through the advancement of next generation sequencing (NGS), though it is easy to develop genome wide single nucleotide polymorphisms (SNPs), usefulness of single sequence repeats (SSRs) or microsatellites is still high owing to easy scoring, high polymorphism, relatively high information generated through genetic analysis (Mayer et al. 2017). Many NGS projects have made transcriptome and the development of SSRs is relatively easier (Zhang et al. 2017). Besides, availability of transcriptomes aid in the development of genic SSRs (or EST-SSRs) which are more transferable among related species (Mathi Thumilan et al. 2016). In this background, a study was conducted to search for EST-InDels and develop EST-SSR markers from the transcriptome sequence of C. longa.

MATERIALS AND METHODS

The transcriptome of C. longa was retrieved from NCBI SRA database (SRR495223). Quality checking of the downloaded sequences was carried out using FASTQC1 toolkit which was followed by de novo assembly utilizing the tools available at DDBJ2 public server. The assembled sequence was downloaded from the server and searched for the ESTs using a Perl script, viz. EST_trimmer.pl3. The syntax which was used for running est_trimmer.pl was <FASTA file> [-amb=, win] [-tr5= (A|C|G|T), n, win] [-tr3= (A|C|G|T), n, win] [-cut=min, max] [-id=name][-help]. Identification of SSR regions was performed using Micro Satellite Identification Tool (MISA v1.0). Minimum unit size cut-off of 6 was used to report a di-nucleotide repeat, 4 for a trinucleotide repeat and 3 for SSRs of sizes 4, 5 and 6. A maximum distance of 100 nucleotides was allowed between two SSRs. In order to specify the search criteria, an additional file containing the microsatellite search parameters was saved as ‘misa.ini’. WebSat4 was used for microsatellite primer prediction. This was followed by submitting the primer sequences to Net Primer5 which analyzed secondary structures including hairpins, self-dimers, and cross-dimers in primer pairs. A rating of 100% indicates the absence of self and cross dimerization of primers. Twenty primers that satisfied the above criteria were selected for wet-lab validation and were custom synthesized at Sigma-Aldrich, Bengaluru. For validation of the EST-SSR primers, genomic DNA was isolated from young leaves of C. longa, Amomum subulatum Roxb. (large cardamom) and Elettaria cardamomum (L.) Maton (small cardamom) using Plant Genomic DNA Kit (Origin Labs, Kerala). Isolated DNA was loaded with 0.8% agarose gel and examined the gel under UV light. The quality of the DNA was also analyzed using Biophotometer (Eppendorf India Ltd). Polymerase chain reactions

---

1 http://www.bioinformatics.batrahm.ac.uk/projects/fastqc
2 http://www.ddbj.nig.ac.jp
3 http://www.pgrc.ipk-gatersleben.de/misa/download/est_trimmer.pl
4 http://www.wsmartins.net/websat
5 http://www.premierbiosoft.com/netprimer
were carried out in Veriti® Thermal Cycler (Applied Biosystems, USA). The final volume of 25 µl PCR reaction mixture comprised 50 ng of genomic DNA, 200 µM of each four dNTPs, 15 picomole of both forward and reverse primers, 1x Taq buffer and 1 unit Taq polymerase enzyme. The PCR conditions were: 94°C for 2 min, followed by 35 cycles of 94°C for 30 sec, specific annealing temperature for 1 min, 72°C for 2 min, and 72°C for 7 min for the final extension. The amplified products were separated using agarose (3%) in horizontal electrophoresis unit at 70V and visualized by ethidium bromide under UV using the gel documentation system. The bands obtained were scored by following expected product size as reference (Table 1).

For the identification of EST-InDels, retrieval of EST sequences from NCBI-EST under accession numbers, H0002100.1, H0002099.1, H0002097.1 and H0002094.1 were downloaded in FASTA format and subjected to multiple sequence alignment using Clustal W. Subsequently, SNP2CAPS tool was used to identify INDELs (INsertion/DELetion) where a single base has been deleted or inserted into one genome relative to another which is a symmetrical relationship, as a deletion in one corresponds to an insertion in another. The SNP2CAPS facilitated the computational conversion of SNPs or InDels into CAPS markers.

