IN SILICO DATAMINING FOR ELUCIDATION OF REPEATS BIOLOGY AND FUNCTIONAL ANNOTATION IN SORGHUM [SORGHUM BICOLOR (L.) MOENCH.]

Thesis submitted to the
University of Agricultural Sciences, Dharwad
In partial fulfillment of the requirements for the
Degree of

MASTER OF SCIENCE (AGRICULTURE)

In

PLANT BIOTECHNOLOGY

By

ARUN S.S

DEPARTMENT OF BIOTECHNOLOGY
COLLEGE OF AGRICULTURE, DHARWAD
UNIVERSITY OF AGRICULTURAL SCIENCES,
DHARWAD – 580 005

MARCH - 2006

ADVISORY COMMITTEE

DHARWAD
MARCH 2006

(B. FAKRUDIN)
Major Advisor

Approved by:

Chairman: _____
(B. FAKRUDIN)

Members: _____
(M.S. Kuruvinashetti)

_____
(R.L. Ravi Kumar)

_____
(H.L. Nadaf)

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF PLATES

LIST OF APPENDICES

I INTRODUCTION

Sorghum [*Sorghum bicolor* (L.) Moench] is the dietary staple of more than 500 million people in over 30 countries, next only to rice, wheat, maize and potatoes in importance (Bedell *et al.,* 2005). It has the ability to endure drought, water-logging and grows well in marginal lands compared to other food crops. For millions of resource poor people, it has gained increasing importance as a food grain, green/dry fodder and feed crop during the past decades in India and Africa. Sorghum is cultivated in an area of 43.73 m ha in the world, producing 58.8 mt of grain with a productivity of 1347 kg/ha (Anonymous, 2004). India is the third largest producer of this crop, producing 7.53 mt grains from 9.4 m ha. Low productivity (801 kg/ha) is mainly due to large area (92.3 %) under rainfed conditions. Maharashtra, Karnataka, Andhra Pradesh, Madhya Pradesh, Rajastan and Tamil Nadu are the top six sorghum growing states in India, which together account for about 87 per cent area and 90 per cent of the total production (Anonymous, 2004).

A member of poaceae family and panicoid grass subfamily, sorghum is closely related to maize, millets and sugarcane (Bedell *et al.,* 2005). It is considered as a model grass genome, a source of genes for some valuable traits including adaptation to the extreme conditions of semiarid tropics (Mullet *et al.,* 2002), ability to secrete allelochemicals inhibiting weed growth (Chema *et al.,* 2000) and the presence of phenolic acids in tissues, which reduce feeding by *Locusta migratoria* (Woodhead *et al.,* 1979). If genes responsible for the resilience of this plant can be isolated, it is possibile to understand the plasticity of this crop species and to develop even more productive genotypes under adverse conditions and also produce transgenic cultivars of value.

The analysis of DNA sequence variation is of major importance in genetic studies. In this context, molecular markers have greatly contributed to genetic analysis of crop plants. A variety of molecular markers have been developed in different crop plants. Among the molecular markers, SSRs are distinctly useful for a variety of applications in plant genetics and breeding because of their reproducibility, multi allelic nature, co-dominant inheritance, relative abundance and good genome coverage (Powell *et al.,* 1996). SSRs are stretches of genomic DNA, consisting of tandemly repeated short units of 1-6 base pairs in length. They are ubiquitous in eukaryotic genomes and can be analyzed through simple PCR technology. The sequences flanking specific microsatellite loci in a genome are shown to be conserved within a particular species, across species within a genus and even among related genera (Varshney *et al.,* 2002). SSR markers have been useful for integrating the genetic, physical and sequence-based physical maps in plant species, and have also provided breeders and geneticists with an efficient tool to link phenotypic and genotypic variation (Gupta and Varshney 2000). However, development of genomic SSR markers is expensive, labour intensive and time consuming, particularly if they are being developed from genomic libraries. Despite cost, due to their importance, SSRs have been developed in a large number of plants including major cereal species such as barley, maize, oats, rice, rye, sorghum, and wheat (Varshney *et al.,* 2002)

In recent years, over seven million Expressed Sequence Tags (EST) from about 200 plant species have been deposited in public databases. Primarily the byproduct of cDNA based expression analysis, ESTs have served as an alternative to complete genome sequencing in low priority crop species. ESTs are a valuable resource for the analysis of biodiversity and gene discovery. Bioinformatics based sequence analysis tools have extended the scope of ESTs into the fields of proteomics, marker development and genome annotation. Although a collection of ESTs is not a substitute for the whole genome sequence, this 'poor man's genome' resource is a foundation for various genome-scale experiments in plants where genomes are yet to be unsequenced. The generation of ESTs facilitates gene discovery as they are direct representatives of the transcribable part of the genome and produced faster and more cheaply (Rudd, 2003).

