Genes, Genomes and Genomics ©2009 Global Science Books

SpicEST – An Annotated Database on Expressed Sequence Tags of Spices

Arumugam Chandrasekar • Aikkal Riju • Nambissan V. Sathyanath • Santhosh J. Eapen*

Bioinformatics Centre, Indian Institute of Spice Research, P. B. No 1701, Marikunnu P.O Calicut -673012, Kerala, India

Corresponding author: * **sjeapen@spices.res.in**

ABSTRACT

SpicEST is an attempt to develop a comprehensive database on ESTs of two spice plants, ginger and turmeric. For this, EST records were downloaded from NCBI open source database. All EST records for *Curcuma longa* and *Zingiber officinale* were mined and stored in MYSQL database. These ESTs were assembled using CAP3 sequence assembly program. The resultant contigs were stored in the database tissue- and crop-wise. By navigating the menus one can retrieve the results of SSRs identified from these ESTs by using five different SSR identification tools – MISA, ETRA, SSR PRIMER, SSRIT and WEB TROLL. Using these, SSR primers were designed with the help of PRIMER3 software and stored in the database after checking their quality using FAST PCR. Primers are being validated in wet lab studies. All the ESTs were annotated using the ESTPASS server. These results were used to find the putative genes and the respective metabolic pathways. The database also contains information on single nucleotide polymorphism in both these spices. Two different programs like CAP3 and AUTOSNP were used for this analysis and results can be retrieved by clicking on the respective tab. 'SpicEST' database is an initiative to promote spice genomics studies and is envisaged as an online tool for spice researchers. It is available as an online database at www.spices.res.in/spiceest.

Keywords: annotation, *Curcuma longa*, expressed sequence tag ginger, primers, single nucleotide polymorphism, turmeric, *Zingiber officinale*

INTRODUCTION

Spices include dried seeds, fruits, roots, bark or vegetative substances used in nutritionally insignificant quantities as a food additive for the purpose of flavoring, and sometimes as a preservative by killing or preventing the growth of harmful bacteria. Many of these substances are also used for other purposes, such as medicine, religious rituals, cosmetics, perfumery or eating as vegetables. Spices constitute an important group of agricultural commodities which are virtually indispensable in the culinary art. In India, spices are important commercial crops from the point of view of both domestic consumption and export.

Due to their numerous medicinal properties there is a need to study the gene expression in spice plants. For genomic studies in spices mainly expressed sequence tags (ESTs) are used. ESTs are partial sequences of a clone, randomly selected from a cDNA library and used to identify genes expressed in a particular tissue. An EST is a tiny portion of an entire gene that can be used to help to identify unknown genes and to map their positions within a genome (Carson and Botha 2000) and is a short sub-sequence of a transcribed spliced nucleotide sequence. ESTs represent a snap-shot of what is expressed in a given tissue, and/or at a given developmental stage. They represent tags of expression for a given cDNA library (Brenner *et al.* 2003). One of the many interesting applications of EST database (dbEST) is gene discovery where many new genes can be found by querying the dbEST with a protein or DNA sequences. In addition, EST collections are good sources of simple sequence repeats (SSRs) and single-nucleotide polymorphisms (SNPs) that can be used for creating saturated genetic maps (Fei *et al.* 2006; Newcomb *et al.* 2006) Thus, EST collections have been generated for many plant species, being the most comprehensive those of *Arabidopsis* and rice (Heiko *et al*. 2003). Spices have been less extensively surveyed, but important large numbers of EST collections are publicly available for turmeric and ginger only. The dbEST database of the National Centre for Biotechnology Information (NCBI) contains (August 14, 2009) 62,838,171 ESTs. dbEST is a new resource that contains data from laboratories generating incomplete "single-pass" cDNA sequences.

MATERIALS AND METHODS

Data: *Curcuma longa* have 12593 sequences and *Zingiber officinale* 38139 EST sequences in dbEST. In this study the turmeric and ginger EST sequences were used for creating a database and to discover SSRs, SNPs, primers and putative genes.

