**Bioinformation**
*www.bioinformation.net*
*open access*
***Hypothesis***

# Identification of single nucleotide polymorphism in ginger using expressed sequence tags

**Arumugam Chandrasekar[1], Aikkal Riju[1], Kandiyl Sithara[3], Sahadevan Anoop[2] and Santhosh J Eapen[1,*]**

[1]Bioinformatics centre, Indian Institute of spice research, P.B.No.1701, Marikunnu P.O, Calicut -673012, Kerala, India; [2] Ottankulam, Athicode (P.O), Palakkad (Dist), Kerala – 678554, India; [3]Kandiyl House, Makkada (P.O), Kakkodi, Calicut – 673617. Kerala, India; Santhosh J Eapen - Email: sjeapen@spices.res.in; *Corresponding author

Received April 28, 2009; Revised May 10, 2009; Accepted June 08, 2009; Published September 30, 2009

**Abstract:**
Ginger (*Zingiber officinale* Rosc) (Family: Zingiberaceae) is a herbaceous perennial, the rhizomes of which are used as a spice. Ginger is a plant which is well known for its medicinal applications. Recently EST-derived SNPs are a free by-product of the currently expanding EST (Expressed Sequence Tag) databases. The development of high-throughput methods for the detection of SNPs (Single Nucleotide Polymorphism) and small indels (insertion/deletion) has led to a revolution in their use as molecular markers. Available (38139) Ginger EST sequences were mined from dbEST of NCBI. CAP3 program was used to assemble EST sequences into contigs. Candidate SNPs and Indel polymorphisms were detected using the perl script AutoSNP version 1.0 which has used 31905 ESTs for detecting SNPs and Indel sites. We found 64026 SNP sites and 7034 indel polymorphisms with frequency of 0.84 SNPs / 100 bp. Among the three tissues from which the EST libraries had been generated, Rhizomes had high frequency of 1.08 SNPs/indels per 100 bp whereas the leaves had lowest frequency of 0.63 per 100 bp and root is showing relative frequency 0.82/100bp. Transitions and transversion ratio is 0.90. In overall detected SNP, transversion is high when compare to transition. These detected SNPs can be used as markers for genetic studies.

