MAIN

OJB™

Online Journal of Bioinformatics ©

Volume 10 (2): 224-232, 2009.

Single nucleotide polymorphisms (SNPs) and indels in expressed sequence tag libraries of turmeric (*Curcuma longa* L.)

Chandrasekar A[1], Riju A[1], Sithara K[2], Anoop S [3], Santhosh J Eapen*[1]

[1]Bioinformatics Centre, Indian Institute of Spices Research, Marikunnu P.O, Calicut – 673012, Kerala, India, [2] Kandiyl House, Makkada (P.O), Kakkodi, Calicut – 673617. Kerala. India. [3]Ottankulam, Athicode (P.O), Palakkad (Dist), Kerala – 678554. India.
*Corresponding author: sjeapen@spices.res.in

ABSTRACT

Chandrasekar A, Riju A, Sithara K, Anoop S , Santhosh J Eapen Single nucleotide polymorphisms (SNPs) and indels in expressed sequence tag libraries of turmeric (Curcuma longa L.), Online J Bioinformatics, Volume 10 (2): 224-232, 2009. Expressed sequence tag (EST) sequencing programs provide methods to identify novel genes. EST's are a source of biologically useful SNPs due to the relatively high redundancy of gene sequences and diversity of genotypes represented within databases. EST based SNPs are potential molecular markers for genetic improvement. 27599 and 2506 polymorphic sites (SNPs and indels) were found in turmeric, a medicinal spice plant. Frequencies of SNPs and indels were 1/91 bp and 1/998 bp, respectively. The transition to transversion ratio was 0.713, which showed a relative increase in transversion type mutation. The program is available (http://spices.res.in/spicesnip).

Keywords: *Curcuma longa*, expressed sequence tag, mutation, nucleotide, transition, transversion, turmeric

INTRODUCTION

Turmeric (*Curcuma longa* L.) is a rhizomatous herbaceous perennial plant of the ginger family, Zingiberaceae which is native to tropical South Asia. It needs temperatures between 20 and 30°C and a considerable amount of annual rainfall to thrive. The family Zingiberaceae contains 49 genera and 1400 species. The main constituent of turmeric contains an essential oil (max. 5%) which contains a variety of sesquiterpenes, many of which are specific for the species. Most important ones are turmerone (max. 30%), ar-turmerone (25%) and Zingiberene (25%). Conjugated diarylheptanoids or curcumin are responsible for the orange colour and probably also for the pungent taste. Origin of turmeric is believed to be from South East Asia. The genus *Curcuma* belonging to the family *Zingiberaceae* has a wide spread occurrence in the tropics of Asia to Africa and Australia. Apart from *Curcuma longa* or C. *domestica*, the common culinary turmeric, there are several other species of *Curcuma* which are mainly used as coloring agents, for production of arrowroot and also for medicinal purposes.

Expressed Sequence Tags (ESTs) provide researchers with a quick and inexpensive route for discovering new genes, for obtaining data on gene expression and regulation, and for the construction of the genome maps. A single nucleotide polymorphism (SNP) is a DNA sequence variation occurring when a single nucleotide - A, T, C, or G - in the genome differs between members of a species (or between paired chromosomes in an individual). SNPs may fall within coding sequences of genes, non-coding regions of genes, or in the intergenic regions. SNPs within a coding sequence will not necessarily change the amino acid sequence of the protein that is produced, due to degeneracy of the genetic code. If DNA sequence in which any change in the base pairs do not result in the change of polypeptide sequence, then it is termed synonymous (sometimes called a silent mutation) - if a different polypeptide sequence is produced, they are non-synonymous. SNPs that are not in protein-coding regions may still have consequences for gene splicing, transcription factor binding, or the sequence of non-coding RNA. The study of single nucleotide polymorphisms is also important in crop and livestock breeding programs. Therefore, ESTs are an important resource for identifying polymorphisms in transcribed regions.

