BIOINFORMATION
Discovery at the interface of physical and biological sciences
open access www.bioinformation.net Hypothesis
Volume 11(2)

High quality SNPs/Indels mining and characterization in ginger from ESTs data base

Mahendra Gaur, Aradhana Das & Enketeswara Subudhi*

Centre of Biotechnology, Siksha 'O' Anushandhan University, Kalinga Nagar, Ghatikia, Bhubaneswar-751003, Odisha; Enketeswara Subudhi – Email: enketeswarasubudhi@soauniversity.ac.in; Phone: +91-9861075829; *Corresponding author

Received January 07, 2015; Accepted January 29, 2015; Published February 28, 2015

Abstract:
Ginger (*Zingiber officinale* Rosc.) is an important herb of the family Zingiberaceae. It is accepted as a universal cure for a multitude of diseases in Indian systems of medicine and its rhizomes are equally popular as a spice ingredient throughout Asia. SNPs, the definitive genetic markers, representing the finest resolution of a DNA sequence, are abundantly found in populations having a lower rate of mutation and are used for genomic analysis. The public ESTs sequences mostly lack quality files, making high quality SNPs detection more difficult since it is exclusively based on sequence comparisons. In the present study, current dbESTs of NCBI was mined and 38115 ginger ESTs sequences were obtained and assembled into contigs using CAP3 program. In this analysis, recent software tool QualitySNP was used to detect 11523 potential SNPs sites, 8810 high quality SNPs and 1008 indels polymorphisms with a frequency of 1.61 SNPs / 10 kbp. Of ESTs libraries generated from three ginger tissues together, rhizomes had a frequency of 0.32 SNPs and 0.03 indels per 10 kbp whereas the leaves had a frequency of 2.51 SNPs and 0.23 indels per 10 kbp and root is showing relative frequency of 0.76/10 kbp SNPs and 0.02/10 kbp indels. The present analysis provides additional information about the tissue wise presence of haplotypes (222), distribution of high quality exonic (2355) and intronic (6455) SNPs and information about singletons (7538) in addition to contigs transitions and transversions ratio (0.57). Among all tissue detected SNPs, transversions number is higher in comparison to the number of transitions. Quality SNPs detected in this work can be used as markers for further ginger genetic experiments.

Keywords: *Zingiber officinale*, Ginger, QualitySNP, ESTs, *in silico,* Indels.

Background:

Ginger (*Zingiber officinale* Rosc.) is a valued crop of family *Zingiberaceae,* rhizomes of which are used in medicine as well as in spice. India is a leading ginger producer, accounting for about 30% of the global share and during 2012-13 it produced 7.45 lakh tonnes from an area of 1.57 lakh hectares out of total global production of 20.95 Lakh tons **[1].** Ginger is cultivated as a spice crop in many states of India. Out of the country's total ginger production, Kerala, Meghalaya, Orissa, Gujarat, Assam, and Arunachal Pradesh together contribute 65 per cent. There is an international market for Indian ginger at current selling price of around $2800-$2850 per ton **[2].** Reports on evaluation of existing rich diversity in indian ginger germplasm are based on phenotypic and phytochemical characteristics which exhibit plasticity and sensitivity to environmental conditions **[3].**