Table 1: Forward and reverse primers of the EST-SSR markers, their product sizes and annealing temperatures

| Primer Name | Forward and Reverse Sequence | Product Size (BP) | Annealing Temp. |
|-------------|------------------------------|-------------------|-----------------|
| F2R2        | CTTGAAATGTCTTGAGCACGAT       | 115               | 52.0°C          |
|             | ATGAGTCCGAAAGTCGTTT         |                   |                 |
| F4R4        | GATGAGGAGCCGAGGGGAG         | 161               | 60.0°C          |
|             | ACAGCATTTTACCTCACCAGT       |                   |                 |
| F5R5        | TCTTTGTCACAAATCCTGC         | 192               | 62.2°C          |
|             | AAGTGCCGGAATCACAAGGA        |                   |                 |
| F6R6        | TCTAAAACAGCCGAGCTCAACAG     | 205               | 60.0°C          |
|             | CAGAACGGAGAGAGAGATGAT       |                   |                 |
| F7R7        | GAGGTAAGATGCTGTTGAG         | 308               | 63.4°C          |
|             | CTTCCTTCCCTCTCCCTCTCTC      |                   |                 |
| F8R8        | CCTCATTTTATCTCGACAGTTT      | 115               | 66.0°C          |
|             | CAGCCAACACCTCCTAAGCAT       |                   |                 |
| F9R9        | TCTTTCCCCCTCTTGCCTC         | 313               | 63.4°C          |
|             | ACAGAAAGACGGAGAGGTGACCT     |                   |                 |
| F10R10      | TGAAAGGAGGTAAGAAGACGA       | 313               | 60.9°C          |
|             | GAGAAGCAGGTGAGGGAGTAGT      |                   |                 |

RESULTS AND DISCUSSION

The raw transcriptome sequence (SRA id: SRR4955223; size: 3.2Gb) of Curcuma longa checked using FASTQC showed that the sequence quality was satisfactory for further analysis. After assembly using Trinity tool in DDBJ the sequence size was reduced to 99Mb. The MISA tool helped to identify 6206 SSRs where mono nucleotides represent 40%, di nucleotides 18%, tri nucleotide 38%, tetra nucleotides 1.7%, penta nucleotide 0.43% and hexa nucleotide 0.5%. A total of 337 primers were designed using Websat. Then these primers were used for checking primer efficiency using the tool NetPrimer. This resulted in 30 primers having 100% efficiency and of which 20 were selected for wet-lab validation with C. longa, A. subulatum and E. cardamomum. About 85% of the primers generated PCR products with C. longa and 50% of the primers worked in both the cardamom species (Fig. 1). These primers can be used to study population genetics including phylogenetic analyses and estimation of genetic diversity of C. longa and other related genera.

In order to identify EST-InDels, the sequences were retrieved from NCBI-EST under accession numbers HO002100.1, HO002099.1, HO002097.1 and HO002094.1. The sequences were aligned together and saved as Fasta format for further analysis using
ClustalW program. The aligned EST sequences showed high number of deletions and 18 additions which are called InDels. They were screened for the development of CAPS markers using the tool SNP2CAPS which predicated 99 CAPS candidates. These InDels could be screened through the development of CAPS markers using the tool SNP2CAPS and it was found that 93 restriction enzymes can be used for screening the InDels.

CONCLUSION

The developed SSR primers worked with curcuma, large cardamom and small cardamom plants at varying levels. The InDel markers identified in the study can be experimentally validated with a set of common restriction enzymes. Both EST-SSR and EST-InDels are useful to perform genetic mapping, marker-assisted selection in breeding programmes and population genetics including phylogenetic analyses and estimation of genetic diversity in *C. longa* and other related genera.

