Mining SSR and SNP/Indel Sites in Expressed Sequence Tag Libraries of *Radopholus similis*

Riju A[*], Lakshmi PDK[$], Nima PL[#], Reena N[€] and Eapen SJ[¥]
Bioinformatics Centre, Indian Institute of Spices Research, Calicut- 12
Telephone. No. +914952731566
Email id: [*] riju.bioinfo@gmail.com; [$] pri2darshini@gmail.com; [#] nimanov1@gmail.com; [€] reena_narayanan@yahoo.co.in; [¥] sjeapen@spices.res.in

ABSTRACT
The objective of this study is to explore the single sequence repeats (SSRs) and single nucleotide polymorphims (SNPs) in expressed sequence tags (ESTs) of *Radopholus similis*. We retrieved 7380 EST sequences consisting different tissues/condition libraries from dbEST of National Centre for Biotechnology Information (NCBI). A total of 1449 SSRs were detected by MISA perl script. Hexa-nucleotide repeats (836 nos.) followed by mononucleotide repeats (207 nos.) were found to be more abundant than other types of repeats. Putative SNP/Indels were found out with the help of AutoSNP. As many as 1038 SNPs and 108 small indels (insertion/deletion) were found with a density of one SNP/191 bp and one indel/1.8 kbp. Candidate SNPs were categorized according to nucleotide substitution as either transition (C↔T or G↔A) or transversion (C↔G, A↔T, C↔A or T↔G). We observed a higher number of transversions type substitution (537) than transitions (501). However considering the individual substitutions, G↔A (281) and C↔T (220) were found to be predominant than purine to pyrimidine base substitutions. Since the SSR and SNP markers are invaluable tools for genetic analysis, the identified SSRs and SNPs of *R. similis* could be used in diversity analysis, genetic trait mapping, association studies and marker assisted selection.

Categories and Subject Descriptors
A.0 [General]: Conference proceedings

General Terms
Experimentation

Keywords
Single Nucleotide Polymorphism, Simple Sequence Repeats, Expressed Sequence Tags, Molecular Marker, Insertion and Deletion, Mutation, Transition, Transversion.

1. INTRODUCTION
Aim of the present study is to mine the expressed sequence tags (ESTs) of *Radopholus similis* to detect Simple Sequence Repeats (SSRs), Single Nucleotide Polymorphisms (SNPs) and Insertions and Deletions (Indels). Burrowing nematode or banana root nematode (*Radopholus similis*) is an important parasite of fruits, vegetables and other crops of tropical and sub tropical regions. It is especially important in bananas, citrus and black pepper, but will also attack coconut, avocado, coffee, sugarcane and different types of grasses and ornamentals. It is a migratory endoparasite of roots. The nematode causes lesions on plant's root that forms a canker and the plant suffers from malnutrition while the nematode completes its life cycle within the root. The reducing cost of DNA sequencing has led to the availability of large sequence data sets derived from whole genome sequencing and large scale Expressed Sequence Tag (EST) discovery. ESTs are ideal for mining of SSRs and SNPs, which may then be applied to diversity analysis, genetic trait mapping, association studies, and marker assisted selection [8].

ESTs are small pieces of DNA sequence (usually 200 to 500 nucleotides long) that are generated by sequencing either one or both ends of an expressed gene. The idea is to sequence bits of DNA that represent genes expressed in certain cells, tissues, or organs from different organisms and use these "tags" to fish a gene out of a portion of chromosomal DNA by matching base pairs. ESTs provide researchers with a quick and inexpensive route for discovering new genes, for obtaining data on gene expression and regulation, and for constructing genome maps. SNP, pronounced "snip", are one-letter variations in the DNA sequence which contribute to differences among individuals. They are the most common form of DNA sequence variation. They are useful polymorphic markers to analyze the diversity and in QTL mapping. Majority of SNPs produce no effect when they occur in intronic or intergenic regions or as synonymous codon substitutions in exons. But even a single indel in coding region can cause frame shift mutations. A single non-synonymous SNP can convert an amino acid to another which in turn will lead to subtle differences in countless characteristics, like appearance, while some affect the risk for certain diseases.