Molecular characterization using PCR based markers are established as robust, reproducible and reliable when compared to morphological and phytochemical markers. Use of these markers reduces the cost, time and labour during analysis **[4, 5].** Markers ISSR, SSR, IRAP etc. has been reported for assessing genetic variation in Zingiber or other members of the Zingiberaceae family **[6-8]** that exclude reports on the discovery and use of quality SNPs to understand the genetic analysis of ginger. Single nucleotide polymorphism is basically single-base allelic variation between two haplotype sequences or between any of the homologous pair of chromosome. SNPs are abundantly found among variations prevalent in genomic DNA both in coding and non-coding regions **[9].** These are responsible for various genetic traits, conserved through evolution and compatible for next generation high-throughput genotyping. However, sequencing of selected DNA fragments for SNPs identification has been subjected to limitations like; higher rate of sequencing error and intensive cost incurred during sequencing the fragments amplified. The alternate cost effective option for SNPs discovery from public database is by using the most extensive resource of ESTs data (dbESTs) hosted by The National Center for Biotechnology Information (NCBI) **[10].** ESTs database has been commonly used for discovery of new genes, exon-intron structure verification, cDNA array construction and gene mapping **[11].** These polymorphisms obtained from ESTs represent numerous functional genes which controls many genetic traits **[12-15].** Many bioinformatics tools, programs as well as pipelines are developed for mining of SNPs by using several input and/or output formats, computational algorithms, filtration and evaluation strategies for getting quality SNPs. There are many programs and pipelines for detection of SNPs viz; SEAN **[16]** PolyPhred **[17]** PolyBayes **[18]** TRACE_DIFF **[19]**, HarvEST **[20]**, AutoSNP **[21]**, QualitySNP **[15]** QualitySNPng **[22].** QualitySNP has three filtering system to eliminate unreliable variations and to handle typical sequencing errors in absence of sequenced reference genome. Extracted SNPs information has been useful for QTL mapping and genome-wide association studies **[9, 23].** Since its publication, the QualitySNP has been successfully used for the identification of SNPs markers in dozens of projects viz; in crop plants **[24]**, zebra finch **[25]**, water fleas **[26],** snakes **[27]** and scallops **[28].** The haplotype-based strategy can make full use of redundancy in sequences by re-clustering them, so that the influence of sequencing errors is avoided and poor quality sequences are removed. QualitySNP pipeline identifies paralogs and quality SNPs on heterozygous diploid as well as polyploid species. Therefore, in the present attempt, an effort is taken to utilize the existing updated ESTs tissue libraries of *Zingiber officinale* to find the SNP/Indels polymorphisms using 38115 ESTs and categorized into three tissue libraries leaves (13282), rhizomes (12763) and roots (12092) ESTs **[29, 30]**. High quality SNPs detecting tool QualitySNP is used to identify the high quality exonic and intronic polymorphisms, haplotypes and DNA substitutions like transitions, transversions and Indels.

Methodology:

Data Mining:

A total of 38115 *Z. officinale* ESTs sequence (13282 Leaves ESTs (DV544275.1-ES560515.1), 12763 Rhizomes ESTs (DY363350.1-DY363469.1) and 12092 Roots ESTs (DY375442.1-DY375561.1)) were retrieved from the dbEST **[31]** hosted by GeneBank (NCBI) using the keywords "Zingiber officinale" and grouped into the respective tissues library.

Sequence Pre-Processing and Clustering:

To get high quality ESTs, Poly-A/T tails and unexpected vector sequences were filtered and trimmed using Trimest **[32]** online program of EMBOSS suite and SeqClean software **[33]** with reference of the UniVec database of NCBI. ESTs having length ≥ 50 bp are traced out if any found to increase assembly quality. The ESTs of high quality were then assembled into contigs using CAP3 **[34]** program at 90% identity. Tissue-wise ESTs assemblies were conducted to reduce redundancy.

High Quality SNPs Discovery:

The Linux based command line program QualitySNP pipeline used for extraction of SNPs **[35].** QualitySNP detected the haplotypes present in the contigs through ESTs re-clustering and discrepancies in nucleotide were extracted between identified haplotypes of a contig. These are considered as potential SNPs (pSNP). Basing on confidence scores of SNPs and allele, quality SNPs (qSNP) were identified **[15].** The nucleotide discrepancies percentage is obtained from qSNP/pSNP ratios using the QualitySNP algorithm.