In addition to ESTs from several projects, sequence data for many fully characterized genes and full-length cDNA clones are availble in crop plants such as rice (Kikuchi *et al.,* 2003). By using specific computer programs, the sequence data for ESTs, genes and cDNA clones can be downloaded from GenBank and scanned for detection and development of SSRs, which are typically referred to as EST-SSRs, genic SSRs or genic microsatellites. Subsequently, locus-specific primers flanking repeat domains can be designed to amplify the genic SsR alleles. Thus, the development of genic SSR markers is relatively easy and inexpensive as they are a byproduct of the sequence data from genes or ESTs that are

publicly available. Genic SSRs have some intrinsic advantages over genomic SSRs, because they are quickly obtained by electronic sorting, and have a higher level of transferability among related species as they are located in more conserved functional regions of the genome. Being the part of the transcribed part of the genome, EST-based SSR markers lead to the direct mapping of genes. Presently EST-SSRs are being used in only a few crops, as these markers are accessible only in those species for which a sufficient number of ESTs exist in public databases. In Sorghum between 1998 and 2001, close to 108,000 ESTs were deposited in GenBank. According tho the latest information there are over 2,32,921 such EST sequences (Benson *et al.,* 2006).

Availbility of ESTs with repeat motifs is the key for developing 'functionally characterized sequence motifs' or 'functional markers'. Computational tools are now regularly used to infer function based upon significant sequence similarity to experimentally verified and putative proteins. These analyses implement FASTA and BLAST comparisons against non-redundant databases as well as 'Gene Ontology' annotation (Folta *et al.,* 2005). Hence, deducing the function for SSR containing ESTs based on homology is usefull in knowing their possible function, which can be experimentally verified later.

Keeping these in view, the present investigation was undertaken with the following objectives in sorghum.

1. To construct a querying database of repeat containing genes/ ESTs of sorghum.
2. To elucidate patterns and biological nature of repeats in genes/ESTs.
3. To functionally annotate SSR containing genes and to validate some of the SSRs in sorghum.

II REVIEW OF LITERATURE

The key issues of structural genomics include analysis of genome structure, organization and evolution, which are addressed by linkage and physical mapping, genome sequencing and comparative genome analysis. The advent of DNA marker technology, pivotal to structural genomics, has recently opened up many new vistas in understanding complex biological problems (O'Brien *et al.,* 1991). Application of DNA markers and resulting molecular linkage maps allow dissection of genetic variation of complex phenotypes. In the last decade, linkage maps have been constructed in several important crop species, including sorghum. Majority of them are based on SSRs derived from the non coding part of the genome (Taramino *et al.,* 1997, Bhattramakki, 2000). The use of this type of SSRs seriously limits their utility in comparitive mapping. Contrary to this, the use of EST derived SSRs which are conserved because of their presence in genes allow us to use them in comparitive mapping. Under this assumption, closer the evolutionary relationship between species, more similar their genomes are expected to be in terms of structural and functional organization. Genomes can be mapped in a comparative way, allowing exploitation of the progress made in model species. Further, rapid accumulation of sequence data has necessiated intelligent organization of data and to facilitate retrevial of valuable information. Construction of a retrievable database has become an essential step in any comprehensive research programme in genomics. Many of the non target sequences resulting from EST/ gene discovery projects go functionally unassigned. Functional assignment/ annotation based on homology search have allowed researchers to economize on both time and effort.