SNPs are molecular markers of choice in recent years for genome mapping and diversity analysis in many crop plants, soybean [33], rye [34], cassava [19]. SNPs are invaluable as a tool for genome mapping, offering the potential for generating high-density genetic maps, which can be used to develop haplotyping system for genes or regions of interest [23]. The low mutation rate of SNPs also makes them excellent markers

for studying complex genetic traits and as a tool for the understanding of genome evolution [29]. Unlike random amplified polymorphic DNAs (RAPDs) and restriction fragment length polymorphisms (RFLPs), SNPs are direct markers because sequence information provides the exact nature of the allelic variation. They are far more prevalent than SSRs and, therefore may provide a high density of markers near a locus of interest. One of the limitations of SNPs is the initial cost associated with their development. Many cost and time effective technologies have been developed in recent years but only limited work has been carried out to examine the occurrence of SNPs in nematodes. Latest release of dbEST contains only 1065 SNP site information belonging to the free-living nematode, *Caenorhabditis elegans*.

Microsatellites or SSRs are tandemly repeated 1-6 DNA motifs that are abundant in eukaryote genome and can mutate rapidly by loss or gain of repeat units. EST-SSRs have received much attention as the increasing amounts of ESTs being deposited in databases for many crops such as rice, wheat etc. [17], [28], [36]. EST-SSRs can be rapidly developed from EST database by data mining tools at low cost, and due to their existence in transcribed region of genome, they can lead to the development of gene-based maps which help to identify candidate genes and increase the efficiency of marker-assisted selection (MAS) [11]. In addition, EST-SSRs show a higher level of transferability to closely related species than genomic SSR markers [26], [28] and can be served as anchor markers for comparative mapping and evolutionary studies [35]. SSRs are one of the most widely used molecular markers in crop breeding, population genetics, mapping and MAS. Being gene specific functional markers, they provide an efficient tool to link phenotypic and genotypic variation [12], [22]. Based on the length of SSR tracts and their potential as genetic markers, they are categorized into two groups – Class I or hypervariable markers (length of SSR ≥20bp) and Class II (length of SSR vary 12 bp ≥ and <20 bp) or potentially variable markers [30].

2. MATERIALS AND METHODS

The GenBank accession numbers ES584514 - ES584533, EY189778- EY195983, CO897542- CO898032, CO960948-CO961502 and CV130236-CV130343 were retrieved from dbEST (http://www.ncbi.nlm.nih.gov/dbEST/) of National Centre for Biotechnology Information (NCBI). These 7380 ESTs (dbEST release 100909) were represented different tissue/condition libraries. ESTs were subjected to pre-processing to remove poly A/T tails using trimmest (http://mobyle.pasteur.fr/cgi-bin/portal.py?form=trimest) and contaminations were removed using VecScreen utility of NCBI. AutoSNP [2] was used to find the SNPs and indels. Transition to transversion ratio and each type of substitutions were calculated. To find the SSRs from non redundant set of sequences, EST sequences were trimmed and clustered using Phrap [10] and the resulting sequences were used as an input of SSRs detecting perl script, MISA [31]. It is capable to detect mono- to hexa-nucleotide repeats and compound repeats. We identified two types of SSRs: (i) perfect SSRs, with an exact repeat of any of motif with length 12 bp or more, e.g. $(AT)_{15}$; (ii) compound repeats, combinations of two or more repeated motifs with length 20 or more, e.g. $(CA)_{16}(TC)_{10}$. For mononucleotides, although A, T, C and G are possible, A and T are grouped into a single category, since an A repeat on a strand is the same as a T repeat on the opposite strand. C on a strand is the same as a G on the opposite strand, resulting in two unique classes of mononucleotides, A/T and C/G [18]. Similarly, all dinucleotide motifs were grouped into the 4 following unique classes: (*i*) AT/TA; (*ii*) AG/GA/CT/TC; (*iii*) AC/CA/TG/GT; and (*iv*) GC/CG. The trinucleotide repeats are grouped into 10 unique classes as per the SSR classification [16]; Katti *et al.* 2001). Primer pairs were designed flanking the detected SSR sites using Primer3 tool [25].