ACKNOWLEDGMENTS

Authors thank the Director, JNTBGRI for the facilities extended to carry out this research programme. SMD and AJ acknowledge their gratitude to members of the Plant Molecular Biology Laboratory of Biotechnology and Bioinformatics Division, JNTBGRI for all academic and technical help. Authors are deeply indebted to Jinu Thomas (JNTBGRI) for her assistance offered to develop the bioinformatics pipeline required for the marker development.

Author contributions: SMD and AJ carried out the bioinformatics analysis and EST-SSR primer validation experiments. SMD prepared the first version of the manuscript. KKS designed the study, supervised the project and prepared the final version of the manuscript.

REFERENCES

Gupta, S. and Prasad, M. 2009. Development and Characterization of Genic SSR Markers in *Medicago truncatula* and Their Transferability in Leguminous and Non-Leguminous Species. *Genome*, 52: 761-771.

Haddad, M., Sauvain, M. and Deharo, E. 2011. Curcuma as a parasiticidal agent: a review. *Planta Med.*, 77: 672–678.

Hamaguchi, T., Ono, K. and Yamada, M. 2010. REVIEW: Curcumin and Alzheimer’s disease. *CNS Neurosci. Ther.*, 16: 285–97.

Hogeweg, P. 2011. The Roots of Bioinformatics in Theoretical Biology. *PLoS Computational Biology* 7: e1002021.

Jain, J., Patil, G.B., Bhargava, P. and Nadgauda, R.S. 2014. In Silico Mining of EST-SSRs in *Jatropha curcas* L.
Transcriptome Mining for EST-InDels and Development of EST-SSR Markers in Turmeric (Curcuma longa L.)

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.

Nagpal, M. and Sood, S. 2013. Role of curcumin in systemic and oral health: An overview. *J. Nat. Sci. Biol. Med.*, **4**: 3-7.

Ravindran, P.N. 2007. Turmeric: the genus Curcuma. CRC Press, London.

Schlötterer, C. and Tautz, D. 1992. Slippage synthesis of simple sequence DNA. *Nucleic Acids Res.*, **20**: 211-215.

Skornickova, J. and Sabu, M. 2005. *Curcuma roscoeana* Wall. (Zingiberaceae) in India. *Gardens’ Bulletin Singapore*, **57**: 187-198.

Skornickova, J., Sida, O., Jarolimova, V., Sabu, M., Fer, T., Travnic, P. and Suda, J. 2007b. Chromosome Numbers and Genome Size Variation in Indian Species of *Curcuma* (Zingiberaceae). *Annals of Botany*, **100**: 505-526.

Skornickova, J., Rehse, T. and Sabu, M. 2007a. Other economically important *Curcuma* species. In: Turmeric: the genus *Curcuma*. Ravindran PN, Babu KN, Sivaraman K (Eds). Boca Raton, FL: CRC Press, pp. 451-467.

Varshney, R.K., Graner, A. and Sorrells, M.E. 2005. Genic Microsatellite Markers in Plants: Features and Applications. *Trends in Biotechnology*, **23**: 48-55.

Wong, K.C. 2016. Computational Biology and Bioinformatics: Gene Regulation. CRC Press Taylor & Francis Group.

Meyer, L., Causse, R., Pernin, F., Scalone, R., Bailly, G. and Chauvel, B. *et al.* 2017. New gSSR and EST-SSR markers reveal high genetic diversity in the invasive plant *Ambrosia artemisiifolia* L. and can be transferred to other invasive *Ambrosia* species. *PLoS ONE*, **12**(5): e0176197.

Zhang, Y., Zhang, X., Wang, Y.H. and Shen, S.K. 2017. *De Novo* Assembly of Transcriptome and Development of Novel EST-SSR Markers in *Rhododendron rex* Lévl. Through Illumina Sequencing. *Frontiers in Plant Science*, **8**: 1664.

Mathi Thumilan, B., Sajeevan, R.S., Biradar, J., Madhuri, T., Nataraja, N.K. and Sreeman, S.M. 2016. Development and Characterization of Genic SSR Markers from Indian Mulberry Transcriptome and Their Transferability to Related Species of Moraceae. *PLoS ONE*, **11**(9): e0162909.