**Availability:**
The results of the present study hosted in our webserver www.spices.res.in/spicesnip

**Keywords:** *Zingiber officinale* Rosc, Expressed Sequence Tag, *in silico*, Ginger, SNPs, Indel

**Background:**
Ginger (*Zingiber officinale*) is a perennial plant in the family Zingiberaceae - its rhizome is commonly used as a cooking spice throughout the world. The ginger plant has a long history of cultivation known to originate in China and then spread to India, Southeast Asia, West Africa, and the Caribbean. India is a leading producer of ginger in the world. Ginger is cultivated in most of the states in India. Kerala and Meghalaya are major ginger growing states in the country. The rhizomes and stems of ginger have assumed significant roles in Chinese, Japanese, and Indian medicine since the 1500s **[1]**.The oleoresin of ginger is often contained in digestive, antitussive, antiflatulent, laxative, and antacid compounds **[2]**. Ginger has a large genome of 23618 Mbp distributed in 2n=22 chromosomes. The phytochemistry and pharmacology of this is well studied but the molecular biological process involved in this is not yet studied. Single-pass sequencing of the 5' and/or 3' ends of randomly selected cDNA clones, is an effective approach to provide genetic information of an organism. These sequences can serve as markers or tags for transcripts, and have been used in the development of SNP markers for reference genetic map and recovery of full-length cDNA and genomic sequences. Expressed sequence tags (ESTs) are also useful for the discovery of novel genes, investigation of genes of unknown function, comparative genomic study, and recognition of exon/intron boundaries. Currently, there are 38139 available ginger sequences in the GenBank, and majority of these sequences are ESTs which had been deposited at NCBI (dbEST) http://www.ncbi.nlm.nih.gov/dbEST/. The lack of sequence information has limited the progress of gene discovery and characterization, global transcript profiling, probe design for development of gene arrays, and generation of molecular markers for Ginger. In this study, we have categorized 38139 ESTs in to three tissue libraries leaves 13274, rhizomes 12763 and roots 12092 ESTs**.** The availability of these EST sequences will allow comparative genomic studies between ginger and other monocotyledonous and dicotyledonous plants, development of molecular markers for the establishment of reference genetic map, design and construction of cDNA microarray for global gene expression profiling. Single nucleotide polymorphisms (SNPs) are a second class of genetic markers that can be mined from sequence data and are useful for characterizing allelic variation, genome-wide mapping, and as a tool for marker-assisted selection. In the field of human genetics, SNPs are a major focus of efforts to increase the efficiency of mapping **[3,4,5,6]** and are already being used for detection and mapping of a variety of diseases **[7-9]**. In many crop plants, SNPs are present with sufficient frequency to offer an alternative for genetic mapping and marker-assisted selection. Although SNPs can be identified by sequencing selected DNA fragments, a practical limitation to this approach for ginger follows from the fact that the sequencing error rate is often higher than the polymorphism rate. The cost of SNP discovery through sequencing amplified fragments is therefore high even with reductions in the cost of sequencing. The objectives of the research described in this paper were to assess the potential of existing public databases for the discovery of single nucleotide polymorphisms. We have mined updated EST tissue libraries of *zingiber officinale* for this analysis to find the SNP / Indel polymorphisms. SNP detecting perl scripts AutoSNP version 1.0 is used indentify the SNP / Indel polymorphisms, DNA substitution like Transversion vs Transition and Indel **[10]**. There are some other SNP detecting software such as SEAN **[11]** PolyPhred **[12]** PolyBayes **[13]** TRACE_DIFF **[14]** and HarvEST (http://harvest.ucr.edu) but AutoSNP provides user friendly approach and interpretable results as html file. 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. We have used three tissue libraries **[15,16]** of *Zingiber officinale*.

**Methodology:**
EST database of NCBI (dbEST release 092509) contains 38139 *Zingiber officinale* Express sequence tag data. We have mined 38139 EST sequences consist of three tissue libraries of leaves 13274 (DV544275-ES560515), rhizomes 12763 (DY350707-DY363469) and roots 12092 (DY363470-DY375561). CAP3 program is used to assemble the EST sequence in to contigs. The SNP detection tool AutoSNP version.1.0 was used to find the candidate SNPs from these libraries. AutoSNP required input as ace or fasta format. But the perl script edited manually to analyse fasta or ace format. Sequence



ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(3): 119-122 (2009)

assembly program CAP3 is integrated in AUTOSNP to make fasta files in to contigs (http://bioweb.pasteur.fr/seqanal/interfaces/cap3.html) **[17].** The DNA substitution like transition (Ts) versus transversion (Tv) ratio of all the libraries in Ginger genome was also calculated.

**Discussion:**

In this study it is discovered that total of 64026 SNP sites and 7034 indel polymorphisms in 38139 ESTs analyzed with an average frequency of 0.84 SNPs / 100 bp. Results of the tissue wise SNP and indel discovery are listed in **Table 1** (**see supplementary material**) and **Figure 1**. In Ginger leaves tissue libraries showing high indels 1983 while comparing other tissues. Rhizome tissues showing the high SNP frequency 1SNP in 100bp. In Ginger a total of 27083 transitions, 29909 transversions and 7034 indels were found while analysis. But we found in tissue wise manner, rhizome transitions are high in number 13433. Rhizome tissues having more SNPs than others. Rhizome part is more expressed than other tissues. While discovering all SNP with DNA substitution overall transitions and transversions ratio is 0.90. When compared to ginger with others, the studies on the occurrence and nature of SNPs are beginning to receive considerable attention, particularly Arabidopsis where over 37,000 SNPs have been identified through the comparison of two accessions **[18]**. It has been reported in maize that there occurs a frequency of one non-coding SNP per 31 bp and 1 coding SNP per 124bp in 18 maize genes assayed in 36 inbred lines **[19]**. Moreover the recent evidence has indicated that SNPs appear to be even more abundant in plant systems than in the human genome. Germano and Klein **[20]** identified five SNPs in 1 kb of cDNA of *Picea rubens* and *Picea mariana*, and also discovered SNPs in the chloroplasts of these species. Recently, in soyabean (*Glycine max*), two SNPs found approximately 400 bp **[21]**. In maize (*Zea mays*), SNP has been detected even more frequently, with one SNP approximately every 48 bp and every 130 bp in 3' untranslated regions and coding regions, respectively **[22,23]**. The SNP analysis on Apple (*Malus domestica*) ESTs the Bi-allelic SNPs were on an average of every 706 bp **[24]** and the study in Maize ESTs **[25]** also showed the relative increase of over transversion and transition sites. This *in silico* analysis will help ginger researchers about the single nucleotide polymorphism related study and nucleotide substitution in this important crop.