3. RESULTS AND DISCUSSION

3.1 SNP /Indel Identification

A total of 7,380 EST sequences of *R. similis* were retrieved from dbEST. It is observed that 1174 sequences contained suspected poly (A/T) tails with minimum length of 4 bp or more and these were removed. Among the 7380 ESTs, 4722 sequences were found to have redundant to at least one sequence and which were grouped as 1152 clusters to predict SNP and indels from redundant data sets. Among the clusters, 644 clusters containing two sequences, 224 clusters containing three sequences and 284 clusters containing four or more sequences were present. Since the EST sequences are prone to sequencing error, we considered the cluster groups containing four or more sequences for final SNP calculation. A total of 1038 SNPs and 108 indels were identified from those sequences (**Table 1**). Candidate SNPs were categorized according to nucleotide substitution as either transition (C↔T or G↔A) or transversion (C↔G, A↔T, C↔A or T↔G). We found transversions (537) as more predominant than transitions (501) in available *R. similis* transcriptiome (**Figure 1**). However, among the individual substitutions the transition type substitutions G↔A and C↔T were found to be most predominant. Our studies indicate that the density of SNP in *R. similis* is 1 SNP/191 bp and that of indel is 1 indel/1.8kbp. Though EST SNP markers are most popular in plant ESTs and human beings, so far there is no evidence for development of a SNP marker and frequency estimation in nematodes based on EST data. The transition to transversion ratio (Ts/Tv) was found to be 0.93 in *R. similis*. In general transitions occur at higher frequencies than transversions such as beet root [27], maize [4] and oil palm [24]. But, Ts/Tv ratio < 1 (more transversions than transitions) was seen in regulatory genes such as endonuclease reverse transcriptase and Tc1 - like transposase [13]. A recent study on grasshopper genome revealed that the majority of transitions of Cytosine residues are at methylated sites (CpG dinucletoide). After accounting for this methylation effect, there was no significant difference between transition and transversion rates [20]. Transversions were seen at higher frequencies than transitions in ginger (*Zingiber officinale* Rosc) EST-SNP [5] and MA (Mutation Accumulated) lines of genomes of nematode *C elegans*. The much lower Ts/Tv ratios observed in MA-line genomes suggest that, genome wide transversions might be more susceptible to selective purging than transitions in *C. elegans* natural populations [7]. The Ts/Tv ratio nearer to one indicates a near neutral selection in *R. similis* which is predominantly parthenogenetic.

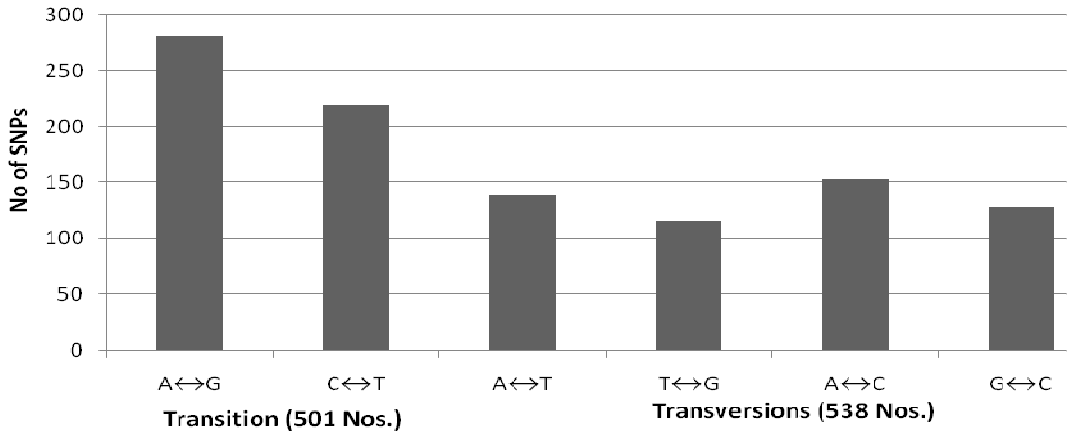

Figure 1. Frequency of SNPs classified as transition and tramsverion substitutions. The two transition classes A↔G and C↔T were found predominat than different types of transversion substitutions.