Prioritizing High Quality SNPs:

Contigs from all assemblies of ginger were processed through ORF finder program **[36]** for locating possible positive or negative open reading frame. From whole ESTs assembly only 101 contigs contains ORF with AUG as a start codon. Out of 8810 qSNP only 2355 qSNP were located in exonic/orf region while 6455 qSNP in non-coding regions. Distribution of qSNP into exonic and intronic region is shown in **Figure 1(a)** & **1(b).**

Discussion:

We obtained 6323 contigs and 17421 singletons from 38115 ESTs from all tissues together accounting to 5455657 consensus sizes of 23708402 base pairs. In this study, a total of 11523 potential SNPs and 8810 high quality SNPs sites and 1008 indels polymorphisms are discovered from 38115 numbers of analysed ESTs with an average frequency of 1.61 SNPs / 10 kbp and 0.18 indels/10 kbp. The size of the contigs varied from 100 to 2954 bp, with an average length of 862 bp. An overview of the tissue wise assembly pertaining to contigs, singletons, indels, and SNPs detection and other parameters are depicted in **Table 1- 4 (see supplementary material).** Ginger tissue wise prevalence of exonic SNPs and intronic SNPs as per their occurrence in coding and non-coding regions is detailed in **Figure 1(a)** and **1(b)** which shows highest in the leaf tissues. When compared tissue wise libraries in ginger, leaves tissue are showing highest SNPs (4895), indels (452), haplotypes (121), transitions (1572) and transversions (2871) in comparison to other tissues. Rhizome tissues are showing the highest ratio of base pairs per SNP (3145), per Indels (28628) and highest ratio of high quality to potential SNPs (0.82). SNPs substitutions obtained from all tissues has a ratio of transitions to transversions at 0.57. Details of DNA substitutions parameters for other tissues are found in **Figure 1 (c) & Table 1**. As compared to the SNPs analysis in ginger, total 31815 potential SNPs, 16772 high quality SNPs and 1815 indels were mined out of total ESTs 83565 in potato **[15]**, 37344 SNPs in Arabidopsis **[37]**, in the maize prevalence of SNPs in non-coding and coding were found to be 1 per 31 bp and 1 per 124 bp respectively **[38].** The average SNPs occurrence in Apple ESTs was found to be one in every 706 bp **[39].** SNPs frequency is higher, ie. 1.61 SNPs per 10 kbp in certain of the genomes like ginger. Similar discoveries on SNPs were also reported in *Arabidopsis* ecotypes viz; one SNP every 3.3 kb in *Landsberg erecta* and one SNP every 6.1 kb in *Columbia* **[40].** One SNP per 20 bp is reported in bread wheat between genes from the A, B and C genomes **[41]** and in maize the SNPs frequency is found to be one Indel /160 bp and one SNP/70 bp **[42].** QualitySNP detected high frequency of transitions in ginger in the present analysis, which corroborates previous SNPs discovery reports **[42]**. High frequency of C to T mutation in this ESTs derived SNPs in ginger may be due to methylation **[42]**. The reported abundance of A/T ratio as well as its reverse complement remains unexplained. Out of total SNPs substitutions detected in this attempt, transversions of all tissues is/are comparatively higher than the transitions which corroborates previous report on the ginger SNPs detection and assigns the reason of ginger is being vegetative propagated crop through rhizome **[43].**