Database implementation: The data was stored in a MySQL database running on an Apache server in a Linux Operating system. The database architecture is shown in **Fig. 1**.

Database interface: The database interface is implemented using PHP, HTML and Java. The sample interface is given in **Figs. 2** and **3**.

RESULTS AND DISCUSSION

SpicEST consists of three databases of SSRs, SNPs and an annotated EST database. In the present study EST resources for database development were downloaded from the NCBI dbEST, and assembled using the sequence assembly program, CAP3 (http://mobyle.pasteur.fr/cgi-bin/MobylePortal/portal.py?form=cap3) (freeware). The assembly of 12593 turmeric ESTs resulted in 3251 contigs and 1630 singletons. In the case of 38,139 ginger ESTs, they were assembled into 7100 contigs and 4890 singletons. The main navigation menus of the database are EST/contigs, spices SSRs, SNPs results, primer results, ESTs annotation and gene search. EST/contigs would show the tables of contigs and singletons of whole ESTs of ginger and turmeric. On clicking the tab SSRs, an interface is provided in the form of a combo box. By selecting the combo box of a crop and any one of the five different SSR finding tools (MISA,

Received: 19 August, 2009. Accepted: 16 September, 2009.

Research Note

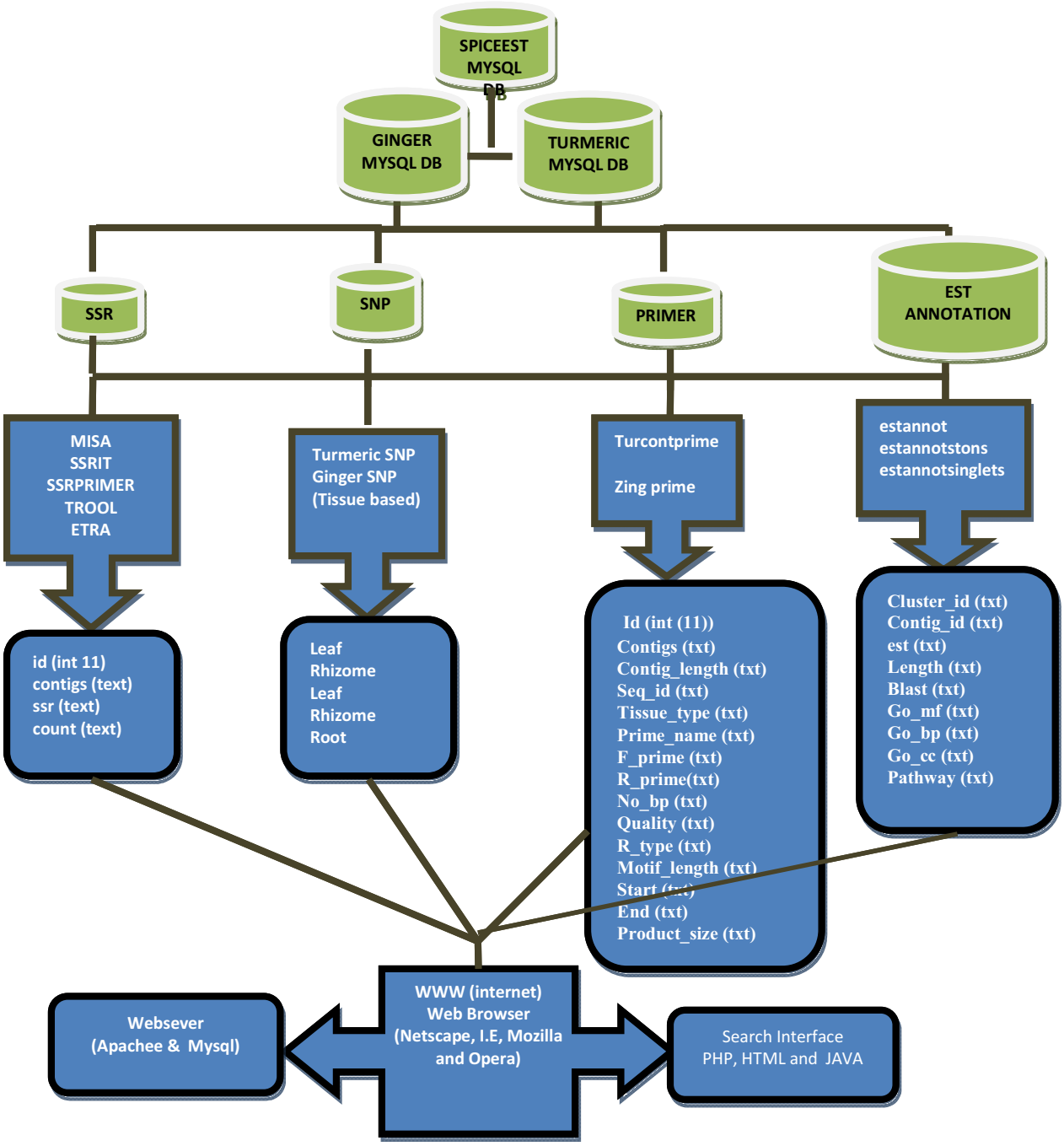

Fig. 1 Architecture of SPICESTdb. Scheme of SPICESTdb describing the outline of the database.

SSRIT, SSRPRIMER, WEBTROLL, ETRA) (Varshney *et al.* 2002; Blenda *et al.* 2006; Riju *et al.* 2009) the details of SSRs found for the respective crop can be obtained. In SNPs the results are provided as a whole and as tissue-based EST libraries. By clicking the arrow mark, we can retrieve the results of SNP. Primers were designed for the EST resources by using the online tool PRIMER3 (http://frodo.wi.mit.edu/) (freeware) (Prasad *et al.* 2005). This software tool will provide a set of forward and reverse primers with additional four oligos in decreasing order of quality. The quality of the primers is determined by checking the GC content (50-60%), melting temperature (55-80°C), without self complimentarity and mispriming (Thiel *et al.