Table 1. Base changes and their frequency of occurrence of SNPs/Indels in *Radopholus similis* transcriptome

Sl. No.	SNP type	Frequency
1	Transition (501)	
	G↔A	281
	C↔T	220
2	Transversion (537)	
	C↔G	129
	A↔T	139
	C↔A	153
	T↔G	116
3	Indels (108)	
	A	46
	T	21
	G	27
	C	14

Indels occurred at very lower frequency (1 indel/1.8kpb) in *R. similis*. Indels may be produced by errors in DNA synthesis, repair, recombination or be due to the insertion and excision of transposable elements that often leaves a characteristic DNA footprint of several nucleotide bases. Adenine involved indels were found to be more abundant than Cytosine indels. Our study indicated several putative SNP/Indel markers for *R. similis* genome which could be useful for marker assisted selection (MAS) and applied in population genetics and high resolution genetic map construction.

3.2 Simple Sequence Repeats

Pre-processed ESTs were separated as 1152 contigs and 2657 singletons using sequence assembly program Phrap (Green, 1999). A total of two contigs and 143 singletons (of size < 100 bp) were found using base_counter perl script and were removed. Using MISA (Thiel *et al.,* 2003) perl script, we found out a total of 1449 SSR sites including 1276 perfect and 173 compound repeats. Among the perfect SSRs, 77 belonged to the class I and 1199 belonged to the class II type markers. We found that SSRs occured in the *R. similis* genome at a frequency of 1 SSR per 1.2 Kb. Though, only less number of SSRs was isolated from nematodes, mining of public databases for nematode SSRs is not reported so far. So the results of the present study could not be compared with other reports. Among the SSRs detected, predominant type was hexa-nucleotide repeats (836 numbers) followed by mononucleotide repeats (207 numbers) (**Table 2**).

Among the hex-nucleotide repeats, 646 different patterns were observed and most of the repeats are abundant in A/T bases (minimum 3 'A' or 'T' nucleotides). In the case of mononucleotide repeats A/T repeats are common (97.5% of mononucleotide repeats). The other possible repeat G/C was very rare (5 nos). Among the dimeric repeat motifs, the AG/GA/TC/CT type was observed 6 times and CA/AC/GT/TG type 5 times. Among the trimeric repeats, AAT/ATA/TAA/TTA/TAT/ATT (22 nos.) was most abundant followed by AAC/ACA/CAA/TTG/TGT/GTT, AGG/GGA/GAG/TCC/CCT/CTC and AGC/GCA/CAG/ TCG/ CGT/GTC (each with 17 times) motifs.

Table 2. Distribution of different types of SSR motifs in ESTs of *Radopholus similis*

Motif	Number of SSRs	Frequency
Mono	207	14.2%
Di	11	0.7%
Tri	122	8.4%
Tetra	95	6.5%
Penta	5	0.3%
Hexa	836	57.7%
Compound	173	11.9%

Primer pairs were successfully designed for 1239 sequences while the remaining SSRs failed to design primers as they were located at either ends of ESTs. Despite their many advantages for studying gene flow in eukaryotes, autosomal markers such as microsatellites have been developed for only a few parasitic nematodes, including parasites of sheep [14], pigs [6] , human [1], [3], [32], rats [9], red grouse [15] and the plant parasite *Meloidogyne artiellia* [21]. The SSRs reported here would provide additional marker information for the plant parasitic nematode, *R. similis.*

4.CONCLUSION

Our findings suggest that the nucleotide substitution towards transversion is not a rare event in molecular evolution. Putative SNP/Indels were categorized according to substitution bias, and as many as 1038 SNPs and 108 small indels (insertion/deletion) were found with a density of one SNP / 191bp and indel frequency was one indel/1.8 kbp. A total of 1449 EST-SSRs were reported in this study. Hexa-nucleotide repeats (836 nos.) followed by mononucleotide repeats (207 nos) were found to be more abundant than other type of repeats. Since SSR and SNP markers are invaluable tools for genetic analysis, the identified SSRs and SNPs could be exploited for diversity analysis, genetic trait mapping, association studies and marker assisted selection of *R. similis*.