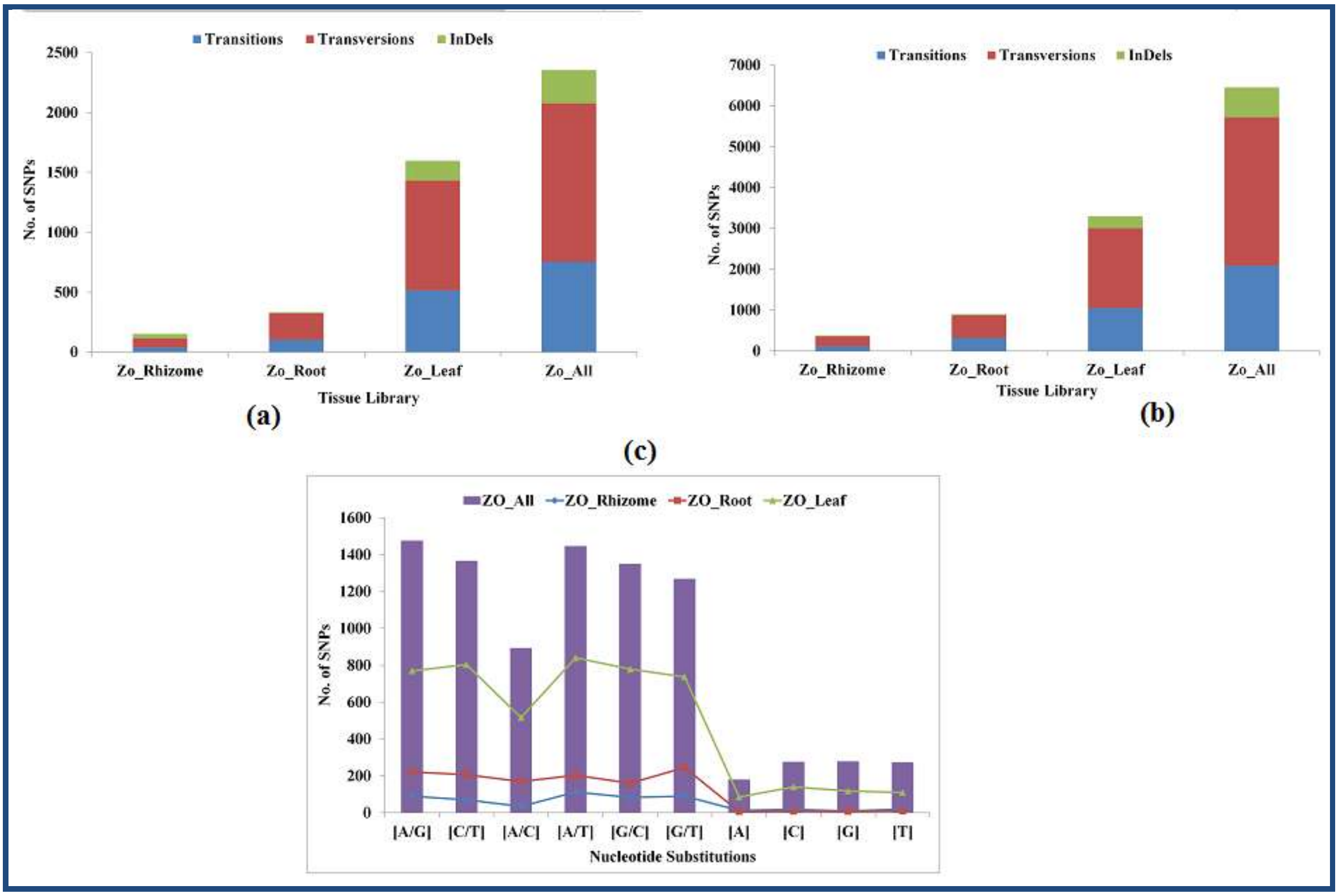

Figure 1: Distribution of High Quality SNPs **(a)** Exonic **(b)** Intronic **(c)** Nucleotide Substitutions

Our study for the first time provides information about high quality SNPs (8810) and 1008 indels polymorphisms in addition to information about potential SNPs (11523) because of the use of triple filtration based on stringent analysis of ESTs database using QualitySNP when compared with the pearl script AutoSNP version 1.0 based analysis of ginger ESTs with higher number of potential 64026 SNPs sites and 7034 indels polymorphisms **[43].** The present analysis provides additional information about tissues wise presence of haplotypes (222) distribution of high quality exonic (2355) and intronic (6455) SNPs and information about singletons (17421) in addition to contigs **(Figure 1(a), 1(b), 1(c)** and **Table 1-4)**. Different software's used for SNPs mining during *in silico* study using the same ESTs database in Sea bass has been reported which exhibits level of stringency implied, and the difference in output in quality SNPs during analysis **[44]** justifies the use of most recent and efficient QualitySNP software integrated with better filtration strategy. We are able to discover the reduced number of quality SNPs in addition to potential SNPs, which may help geneticists and breeders working in/on cultivar identification, germplasm conservation in ginger to a greater extent.