* 2003). These primers designed were checked for above mentioned quality using FAST PCR program (http://www.biocenter.helsinki.fi/bi/Programs/fastpcr.htm) (freeware). The quality primers were stored in the database and the results can be retrieved by clicking the results button. ESTpass (http://estpass.kobic.re.kr/) (freeware) server is used as an online analysis for annotation. The menu EST annotation provides a double combo box for crop and tissues, and the user can retrieve the summary results of annotation by selecting the respective spice and corresponding tissues of leaves or rhizomes. By clicking the results button, a table view of putative genes of the respective contigs, singletons and singlets are provided. In the table BLAST, GO and PATHWAY links are provided which will show the respective annotation of the ESTs (Sterky *et al.* 2004; Missirlis *et al.* 2005). The gene search page will provide information on the type of putative gene. Users can select the tissues and type the known putative gene of interest for searching in the existing database which will provide the details of the gene if there is a suitable match (Aishwarya *et al.* 2007; Archak *et al.* 2007; Gonzalez-Ibeas *et al.* 2007). Additional links to SSR tools and other EST databases are also provided.

CONCLUSIONS

The information content of the database is useful in different areas of research like gene tagging, genome mapping, population genetics, germplasm characterization and in understanding microsatellite dynamics in spice genomes

Fig. 2 Overall search interface of SPICEST database (SSR and SNP database).

Fig. 3 Overall search interface of SPICEST database (Annotated EST database and other tools).

and this will also serve as a reference database on spice ESTs. The database is a good utility for crop improvement programmes in spices. It is freely available in public domain and further information can be obtained from the authors. We plan to include datasets of other spice crops like pepper, nutmeg, clove, chillies, cinnamon and cardamom. An advanced graphical tool will be added to facilitate display and visualization of the underlying data. SpicESTdb is a platform independent relational database publicly available at www.spices.res.in/spiceest/.

ACKNOWLEDGEMENTS

We thank the web resource providers for data collection and database development. This work was supported by a grant from Department of Biotechnology (BTISnet), New Delhi, India.