Detailed review of literature on EST dereived SSR markers, its application and functional annotation through homology are presented here

2.1 EXPRESSED SEQUENCE TAGS (ESTs)

ESTs are typically unedited, automatically processed, single-read sequences produced from cDNA (small DNA molecules, reverse-transcribed from the cellular mRNA population). Libraries of cDNAs are routinely prepared, which contain tens of thousands of clones from a variety of specific tissue types, as a snapshot of gene expression during defined developmental stages and following a specific biotic or abiotic challenge. The relativly low cost of automated EST sequencing has made it an attractive route to broader sampling of the transcriptome (Rudd, 2003). Mark Adams first used the term EST in relation to gene discovery and the human genome project in 1991 (Adams *et al.,* 1991). Subsequently, 33 million ESTs have been sequenced from more than 500 species, representing a wide taxonomic variety of fungi, plants and animals (Benson *et al.,* 2006)

2.1.1 EST sequence availability and biodiversity

With the latest releases at the NCBI sequence database and the weekly updates to the EST database (http://www.ncbi.nlm.nih.gov/dbEST/index.html), there were about 32.75 million ESTs available within the public domain, on 13 January 2006 (Benson *et al.,* 2006). Top 20 plant species, in terms of the total number of ESTs available in the NCBI ESTdb are listed in Table 1

In the EST libraries, most sequences are assigned to either agricultural species, a situation of agronomic interest or model plant species (*Arabidopsis, Chlamydomonas, Physcomitrella*)

2.2 DATABASES, TOOLS AND THEIR USES

Biological data are gathered and stored in a variety of ways all over the world. In order to interpret these data in a biologically meaningful way, we need special tools, techniques and efficient algorithms. Databases and programs allow us to access existing information and explore data to find similarities and differences. The various internet based molecular biology databases have their own unique navigation tools, data storage formats and data retrieval tools (Ignacimuthu, 2003)

Table 1: List of top 20 plant species with ESTs in NCBI ESTdb as on 13 January 2006.

Sl. No.	*Plant species*	*Common name*	*ESTs available*
1	*Zea mays*	Maize	662290
2	*Triticum aestivum*	Wheat	600205
3	*Arabidopsis thaliana*	Thale cress	421027
4	*Oryza sativa*	Rice	407545
5	*Hordeum vulgare* subsp. vulgare	Barley	395065
6	*Glycine max*	Soybean	356780
7	*Pinus taeda*	Loblolly pine	329469
8	*Saccharum officinarum*	Sugarcane	246301
9	*Solanum tuberosum*	Potato	219765
10	*Sorghum bicolor*	Sorghum	208466
11	*Lycopersicon esculentum*	Tomato	199279
12	*Malus spp.*	Apple tree	197973
13	*Vitis vinifera*	Grapes	194176
14	*Picea glauca*	White spruce	132624
15	*Physcomitrella patens* subsp. patens	Moss	120702
16	*Lotus corniculatus* var. japonicus	Lotus	111623
17	*Gossypium hirsutum*	Cotton	108424
18	*Citrus sinensis*	Citrus	92521
19	*Brassica napus*	Oilseed rape	72350
20	*Helianthus annuus*	Sunflower	66098

2.2.1. Importance of Databases

A database is a logically coherent collection of related data with inherent meaning to permit certain applications. It is composed of discrete coherent parcels of information as entries or records that can be processed by a computer program. Contents of a databse can easily be accessed, managed and updated. Databases can be searched or cross-referenced either over the Internet or using downloaded versions on local computers or computer networks by multiple users. The databases are electronic filing cabinets, a convenient and efficient method of storing vast amount of information. They are assemblages of analyzed biological information into central and shareable resources (Hancock, 2002).

Databases are needed to collect and preserve data, standardize presentation to make data easy to access and search and organize into knowledge. The primary goals of databases are; i) minimizing data redundancy and ii) achieving data independence. Databases are essential for managing similar kind of data and developing a network to access them across the globe. A large amount of biological information is available all over the world through www but the data are widely distributed and it is therefore necessary for scientists to have efficient mechanisms for data retrieval (Hancock, 2002).

In order to derive maximum benefit from the vast amount of sequence information that is available today, one must establish, maintain and disseminate databases, provide easy to use software to access the information they contain, and design state-of-the art analytical tools to visualize and interpret the structural and functional clues hidden in the data (Ignacimuthu, 2003).

Databases of nucleic acid and protein sequences maintain facilities for a very wide variety of operations such as retrieval of sequences from the data base, sequence comparison, translation of DNA sequences to protein sequences, simple types of structure analysis and prediction, pattern recognition and molecular graphics. Some examples of such databases are Entrez (http://www. ncbi.nlm.nih.gov/Entrez/) and OMIM. ExPASy (http://www.expasy.ch) is the information retrieval and analysis system.