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 **[26, 27]** the potential for plant breeding to be impacted by new technology has never been greater. In ginger, sequencing projects offer a potential solution to the scarcity of markers that can be used in elite breeding populations. Of special interest is the ability to discover DNA polymorphisms by mining sequence data **[28,29]**. The frequency of single nucleotide polymorphisms that we detected is considerably lower than reported for maize, wheat, barley, and soybean. Not surprisingly it is also lower than the one SNP per approximately 100 bases that was detected in some of tissue libraries **[30]**. There was a relative increase in the proportion of transition (6805) over transversion (7258) in Ginger ESTs except in leaves libraries (**Figure 1**). C / T transition was found to be high in ginger (**Table 1 in supplementary material**). High frequency of the C to T mutation is usually seen due to methylation. We also used the Shannon information index to analyze the proportion of ten possible types of SNP/indels. ESTs from tissues of root showed highest values of indices (0.164) whereas leaves had the least value (0.150) and rhizome showed relatively increased value (0.152). Our study on higher number and Shannon index of SNP/indel sites in root tissue than other tissues also gives the additional information about in genomic variation in genes expressed specifically in root tissue. Ratio of transition to transversion (Ts/Tv) was very useful to compare the genotypes of hepatitis virus C and also differences among the mitochondrial genomes of animals. Our study gives a method, which compares the ten possible types of SNP/ indels in a single index. The results of detected SNPs were accessed through online at www.spices.res.in/spicesnip/