SNPs have been shown to be the most abundant source of DNA polymorphism in human, animal and plant genomes. They are the most common type of alleles found within and between varieties of a crop species. SNPs possess desirable properties as molecular markers. Biallelism makes them easy to score in high throughput genotyping assays. Molecular genetic markers developed from ESTs can be used to examine a group of individuals or populations to estimate various diversity measures and genetic distances, infer genetic structure and clustering patterns, test for Hardy-Weinberg equilibrium and multi-locus equilibrium, and to test polymorphic loci for evidence of selective neutrality. They are useful to plant breeders, germplasm managers, and population geneticists. The use of EST sequence data for the identification of SNPs has many advantages that can be exploited to facilitate the development of highly dense genetic maps and markers assisted breeding programs (Varshney et. al., 2005). SNPs can be used to saturate genetic maps in plants (Bhattramakki and Rafalski, 2001).

Recently EST resources of turmeric cultivar, orange turmeric are being developed at The University of Arizona, US, and deposited at (http://www.ncbi.nlm.nih.gov/dbEST/) dbEST. EST sequencing programs have provided a wealth of information, identifying novel genes from a broad range of organisms and providing an indication of gene expression level in particular tissues (Adams et. al., 1995). EST sequence data may provide the richest source of biologically useful SNPs due to the relatively high redundancy of gene sequence, the diversity of genotypes represented within databases, and the fact that each SNP would be associated with an expressed gene (Picoult-Newberg et. al., 1999). SNP detection perl script AutoSNP version.1.0 was used to find the SNP site information and transition vs transversion analysis (Barker et. al., 2003). EST-SNP can be detected by using other programs or servers such as SEAN (Huntley et. al., 2006), PolyPhred (Nickerson et. al., 1997), PolyBayes (Marth et. al., 1999), TRACE_DIFF (Bonfield et. al., 1998), HaploSNPer (Tang et. al., 2008) and HarvEST (http://harvest.ucr.edu) but AutoSNP provides user friendly approach and interpretable result as html file. SNPs can be classified based on their nucleotide substitution as either transition (C/T or G/A) or transversion (C/G, A/T, C/A or T / G). Indel sites can classified to four groups based on the nucleotide involved (A/T/C/G). Thus there are ten kinds of SNP/indel (two types of transition and four types of transversion and four groups of indels) are possible in the SNP/indel sites in EST libraries. The objective of this study was to identify SNPs in turmeric, a common spice plant, using the AutoSNP tool.

MATERIALS AND METHODS

EST database, dbEST of NCBI release 060509 contains 12593 expressed sequence tags which belong to two tissue libraries, rhizome (6870 ESTs) and young leaves (5723 ESTs) of turmeric. The retrieved sequences were separated tissues wise and clustered in ace format by using contig building package Cap3 (Huang and Madan, 1999) server (http://mobyle.pasteur.fr/cgi-bin/MobylePortal/portal.py#). AutoSNP version.1.0 was used to find the candidate SNPs from these contigs. The transition (Ts) versus transversion (Tv) ratio was also calculated for all the libraries to find the DNA substitution in turmeric genome.

We used Shannon information index for indexing the distribution of SNP/Indels into ten possible categories. Frequency of each of the ten types of SNP/indel sites was scored. From this value, proportion (Pi) of occurrence of each type (nature of transition / transversion / indel) to the total SNP/indels in each tissue library was worked out. Shannon index estimates (Weaver and Shannon, 1949) have been worked out using the formula

$$H' = -\sum_{i=1}^{S} p_i \ln p_i \qquad \text{------ (1)}$$

Where S is the total number of SNP/indel states (10) , pi= proportion of ESTs in the ith type of SNP/indel state. The calculated value is divided by the $\log_2 10$ to get uniformity. We divided the summation value by 0.5 S * ln 0.5 to normalize the index for easy comparison among different contigs where S is the total number of EST sequences used in analysis. The final Shannon's index value thus obtained is multiplied by 100 for easy documentation and comparison.