2.2.2 Types of Databases

Databases are categorised broadly based on the nature of the information being stored and the manner of data storage. Databases are broadly classified as generalized databases include DNA, protein, carbohydrate etc., and specialized databases, which EST, genome survey sequences (GSS), single nucleotide polymorphism (SNP), sequence tagged sites (STS) etc. Other specialized databases include Kabat for immunology proteins and ligand for enzymes reaction. Generalized databases are further classified into sequence and structural databases. Sequence databases contain the individual sequence records of either nucleotides or amino acids or proteins. Structure databases contain the individual sequence records of solved structures of macromolecules (e.g. Protein 3D structure).

Two principal types of databases are:

i. Relational
ii. Object-oriented.

The relational database arranges the data into tables made up of rows giving specific items in the database and columns giving the features as attributes of those items. The object-oriented database includes objects such as genetic maps, genes, or proteins, with an associated set of utilities for analysis, which help in identifying the relationships among these objects (Ignacimuthu, 2003).

2.2.3 EST database

A number of EST based sequence resources have been developed that address the quality, redundancy and partial nature of EST sequences. Sequence resources such as the dbEST database (Boguski *et al.,* 1993) and the EMBL database (Stoesser *et al.,* 2003) archive all the available ESTs and provide methods to search for individual sequences on the basis of species, clone or homology attributes. However, these searches are limited to the sequence features that are supplied when the sequence is submitted.

A range of plant specific EST databases have been described in the last few years in which sequence analysis and annotation has moved beyond the EST sequence and significant value has been added. Value addition for these EST sequences involves clustering and assembling the ESTs into a more manageable datasets in terms of size and quality clusters. A number of tools are also available with such databases, which generally perform processing steps to achieve a common result. The sequences need to be carefully cured of vector and polylinker remnants before a clustering protocol places the ESTs into groups of similar sequences (Heumann and Mewes, 1996). The Assembly step then places the clustered sequences into logical contigs and singletons (Gordon *et al.,* 1998; Huang and Madan, 1999). Finally, the clustering process yields sequences that are typically longer than any individual EST and are of a higher quality, without redundancy. Additionally, cluster consensus sequences bring out valuable information on sequence polymorphisms that would otherwise are not observable (Rudd, 2003).

These cluster consensus and singleton sequences form the core sequence data within several plant specific EST derived databases. A collection of these resources is listed in Table 2. Most of these sequence databases have added further value by attaching additional annotation to the sequences and by providing methods to select specific sequences or groups of sequences that satisfy set criteria. The most valuable annotations and methods are those that assign tentative function and allow retrieval and identification of sequences on the basis of tissue or a specific challenge (Rudd, 2003).

2.3 MICROSATELLITES

Microsatellites containin 1 to 6 bp DNA sequence motifs repeated several times (Tautz and Renz, 1994). Popularly described as simple sequence repeats (SSRs) in plants (Morgante and Oliveri, 1993) and as short tandem repeats (STRs) in animal systems (Edwards *et al.,* 1991). These motifs, known as di-, tri-, tetra-nucleotide repeats, *etc.*, accordingly. Microsatellites are abundant and occur randomly in all eukaryotic DNA examined so far (Gupta *et al.,* 1996; Gupta and Varshney, 2000). Microsatellites can be termed simple (“perfect” or “pure”) if they contain several repeats of two or more nucleotides *i.e.*, $(N_1N_2...N_x)_n$ or they may be more complicated ("imperfect" or "compound") where two or more repeat

Table 2: Important plant specific EST databases.