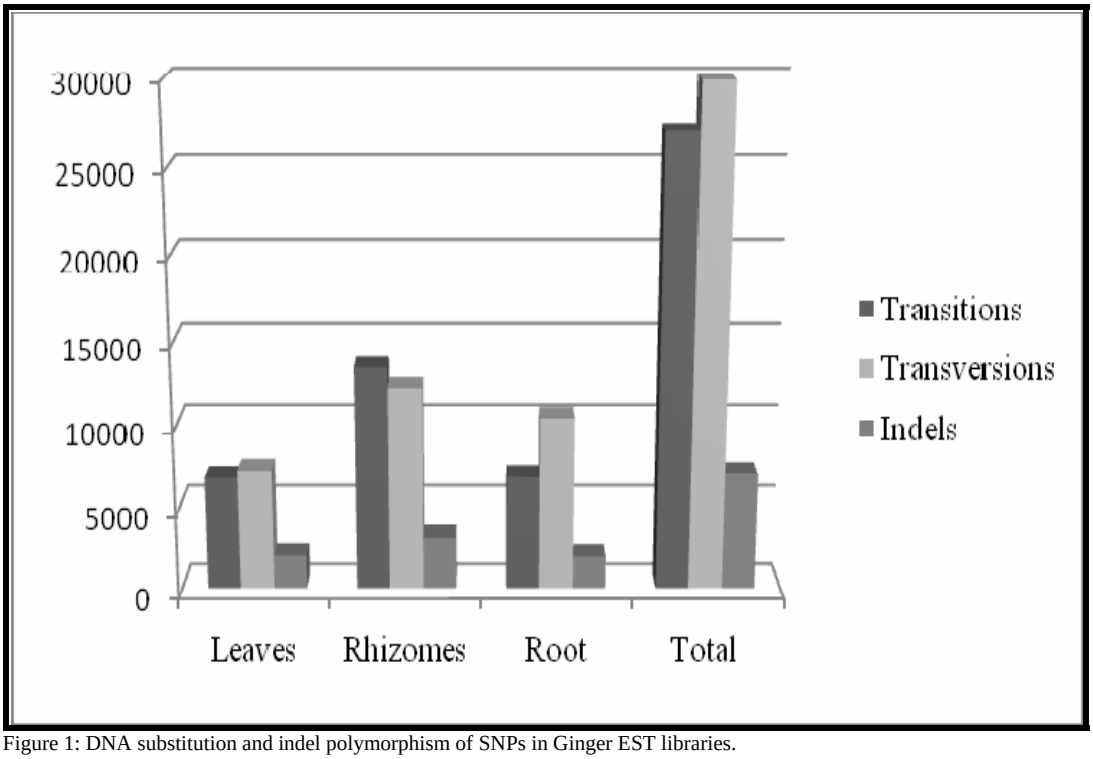


Figure 1: DNA substitution and indel polymorphism of SNPs in Ginger EST libraries.



ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(3): 119-122 (2009)

**Conclusion:**
In total, we have identified over 64026 candidate SNP polymorphisms with frequency of 0.84 SNPs/100bp in Ginger EST sequence data, along with two measures of confidence for each predicted polymorphism. Segregation of these SNPs with haplotype along with validation demonstrates that candidate SNPs with high redundancy and co-segregation confidence scores are likely to represent true SNPs. The transition to transversion ratio and indel size frequencies correspond to those observed by the analysis methods of SNP discovery and suggest that the majority of predicted SNPs and indel identified using this approach represent true genetic variation in ginger. Overall transversion is high because ginger is vegetative propagated through rhizome. This *in silico* analysis on ginger shows the potential SNP markers for use in ginger breeding and the online information we created would help to designing new primers and develop more markers and to saturate the linkage maps.

**Acknowledgement:**
We are grateful to Department of Biotechnology Government of India for the financial support.

**References:**
**[1]** JW Purseglove *et al., Spice* 2 (1981)
**[2]** VS Govindarajan *et al*., *Crit Rev Food Sci Nutr*. **17**:1 (1982) [PMID: 7049579]
**[3]** International SNP Map Working Group *Nature* **409**: 928 (2001) [PMID: 11237013]
**[4]** J Aerts *et al., Human Mutation* **20**:162 (2002) [PMID: 12203988]
**[5]** S Balasubramanian *et al., Pharmacogenomics* **3**:393 (2002) [PMID: 12052146]
**[6]** LYY Chen *et al*., *Genome Research* **12**:1106 (2002) [PMID: 12097348]
**[7]** BAJ Verhage *et al., Int J Cancer.* **100**: 683 (2002) [PMID: 12209606]
**[8]** Y Sugimoto *et al*., *Journal of Viral Hepatitis* **9**:377 (2002) [PMID: 12225333]
**[9]** K Margiotti *et al., The Prostate* **53**:65 (2002) [PMID: 12210481]
**[10]** G Barker *et al*., *Bioinformatics* **19**:421 (2003) [PMID: 12584131]
**[11]** D Huntley *et al., Bioinformatics* **22**:495 (2006) [PMID: 16357032]
**[12]** DA Nickerson *et al., Nucleic Acids Res*. **25**:2745 (1997) [PMID: 9207020]
**[13]** GT Marth *et al*., *Nat. Genet*. **23**:452 (1999) [PMID: 10581034]
**[14]** JK Bonfield *et al., Nucleic Acids Res.* **26**:3404 (1998) [PMID: 9649626].
**[15]** S Jouannic *et al., FEBS*. **579**: 2709 (2005) [PMID: 15862313]
**[16]** CL Ho *et al., BMC Genomics* **8**:381 (2007) [PMID: 17953740]
**[17]** X Huang *et al., Genom Res*. **9**:868 (1999) [PMID: 10508846]
**[18]** G Jander *et al., Plant Physiol*. **129**: 440 (2002) [PMID: 12068090]
**[19]** A Ching *et al*., *BMC Genet*. **3**:19 (2002) [PMID: 12366868]
**[20]** J Germano *et al*., *Theor Appl Genet*. **99**:37 (1999)
**[21]** VH Coryell *et al*., *Theor Appl Genet*. **101**:1291 (1999)
**[22]** MI Tenaillon *et al*., *Proc Natl Acad Sci*. **98**:9161 (2001) [PMID: 12454083]
**[23]** A Rafalski *et al., Curr Opin Plant Biol*. **5**:94 (2002) [PMID: 11856602]
**[24]** RD Newcomb, *et al., Plant Physiology* **141**:147 (2006) [PMID: 16531485]
**[25]** J Batley *et al*., *Plant Physiology* **132**:84 (2003) [PMID: 12746514]
**[26]** SA Goff *et al*., *Science* **296**: 92 (2002) [PMID: 11935018]
**[27]** J Yu *et al., Science* **296**:79 (2002) [PMID: 11935017]
**[28]** MJM Smulders *et al*., *Theor Appl Genet*. **94**:264 (1997)
**[29]** GMM Bredemeijer *et al., Theor Appl Genet*. **105**:1019 (2002) [PMID: 12582929]
**[30]** S Suliman-Pollatschek *et al*., *Cellular and Molecular Biology Letters* **7**:583 (2002) [PMID: 12378264]