RESULTS AND DISCUSSION

It was found that 27599 SNPs and 2506 Indels sites in turmeric (Table 1). SNPs occurred at a frequency of one out of every 91 bp and indels at one in every 998 bp. The SNPs were classified according to the nucleotide substitutions as either transition (Ts) or transversion (Tv). A total of 11492 transitions and 16107 transversions were reported in this study. The transition and transversion ratio is very low in young leaves (0.644) than rhizome (0.804), the overall transition (Ts) vs transversion (Tv) ratio is 0.713, which shows that the relative increase of transversion over transition. Germano and Klein (1999) identified one SNPs in 200 bp of cDNA of *Picea rubens* and *Picea mariana*, and also discovered SNPs in the chloroplasts of these species. Recently 62 SNPs and 12 indels in 21 Unigenes are reported in tomato (Joanne et al., 2005). In soybean (*Glycine max*), two SNPs were found approximately every 400 bp (Coryell et al., 1999). In maize (*Zea mays*), SNPs occur frequently, with one SNP in every 48 bp and every 130 bp in 3' untranslated regions and coding regions, respectively (Tenaillon et al., 2001), (Rafalski, 2002). SNP frequency in apple (*Malus domestica*) ESTs is one in every 706 bp (Newcomb et al., 2006) while in beetroot it was one in every 130 bp. EST-SNPs in oil palm showed a frequency of 1.36 SNPs / 100 bp (Riju et al., 2007). In rye transcriptome the SNP frequency was found to be one SNP per 58 bp and one indel per 214 bp (Varshney et al., 2007). The SNP frequencies in wheat (1/540bp) (Somers et al., 2003), maize (1/61bp) (Ching et al., 2002), barley (ranging between 1/78 and 1/189bp) (Kanazin et al., 2002; Bundock et al., 2003; Russell et al., 2004) and sorghum (1/123bp) (Hamblin et al., 2004) are also available. SNP frequency in turmeric was higher than soybean, apple, beetroot, wheat, sorghum etc. but comparable to that of maize, oil palm, rye, barley etc.

Table 1. Incidence and frequency of SNPs and indels in tissue-specific EST libraries of turmeric

Tissue Name	No of ESTs	No of contigs	Consensus size (bp)	SNP and Indel sites	Transitions (Ts)	Transversions (Tv)	Indels	Ts / Tv	Frequency of indels in bp	Frequency
Leaves	4741	1632	1173735	16264	5875	9117	1272	0.644	1Indel/923	1SNP/78bp
Rhizome	5909	1774	1326902	13841	5617	6990	1234	0.804	1Indel/1075	1SNP/105bp
Total	10650	3406	2500637	30105	11492	16107	2506	0.713	1Indel/998	1SNP/91bp

EST analysis of beetroot, oil palm and maize reported a higher rate of transition than transversion. According to several authors (Wakeley, 1994; 1996; Purvis & Bromham, 1997; Ina, 1998; Yang & Yoder, 1999; Strandberg & Salter, 2004) during DNA sequence evolution the rate of transitions differs from the rate of transversions, with transitions occurring more frequently than transversions. This is important because it provides insight into the process of molecular evolution. However, in this study, we found that the Ts vs Tv ratio was 0.713. Because transitions are thought to be greatly more frequent than transversions, the present finding is quite surprising. Similar observations were made in a study of genes of rice chromosome 8 [Wu *et al.,* 2004], but the reason for these difference in the substitution pattern are unclear. Similar trends have been observed in Rafnia Thunb.(Fabaceae, Crotalarieae), genes on plastid regions there ts:tv for *trnL-F* (0.48), *rps 16* (0.5) and *accD-psa1* (0.56) were more or less (Motsi et al., *2008).* This is in contrast with the literature where low levels of genetic divergence are usually associated with high Ts: Tv ratios. Zhao *et al.* (2002) observed that the probability of a transversion increased when the number of purines increased (i.e., 0 → 1 → 2) at the immediate adjacent sites while studying human genome SNPs. Putative functions of the entire cluster have been identified using the standard homology searches within the non-redundant protein database (NCBI BlastX) and most of them have shown known functions. The known function can give added value to eSNP marker, since the possibility exists that SNPs can then be directly associated with variants for a specific function (Holton et al., 2002; Gao et al., 2004). The SNPs reported here can facilitate the development of SNP markers for both turmeric genetics and breeding programs. The tissue wise EST clusters and their SNP and indel site information is made available at http://spices.res.in/spicesnip.