Conclusion:

We have detected 11523 potential SNPs sites with 8810 high quality, and 1008 indels polymorphisms as well as 1.61 SNPs / 1000 bp frequency using recent software QualitySNP. Of ESTs libraries collected from 3 tissues, rhizomes had a frequency of 0.32 SNPs and 0.03 indels per 1000 bp, but the leaves had 2.51 SNPs and 0.23 indels per 1000 bp and root is showing relative frequency 0.76/1000 bp SNPs and 0.02/1000bp indels. The present analysis provides additional information about tissues wise presence of haplotypes (222) distribution of high quality exonic (2355) and intronic (6455) SNPs and information about singletons (7538) in addition to contigs transitions and transversions ratio was 0.57. Among all tissue detected SNPs, transversions number is higher in comparison to the transition. The quality SNPs detected can be used as potential markers for ginger genetic studies.

Acknowledgement:

We are thankful to Prof. Manoj Ranjan Nayak, President, Siksha 'O' Anusandhan University for providing necessary infrastructure and encouragement throughout.

References:

[1] http://agropedia.iitk.ac.in/content/ginger-price-forecast-store-and-sell-ginger-after-may-2014
[2] http://www.factfish.com/statistic/ginger%2C%20area%20harvested
[3] Das A *et al. JCHP Special Issue.* 2014 **2**: 16
[4] Zietkiewicz E *et al. Genomics* 1994 **20**: 17618 [PMID: 8020964]
[5] Kalendar R *et al. Theor Appl Genet.* 1999 **98:** 704 [PMID: 17406494]
[6] Jatoi SA *et al. Breed Sci*. 2008 **58:** 261
[7] Kizhakkayil J & Sasikumar B, *Scientia Horticulturae* 2010 **125:** 73
[8] Pandotra P *et al Scientia Horticulturae* 2013 **160**: 283
[9] Brookes AJ *Gene* 1999 **234**: 177 [PMID: 10395891]
[10] Boguski MS *et al. Nat Genet.* 1993 **4:** 332 [PMID: 8401577]
[11] Lee B *et al. Nucleic Acids Res.* 2007 **35**: W159 [PMID: 17526512]
[12] Garg K *et al Genome Res.* 1999 **9:** 1087 [PMID: 10568748]
[13] Picoult-Newberg L *et al. Genome Res* 1999 **9:** 167 [PMID: 10022981]
[14] Batley J *et al. Plant Physiol* 2003 **132**: 84 [PMID: 12746514]
[15] Tang J *et al. BMC Bioinformatics* 2006 **7:** 438 [PMID: 17029635]
[16] Huntley D *et al. Bioinformatics* 2006 **22:** 495 [PMID: 16357032]
[17] Nickerson DA *et al. Nucleic Acids Res.* 1997 **25:** 2745 [PMID: 9207020]
[18] Marth GT *et al. Nat Genet.* 1999 **23**: 452 [PMID: 10581034]
[19] Bonfield JK *et al. Nucleic Acids Res.* 1998 **26:** 3404 [PMID: 9649626]
[20] Close TJ *et al. Methods Mol Biol.* 2007 **406:** 161 [PMID: 18287692]
[21] Barker G *et al. Bioinformatics* 2003 **19:** 421 [PMID: 12584131]
[22] Nijveen H *et al. Nucleic Acids Res.* 2013 **41**: W587 [PMID: 23632165]
[23] Davey JW *et al. Nat Rev Genet.* 2011 **12**: 499 [PMID: 21681211]
[24] Anithakumari AM *et al. Mol. Breed.* 2010 **26**: 65 [PMID: 20502512]
[25] Stapley J *et al. Genetics* 2008 **179:** 651 [PMID: 18493078]
[26] Orsini L *et al. BMC Genomics* 2011 **12**: 309 [PMID: 21668940]
[27] Cardoso KC *et al. BMC Genomics* 2010 **11:** 605 [PMID: 20977763]
[28] Hou R *et al. PLoS One* 2011 **6**: e21560 [PMID: 21720557]
[29] Jouannic S *et al. FEBS Lett.* 2005 **579:** 2709 [PMID: 15862313]
[30] Ho CL *et al. BMC Genomics* 2007 **8:** 381 [PMID: 17953740]
[31] *http://www.ncbi.nlm.nih.gov/dbEST/index.html*
[32] Rice P *et al. Trends Genet.* 2000 **16:** 276 [PMID: 10827456]
[33] *http://sourceforge.net/projects/seqclean/files/*
[34] Huang X & Madan A, *Genome Res.* 1999 **9:** 868 [PMID: 10508846]
[35] Chunxian C & Gmitter FG, *BMC Genomics* 2013 **14:** 746 [PMID: 24175923]
[36] *http://www.ncbi.nlm.nih.gov/gorf/gorf.html*
[37] Jander G *et al. Plant Physiol.* 2002 **129:** 440 [PMID: 12068090]
[38] Ching A *et al. BMC Genet.* 2002 **3:** 19 [PMID: 12366868]
[39] Newcomb RD *et al. Plant Physiol.* 2006 **141:** 147 [PMID: 16531485]
[40] *www.arabidopsis.org*
[41] Wolters P *et al. Plant and Animal Genome VIII Conference, San Diego* 2000 9-12
[42] Bhattramakki D *et al. Plant Mol. Biol.* 2002 **48:** 539 [PMID: 12004893]
[43] Chandrasekar A *et al. Bioinformation* 2009 **4**: 119 [PMID: 20198184]
[44] Souche EL *et al. Journal of Integrative Bioinformatics* 2007 **4**: 73