**Edited by P. Kangueane**
**Citation: Chandrasekar et al.,** Bioinformation 4(3): 119-122 (2009)




ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(3): 119-122 (2009)

**Bioinformation**
*www.bioinformation.net*
*open access*
***Hypothesis***

## Supplementary material

**Table 1:** Discovered SNP results of Zingiber officinale.

| Tissue Name | Sequences analysed | No of contigs | Consensus size | SNPs detected | Transitions (Ts) | Transversions (Tv) | Indels | Ts / Tv | SNP frequency |
|---|---|---|---|---|---|---|---|---|---|
| Leaves | 11307 | 2660 | 2555383 | 16046 | 6805 | 7258 | 1983 | 0.94 | 0.63/100 bp |
| Rhizomes | 10123 | 2921 | 2668324 | 28809 | 13433 | 12249 | 3127 | 1.10 | 1.08/100 bp |
| Root | 10475 | 2972 | 2332434 | 19171 | 6845 | 10402 | 1924 | 0.66 | 0.82/100 bp |
| **Total** | 31905 | 8553 | 7556141 | 64026 | 27083 | 29909 | 7034 | 0.90 | 0.84/100 bp |

**A total of 31905 consensus EST sequence are used to predict the SNP site from Ginger species, which made 8553 cluster groups and found 64026 SNP/indel sites.**

**Table 2:** Summary of SNPs and indels detected in the Ginger EST libraries

| Results | Leaves | Rhizomes | Root | Total |
|---|---|---|---|---|
| Total No. of ESTs | 13274 | 12763 | 12092 | 38129 |
| Total sequences analysed | 11307 | 10123 | 10475 | 31905 |
| No. of contigs | 2660 | 2921 | 2972 | 8553 |
| Total SNPs detected | 16046 | 28809 | 19171 | 64026 |
| Total consensus size (bp) | 2555383 | 2668324 | 2332434 | 7556141 |
| Frequency of SNP per 100bp | 0.63 | 1.08 | 0.82 | 0.84 |
| Transitions | | | | |
| G/A | 3595 | 6901 | 3278 | 13774 |
| C/T | 3210 | 6532 | 3567 | 13309 |
| Tranversions | | | | |
| A/C | 1822 | 3200 | 2617 | 7639 |
| A/T | 1678 | 2550 | 2704 | 6932 |
| G/T | 1917 | 3065 | 2553 | 7535 |
| G/C | 1841 | 3434 | 2528 | 7803 |
| Indel | | | | |
| A | 564 | 818 | 623 | 2005 |
| G | 504 | 772 | 547 | 1823 |
| T | 433 | 743 | 361 | 1537 |
| C | 482 | 794 | 393 | 1669 |
| Shannon Index | 0.150 | 0.152 | 0.164 | 0.052 |

$H' = \sum_{i=1}^{n} piLog_2Pi$

Where n is the total number of SNP/indel states (10) pi= proportion of ESTs in the $i^{th}$ type of SNP/indel state. The calculated value is divided by the $log_2 10$ to get uniformity.



ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(3): 119-122 (2009)