5. ACKNOWLEDGEMENTS

This work was supported by a financial grant from Department of Biotechnology (BTISnet), Government of India, New Delhi, India.

6. REFERENCES

[1] Anderson, J.D., Williams-Blangero, S., and Anderson, T.J.C. Spurious genotypes in female nematodes resulting from contamination with male DNA. *J Parasitol.* 89 (2003), 1232–4.

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

[3] Barker, G.C., and Bundy, D.A.P. Isolation and characterization of microsatellite loci from the human whipworm *Trichuris trichiura*. *Mol. Ecol.* 9 (2000), 1181–3.

[4] Batley, J., Barker, G., Helen, O'Sullivan., Edwards, K.J., and Edwards, D. Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiol.* 132 (2003), 84-91.

[5] Chandrasekar, A., Riju, A., Sithara, K., Anoop, S., and Eapen, S.J. Identification of single nucleotide polymorphism in ginger using expressed sequence tags. *Bioinformation* 4(3) (2009), 119-122.

[6] Conole, J.C., Chilton, N.B., Jarvis, T., and Gasser, R.B. Mutation scanning analysis of microsatellite variability in the second internal transcribed spacer (precursor ribosomal RNA) for three species of *Metastrongylus* (Strongylida: Metastrongyloidea). *Parasitol.* 122 (2001), 195–206.

[7] Denvera, D.R., Dolan, P.C., Wilhelm, L.J., Sung. W., Lucas-Lledo, J.I., Howe, D.K., Lewis, S.C., Okamoto, K., Thomas, W.K., Lynch, M., and Baer, C.F. A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. *Proc. National Acad. Sci.* (USA) 106 (38) (2009), 16310–16314.

[8] Duran, C., Appleby, N., Edwards, D. and Batley, J. Molecular genetic markers: discovery, applications, data storage and visualisation. *Curr. Bioinformatics* 4 (2009), 16-27.

[9] Fisher, M.C., and Viney, M.E. Microsatellites of the parasitic nematode *Strongyloides ratti*. *Mol. Biochem. Parasitol.* 80 (1996), 221–224.

[10] Green, P. (unpublished), (1999) http://www.phrap.org

[11] Gupta, P.K., and Rustgi, S. Molecular markers from the transcribed/expressed region of the genome in higher plants. *Functional Int. Genomics* 4 (2004), 139-162.

[12] Gupta, P.K., and Varshney, R.K. The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. *Euphytica* 113 (2000), 163-185.

[13] Hale, M.C., McCormick, C.R., Jackson, J.R., and DeWoody, J.A. Next-generation pyrosequencing of gonad ranscriptomes in the polyploid lake sturgeon (*Acipenser fulvescens*): the relative merits of normalization and rarefaction in gene discovery. *BMC Genomics* 10 (2009), 203.

[14] Hoekstra, R., Criado-Fornelio, A., Fakkeldij, J., Bergman, J., and Roos, M.H. Microsatellites of the parasitic nematode *Haemonchus contortus*: polymorphism and linkage with a direct repeat. *Mol Biochem. Parasitol.* 89 (1997), 97–107.

[15] Johnson, P.C.D., Webster, L.M.I., Adam, A., Buckland, R., Dawson, D.A., and Keller, L.F. Abundant variation in microsatellites of the parasitic nematode *Trichostrongylus tenuis* and linkage to a tandem repeat. *Mol. Biochem. Pasaitol.* 148 (2006), 210-218.

[16] Jurka, J., and Pethiygoda, C. Simple repetitive DNA sequences from primates: Compilation and analysis. *J Mol. Evol.* 40 (1995), 120-126.