Edited by P Kangueane
Citation: **Gaur *et al.*** Bioinformation 11(2): 085-089 (2015)

Supplementary material:

Table 1: Distribution of High Quality Exonic SNPs in Ginger

Tissue	Transitions (t_s)	Transversions (t_v)	InDels	Total
Zo_Rhizome	39	77	37	153
Zo_Root	102	224	6	332
Zo_Leaf	516	917	165	1598
Zo_All	755	1322	278	2355

Table 2: Distribution of High Quality Intronic SNPs in Ginger

Tissue	Transitions (t_s)	Transversions (t_v)	InDels	Total
Zo_Rhizome	120	242	22	384
Zo_Root	325	553	24	902
Zo_Leaf	1056	1954	287	3297
Zo_All	2088	3637	730	6455

Table 3: Distribution of Nucleotide Substitutions in Ginger

Variation	[A/G]	[C/T]	[A/C]	[A/T]	[G/C]	[G/T]	[A]	[C]	[G]	[T]
ZO_Rhizome	89	70	34	112	83	90	12	18	9	20
ZO_Root	221	206	169	203	160	245	6	8	7	9
ZO_Leaf	769	803	516	840	779	736	85	140	118	109
ZO_All	1476	1367	894	1446	1350	1269	180	276	279	273

Table 4: Extract of SNPs and indels discovery in the Ginger ESTs Sequences

	ZO_All	ZO_Leaf	ZO_Rhizome	ZO_Root
Total ESTs	38115	13282	12763	12069
Total Contings	6323	2189	2009	2122
Total Consensus Size (bp)	5455657	1946842	1689089	1621996
Average Contig Length	862	889	840	764
Average ESTs per Contig	3.27	3.47	2.60	2.71
Maximum Contig Size	2954	2954	2395	2430
Minimum Contig Size	100	100	178	153
Contings with 2 ESTs	3793	1380	1465	1449
Contings with >2 ESTs	2530	809	544	673
Singletons	17421	5673	7538	6315
Potential SNPs (pSNP)	11523	6417	658	1517
High Quality SNPs (qSNP)	8810	4895	537	1234
Haplotypes	222	121	21	36
Ratio of qSNP to pSNP	0.76	0.76	0.82	0.81
No of SNPs per 10 kbp	1.61	2.51	0.32	0.76
No of SNPs per Contig	1.39	2.23	0.27	0.58
No of bp per SNP	619	397	3145	1314
Transitions (t_s)	2843	1572	159	427
Transversions (t_v)	4959	2871	319	777
Ratio of Transitions to Transversion	0.57	0.54	0.50	0.55
InDels	1008	452	59	30
InDels per 10 kbp	0.18	0.23	0.03	0.02
InDels per Contig	0.15	0.21	0.03	0.01
Base pairs per InDels	5412	4307	28628	54066