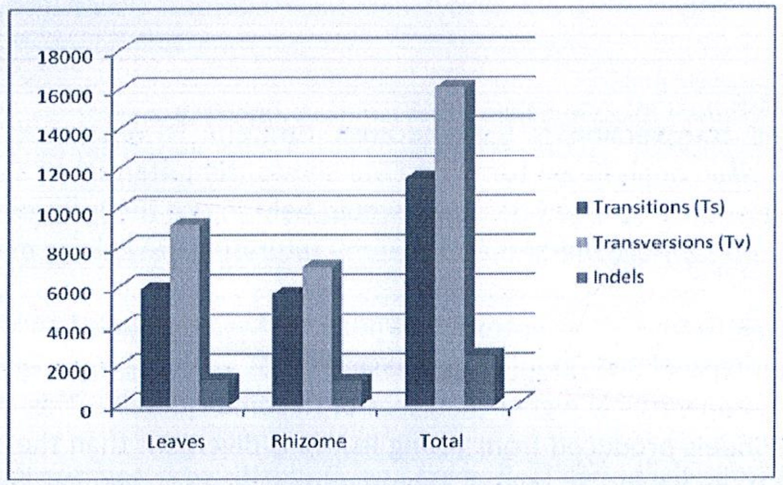

Figure 1: DNA substitution of SNPs and indel polymorphisms in EST libraries of different tissues of turmeric.

CONCLUSION

Generally single nucleotide polymorphism (SNP) discovery and genotyping are essential to genetic mapping. SNPs are indispensable in such applications as association mapping and construction of high-density genetic maps. The present study represents the initiation of SNP discovery and development in turmeric. In total, we have reported 27599 SNPs and 2506 indel sites in turmeric ESTs. SNP frequency is observed one every 91 bp and indels at 998 bp. Transition to transversion ratio shows that the transversion is occurring in this genome more frequent than transition. Overall transversion is high because turmeric is vegetative propagated through rhizomes. This *in silico* analysis on turmeric ESTs shows the potential of SNP markers for use in turmeric breeding and the created database would help to use the information in designing new primers and developing more markers and thereby saturating the linkage maps.

The Shannon index (Information entropy) formula (1) was employed to find the probability in distribution of ten type of SNP/indel in tissues and crops wise. Shannon index reported in young leaves (0.409) is comparatively higher than that of rhizomes (0.329). This showed distribution of ten types of SNP/indels produced from young leaves is divergent than the other tissue library of rhizomes (Table 2). By having an overall Shannon index (0.182) and SNP frequency (1SNP/91bp) we can conclude that the ten types of nucleotide substitutions occurring in this crop are more frequent and have a similar distribution.

Table 2. The nucleotide substitution pattern and Shannon indices of SNPs observed in tissue specific EST libraries of turmeric.

Nucleotide Substitution	Leaves	Rhizome	Total
C/T	2773	2646	5419
G/A	3102	2971	6073
Total (Transition)	5875	5617	11492
A/T	1829	1458	3287
C/G	2638	1818	4456
G/T	2350	1738	4088
A/C	2300	1976	4276
Total (Transversion)	9117	6990	16107
A	358	365	723
C	307	304	611
G	284	253	537
T	323	312	635
Total(Indel)	1272	1234	2506
Shannon index (x100)	0. 409	0. 329	0. 182

REFERENCES

Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.G., Gocayne, J.D. and White, O. (1995). Initial assessment of human gene diversity and expression patterns based upon 83-million nucleotide of cDNA sequence. Nature, 377: 3

Barker, G., Batley, J., O'sullivan, H., Edwards, K.J. and Edwards, D. (2003). Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. Bioinformatics, 19: 421-422.