REFERENCES

Aishwarya V, Grover A, Sharma PC (2007) EuMicroSatdb: A database for microsatellites in the sequenced genomes of eukaryotes. *BMC Genomics* **8**, 225

Archak S, Meduri E, Kumar PS, Nagaraju J (2007) InSatDb: A microsatellite database of fully sequenced insect genomes. *Nucleic Acids Research* **35**, D36-D39

Blenda A, Scheffler J, Scheffler B, Palmer M, Lacape JM, Yu JZ, Jesudurai C, Jung S, Muthukumar S, Yellambalase P, Ficklin S, Staton M, Eshelman R, Ulloa M, Saha S, Burr B, Liu S, Zhang T, Fang D, Pepper A, Kumpatla S, Jacobs J, Tomkins J, Cantrell R, Main D (2006) CMD: A cotton microsatellite database resource for *Gossypium* genomics. *BMC Genomics* 7, 132

Brenner ED, Stevenson DW, McCombie RW, Katari MS, Rudd SA, Mayer KF, Palenchar PM, Runko SJ, Twigg RW, Dai G, Martienssen RA, Benfey PN, Coruzzi GM (2003) Expressed sequence tag analysis in *Cycas*, the most primitive living seed plant. *Genome Biology* **4**, R78

Carson DL, Botha FC (2000) Preliminary analysis of expressed sequence tags for sugarcane. *Crop Science* **40**, 1769-1779

Fei Z, Tang X, Alba R, Giovannoni J (2006) Tomato Expression Database (TED): A suite of data presentation and analysis tools. *Nucleic Acids Research* **34**, D766-D770.

Gonzalez-Ibeas D, Blanca J, Roig C, González-To M, Picó B, Truniger V, Gómez P, Deleu W, Caño-Delgado A, Arús P, Nuez F, Garcia-Mas J, Puigdomènech P, Aranda MA (2007) MELOGEN: an EST database for melon functional genomics. *BMC Genomics* **8**, 306-322

Heiko S, Wojciech MK (2003) Comparison of rice and *Arabidopsis* annotation. *Current Opinion in Plant Biology* **6**, 106-112

Missirlis PI, Mead CR, Butland SL, Ouellette BF, Devon R S, Leavitt BR, Holt RA (2005) Satellog: A database for the identification and prioritization of satellite repeats in disease association studies. *BMC Bioinformatics* **10**, 145

Newcomb RD, Crowhurst RN, Gleave AP, Rikkerink EHA, Allan AC, Beuning LL, Bowen JH, Gera E, Jamieson KR, Janssen BJ (2006) Analyses of expressed sequence tags from apple. *Plant Physiology* **141**, 147-166

Prasad MD, Muthulakshmi M, Arunkumar KP, Madhu M, Sreenu VB, Pavithra V, Boseb B, Nagarajaram HA, Mita K (2005) SilkSatDb: a microsatellite database of the silkworm, *Bombyx mori*. *Nucleic Acids Research* **33**, D403-D406

Riju A, Rajesh MK, Fasila Sherin PTP, Chandrasekar A, Elain Apsara S, Arunachalam V (2009) Mining of expressed sequence tag libraries of cacao for microsatellite markers using five computational tools. *Journal of Genetics* **88**, 217-225

Sterky F, Bhalerao RR, Unneberg P, Segerman B, Nilsson P, Brunner AM, Charbonnel-Campaa L, Lindvall JJ, Tandre K, Strauss SH, Sundberg B, Gustafsson P, Uhlen M, Bhalerao RP, Nilsson O, Sandberg G, Karlsson J, Lundeberg J, Jansson S (2004) A *Populus* EST resource for plant functional genomics. *Proceedings of the National Academy of Sciences USA* **101**, 13951-13956

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.). *Theoretical Applied Genetics* **106**, 411-422

Varshney RK, Thiel T, Stein N, Langridge P, Garner A (2002) *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cell and Molecular Biology Letter* **7**, 537-546