Plant EST Database	*URL*	*Genomes*	*References*
TIGR Plant Gene Indices	http://www.tigr.org/tdb/tgi/plant.shtml	All large collections of plant ESTs	Quackenbush *et al.,* 2001.
NCBI Unigenes	http://www.ncbi.nlm.nih.gov/UniGene/	11 plants with largest EST collections	Wheeler *et al.,* 2003.
MIPS Sputniks	http://mips.gsf.de/proj/sputnik/	All large collections of plant ESTs	Rudd *et al.,* 2003
PlantGDB	http://www.zmdb.iastate.edu/PlantGDB/	All large collections of plant ESTs	Rudd, 2003
University Minnesota	http://www.ccgb.umn.edu/	*Pinus taeda, Medicago truncatula, Glycine max*	Lamblin *et al.,* 2003
B-EST barley database	http://pgrc.ipk-gatersleben.de/est/login.php	*Hordeum vulgare*	Rudd, 2003
Kazusa EST databases	http://www.kazusa.or.jp/en/plant/database.html	*Lotus japonicus, Arabidopsis thaliana, Porphyra yezoensis, Chlamydomonas reinhardtii*	Rudd, 2003
Solanaceae genomics network	http://sgn.cornell.edu/	Different *Lycopersicon* and *Solanum* species	Hoeven *et al.,* 2002
Chlamydomonas resource centre	http://www.biology.duke.edu/chlamy_genome/	*Chlamydomonas reinhardtii*	Shrager *et al.,* 2003
Arizona Genomics Computational Lab (AGCoL)	http://agcol.arizona.edu/pave/cotton/	A global assembly of cotton ESTs	Udall *et al.,* 2006
Laboratory for Genomics and Bioinformatics (Fungen)	http://www.fungen.org/	EST libraries for *Sorghum spp., Homo sapiens, Ichthyophthirius multifiliis,* Pine and Horse	Pratt *et al.,* 2005

motifs are present, *e.g.,* $(CA)_n(GT)_n$, or $(CA)_n(GT)_n$. Some may have spacers between motifs repeated several times *e.g.,* $(GA)_n$ $(N)_n$ $(CT)_n$ (Chambers and MacAvoy, 2000).

2.3.1 Microsatellite-based markers

SSRs are considered as the most efficient markers, but their use is still limited because of the long and laborious steps needed to develop them. Two general strategies are employed to develop SSR markers: by searching for sequences containing microsatellites in the available databases and screening the genomic or any other library with probes complementary to microsatellite sequences. Exceptionally, some strategies without library construction have also been developed recently (Trojanowska *et al.,* 2004).

2.3.2 Features of microsatellites:

SSRs are thought to arise and develop due to replication slippage and a mutations could expand or contract them. It is also suggested that SSRs undergo a life cycle, having birth, growth and death with life span ranging from tens to even hundreds of millions of years (Messier *et al.,* 1996; Primmer and Ellegren, 1998). In humans about 3 per cent of the genome is occupied by SSRs, distributed throughout the genome in both coding and non-coding regions (Tothet *et al.,* 2000), respectively termed as genomic and genic SSRs (Varshney *et al.,* 2005). The study of repeat density and its distribution pattern in the genome is expected to help in understanding their significance. There is accumulating evidence to suggest that SSRs might as well regulate gene expression (Kunzler *et al.,* 1995; Moxon and Wills 1999). Genic SSRs have high proportion of high-quality markers than genomic SSRs and they show prominent bands and distinct allelic peaks (Thiel *et al.,* 2003; Kota *et al.,* 2001; Yu *et al.,* 2004; Saha *et al.,* 2004; Nicot *et al.,* 2004; Cho *et al.,* 2000; Eujayl *et al.,* 2001). High quality and robustness of amplification patterns, along with other merits associated with EST-SSR markers enhance their value, especially for germplasm characterization (Varshney *et al.,* 2005).

Amplification rate and null alleles:

A success rate of 60–90 per cent amplification for both genomic and EST-SSRs has been reported in different studies (Thiel *et al.,* 2003; Kota *et al.,* 2001; Yu *et al.,* 2004; Saha *et al.,* 2004; Gupta *et al.,* 2003; Cordeiro *et al.,* 2001). This might be due to (i) one or both primers of the EST-SSR extend across a splice site (ii) the presence of large introns in genomic DNA sequence (iii) the use of questionable sequence information for primer development and (iv) design of primers from chimeric cDNA clones. Thus, the quality of the EST-SSR sequence for designing the primer pairs is important (Varshney *et al.,* 2005). In one survey, up to 9.0 per cent of cereal ESTs were of low quality (Sreenivasulu *et al.,* 2002), which should not be considered for designing primers (Thiel *et al.,* 2003). Comparitively, amplicon size of EST-SSRs frequently deviated from expectation (Thiel *et al.,* 2003; Kota *et al.,* 2001; Yu *et al.,* 2004; Cordeiro *et al.,* 2001; Nicot *et al.,* 2004). Which is probably a result of the presence of introns and insertions-deletions (in-dels) in the corresponding genomic sequence, as was substantiated by sequence analysis (Saha *et al.,* 2004). Large in-dels (20 Cbp) in the SSR-ESTs can alter amplicon size sufficiently to enable visualization of polymorphism on agarose gels at a significantly low than acrylamide gels (Yu *et al.,* 2004).