Bhattramakki, D. and Rafalski, A. (2001). Discovery and application of single nucleotide polymorphism markers in plants. In: Henry R.J. (ed.) Plant Genotyping: the DNA Fingerprinting of Plants. CABI Publishing, Oxon, UK, pp. 179-191.

Bonfield, J.K., Rada, C. and Staden, R. (1998). Automated detection of point mutations using fluorescent sequence trace subtraction. Nucleic Acids Research, 26: 3404–3409.

Bundock, P.C., Christopher, J.T., Eggler, P., Ablett, G., Henry, R.J., Holton, T.A. (2003). Single nucleotide polymorphisms in cytochrome P450 genes from barley. Theoretical Applied Genetics, 106: 676–682.

Ching, A., Caldwell, K.S., Jung, M., Dolan, M., Smith, O.S., Tingey, S., Morgante, M., Rafalski, A.J. (2002). SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC Genetics,* 3: 19.

Coryell, V.H., Jessen, H., Schupp, J.M., Webb, D. and Keim, P. (1999). Allele-specific hybridisation markers for soybean. Theoretical Applied Genetics, 101: 1291–1298.

Gao, L.F, Jing, R.L, Hu, N.X, Li, X.P, Zhou, R.H, Chang, X.P, Tang, J.F and Ma Zy, Jia J.Z.(2004) One hundred and one new microsatellite loci derived from ESTs (EST-SSRs) in bread wheat. Theoretical Applied Genetics, 108:1392–1400.

Germano, J. and Klein, A.S. (1999). Species specific nuclear and chloroplast single nucleotide polymorphisms to distinguish *Picea glauca*, *P. mariana* and *P. rubens*. Theoretical Applied Genetics, 99: 37–49.

Hamblin, M.T., Mitchell, S.E., White, G.M., Gallego, J., Kukatla, R., Wing, R.A., Paterson, A.H., Kresovich, S. (2004). Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of *Sorghum bicolor*. Genetics, 167: 471–483.

Holton, T.A, Christopher, J.T, McClure, L., Harker, N. and Henry, R.J. (2002) Identification and mapping of polymorphic SSR markers from expressed gene sequences of barley and wheat. Molecualr Breeding, 9:63–71.

Huang, X. and Madan, A. (1999). CAP3: a DNA sequence assembly program. Genome Research, 9: 68–877.

Huntley, D., Baldo, A., Johri, S. and Sergot, M. (2006). SEAN: SNP prediction and display program utilizing EST sequence clusters. Bioinformatics, 22 (4): 495-496.

Ina, Y. (1998). Estimation of the transition/transversion ratio. Journal of Molecular Evolution, 46: 521–528.

Joanne, A.L. and Angela,M. B. (2005) Tomato SNP discovery by EST mining and resequencing. Molecular Breeding, 16: 343–349

Kanazin, V., Talbert, H., See, D., Decamp, P., Nevo, E., Blake, T. (2002). Discovery and assay of single nucleotide polymorphism in barley (*Hordeum vulgare*). Plant Molecular Biology. 48: 529–537.

Marth, G.T., Korf, I., Yandell, M. D., Yeh, R. T., Zhijie, G., Zakeri,H., Stitziel, N.O., Hillier, L., Kwok,P.Y. and Gish, W. R. (1999). A general approach to single-nucleotide polymorphism discovery. Nature Genetics, 23: 452–456.

Motsi, Moleboheng, C., (2008). Molecular phylogenetics of the genus Rafnia Thunb.(Fabaceae, Crotalarieae).Master degree Dissertation pp. 75.