[17] Kantety, R.V., La Rota, M., Matthews, D.E., and Sorrells, M.E. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum, and wheat. *Plant Mol. Biol.* 48 (2002), 501-510.

[18] Katti, M.V., Ranjekar, P.K., and Gupta, V.S. Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol. Biol. Evol.* 18 (2001), 1161-1167.

[19] Kawuki, R.S., Ferguson, M., Labuschagne, M., Herselman, L., and Kim, D.J. Identification, characterization and application of single nucleotide polymorphisms for diversity assessment in cassava (*Manihot esculenta* Crantz). *Mol. Breeding* 23(2009), 669-684.

[20] Keller, I., Bensasson, D., and Nichols, R.A. Transition-transversion bias is not universal: A counter example from grasshopper pseudogenes. *PLoS Genetics* 3(2) (2007), e22.

[21] Luca, F.D., Reyes, A., Veronico, P., Vito, M.D., Lamberti, F., and Giorgi, C.D. Characterization of the (GAAA) microsatellite region in the plant parasitic nematode *Meloidogyne artiellia.*. *Gene* 293 (2002), 191-198.

[22] Powell, W., Machray, G.C., and Provan, J. Polymorphism revealed by simple sequence repeats. *Trends Plant Sci.* 1(1996), 215-222.

[23] Rafalski, A. Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opinion Plant Biol.* 5 (2002), 94-100.

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

[25] Rozen, S., and Skaletsky, H.J. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (Eds) *Bioinformatics Methods and Protocols: Methods in Molecular Biology.* Humana Press, Totowa, NJ, (2000) pp. 365-386.

[26] Saha, M.C., Mian, M.A., Eujayl, I., Zwonitzer, J.C., Wang, L., and May, G.D. Tall fescue EST-SSR markers with transferability across several grass species. *Theor Appl Gen* 109 (2004), 783-791.

[27] Schneider, K., Weisshaar, B., Borchardt, D.C., and Salamini, F. SNP frequency and allelic haplotype structure of *Beta vulgaris* expressed genes. *Mol. Breeding* 8(2001), 63-74.

[28] Scott, K.D., Eggler, P., Seaton, G., Rossetto, M., Ablett, E.M., Lee, L.S., and Henry, R.J. Analysis of SSRs derived from grape ESTs. *Theor. Appl. Gen.* 100 (2000), 723-726.

[29] Syvanen, A.C. Accessing genetic variation: Genotyping single nucleotide polymorphisms. *Nature Reviews Genetics* 2 (2001), 930-942.

[30] Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S., and McCouch, S. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): Frequency, length variation, transposon associations, and genetic marker potential. *Genome Research* 11(2001), 1441-1452.

[31] Thiel, T., Michalek, V., and Graner, A. Exploiting EST data-bases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Gen.* 106 (2003), 411-422.

[32] Underwood, A.P., and Bianco, A.E. Identification of a molecular marker for the Y chromosome of *Brugia malayi*. *Mol. Biochem. Parasitol.* 99 (1999), 1–10.

[33] Van, K., Hwang, E.Y., Kim, Y.M., Park, H.J., Lee, S.H., and Cregan, P.B. Discovery of SNPs in soybean genotypes frequently used as the parents of mapping populations in the United States and Korea. *J. Heredity* 96, (2005), 529-535.

[34] Varshney, R.K., Beier, U., Khlestkina, E.K., Kota, R., Korzun, V., Graner, A., and Börner, A. Single nucleotide polymorphisms in rye (*Secale cereale* L.): discovery, frequency and application for genome mapping and diversity studies. *Theor. Appl. Gen.* 114 (2007), 1105-1116.

[35] Varshney, R.K,, Sigmund, R., Börner, A., Korzun, V., Stein, N., Sorrells, M.E., Langridge, P., and Graner, A. Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. *Plant Sci.* 168 (2005), 195-202.

[36] Varshney, R.K., Thiel, T., Stein, N., Langridge, P., and Graner, A. *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cellular Mol. Biol. Lett.* 7(2002), 537-546.