Newcomb, R.D., Crowhurst, R.N., Gleave, A.P., Rikkerink, E.H.A., Allan, A.C., Beuning, L.L., Bowen, J.H., Gera, E., Jamieson, K.R., Janssen, B.J., Laing, W.A., McArtney, A., Nain, B., Ross, G.S., Snowden, K.C., Souleyre, E.J.F., Walton, E. F. and Yauk, Y.K. (2006). Analyses of Expressed Sequence Tags from apple. Plant Physiology 141: 147–166.

Nickerson, D.A., Tobe, V.O. and Taylor, S. L. (1997). PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Research, 25: 2745–2751.

Picoult-Newberg, L., Idenker, T.E., Pohl, M.G., Taylor, S.L., Donaldson, M.A., Nickerson, D.A. and Boyce-Jacino, M. (1999). Mining SNPs from EST databases. Genome Research. 9: 167-174.

Purvis, A. and Bromham, L. (1997). Estimating the transition/transversion ratio from independent pair wise comparisons with an assumed phylogeny. Journal of Molecular Evolution, 44: 112–119.

Rafalski, A. (2002). Applications of single nucleotide polymorphisms in crop genetics. Current Opinion in Plant Biology, 5: 94–100.

Riju, A., Chandraseker, A. and Arunachalam, V. (2007). Mining for single nucleotide polymorphisms and insertions / deletions in expressed sequence tag libraries of oil palm. Bioinformation, 2(4): 128-131.

Russell, J., Booth, A., Fuller, J., Harrower, B., Hedley, P., Machray, G. and Powell, W. (2004). A comparison of sequence-based poly-morphism and haplotype content in transcribed and anonymous regions of the barley genome. Genome, 47: 389–398.

Somers, D.J., Kirkpatrick, R., Moniwa, M. and Walsh, A. (2003). Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. Genome, 46: 431–437.

Strandberg, A.K.K. and Salter L. A. (2004). A comparison of methods for estimating the transition: transversion ratio from DNA sequences. Molecular Phylogenetics Evolution, 32: 495–503.

Tang, J., Leunissen, J.A.M., Voorrips, R.E., Linden, C.G. and Vosman, B. (2008). HaploSNPer: a web-based allele and SNP detection tool. BMC Genetics, 9: 23

Tenaillon, M.I., Sawkins, M.C., Long, A.D., Gaut, R.L., Doebley, J.F. and Gaut, B.S. (2001). Patterns of DNA sequence polymorphism along chromosome 1 of maize *(Zea mays* ssp. *mays* L.). Proceedings of National Academic Science, USA. 98: 9161–9166.

Varshney, R. K., Graner, A. and Sorrells, M. E. (2005). Genic microsatellite markers in plants: features and applications. Trends in Biotechnology. 23: 48-55.

Varshney, R. K., Beier, U., Khlestkina, E. K., Kota, R., Korzun, V., Graner, A. and Borner, A. (2007). Single nucleotide polymorphisms in rye (*Secale cereale* L.): discovery, frequency, and applications for genome mapping and diversity studies. Theoretical Applied Genetics, 114: 1105-1116.

Wakeley, J. (1994). Substitution rate variation among sites and the estimation of transition bias. Molecular Biology Evolution, 11: 436–442.

Wakeley, J. (1996). The excess of transitions among nucleotide substitutions: New methods of estimating transition bias underscore its significance. TREE, 11: 158–163.

Wu, J., Yamagata, H., Hayashi-Tsugane, M., Hijishita, S., Fujisawa, M., Shibata, M., Ito, Y., Nakamura, M., Sakaguchi, M., Yoshihara, R., Kobayashi, H., Ito, K., Karasawa, W., Yamamoto, M., Saji, S., Katagiri, S., Kanamori, H., Namiki, N., Katayose, Y. Matsumoto, T. and Takuji Sasaki (2004). Composition and structure of the centromeric region of rice chromosome 8. The Plant Cell, 16: 967–976.

Yang, Z. and Yoder, A. (1999). Estimation of the transition/transversion rate bias and species sampling. Journal of Molecular Evolution, 48: 274–283.

ACKNOWLEDGEMENT

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.