Null alleles were observed by using EST-SSR markers in studies on kiwi fruit (Fraser *et al.,* 2003), rice (Cho *et al.,* 2000), spruce (Rungis *et al.,* 2004) and wheat (Gupta *et al.,* 2003; Eujayl *et al.,* 2001). Null alleles can occur due to (i) the deletion of microsatellite at a specified locus (Callen *et al.,* 1993) (ii) mutations (in-dels or substitutions) in the primer binding site (Lehman *et al.,* 1996). Occurrence of null alleles complicates the interpretation of data on segregation as the heterozygotes cannot be identified and reaction failure is not detected.

Level of polymorphism:

EST-SSR primers have been reported to be less polymorphic compared to genomic SSRs because of greater DNA sequence conservation in coding regions (Scott *et al.,* 2000; Rungis *et al.,* 2004; Cho *et al.,* 2000; Eujayl *et al.,* 2001; Chabane *et al.,* 2005; Russell *et al.,*

2004). Further, EST-SSRs derived from 3' ESTs were found to be superior to those derived from 5' ESTs (Scott *et al.,* 2000; Gao *et al.,* 2003; Gupta *et al.,* 2003; Holton *et al.,* 2002). Owing to the process of cDNA generation (polyT priming), there is a preferential selection of untranslated regions (UTRs) within 3' ESTs, resulting in more variation than in 5' ESTs (Scott *et al.,* 2000). They also reported that there were differences in polymorphism among microsatellites derived from 3' UTR (most polymorphic at cultivar level), 5' UTR (most polymorphic between cultivar and species) and within the coding sequence (most polymorphic between species and genera).

2.3.3 Frequency of EST-SSR markers:

Due to large-scale genome/EST sequencing projects in several plant species, including cereals, has resulted in large ammount of sequence data which can be utilized for studying the frequency, distribution and organization of microsatellites in the expressed portion of the genome. For development of EST-SSRs, ESTs have been scanned in different plant species, including cereals such as rice (Temnykh *et al.,* 2000; 2001), barley (Kota *et al.,* 2001; Thiel *et al.,* 2003), wheat (Eujayl *et al.,* 2002; Gao *et al.,* 2003; Gupta *et al.,* 2003), and rye (Hackauf and Wehling, 2002). These efforts have allowed estimation of the density of SSRs in expressed regions of the genomes (Table 3). Varshney *et al.* (2002) found an average density of one SSR for every 6.0 kb of ESTs after scanning 75.2 Mb barley, 54.7 Mb maize, 43.9 Mb rice, 3.7 Mb rye, 41.6 Mb sorghum and 37.5 Mb of wheat sequences. However, in another study, the frequency of SSRs was one in every 11.81 kb in rice, 17.42 kb in wheat and 28.32 kb in maize (Gao *et al.,* 2003). Difference in the frequency of SSRs in the ESTs of a particular species in different studies may be attributed to criteria of SSR search and the quantum of data used for this purpose.

2.3.4 Types of repeats in EST SSRs

In a comprehensive study of cereals, Varshney *et al.* (2002) found that the tri nucleotide repeats (TNRs) were the most frequent (54-78 %) followed by the di nucleotide repeats (DNRs) (17.1-40.4 %). The abundance of TNR-SSRs is possibly due to the absence of frameshift mutations in such SSRs (Metzgar *et al.,* 2000). Analysing coding DNA sequences in the whole genomes of fruitfly, the nematode *C. elegans* and the budding yeast, Katti *et al.* (2001) opined that trimeric codon repeats corresponding to small hydrophilic amino acids were more frequent as these are better tolerated than those for hydrophobic and basic amino acids.

In cereal genomes, among the DNRs, the motif AG is the most frequent (38- 59 %) followed by the motif AC (20-34 %) in all the species except rye, where these frequencies were 50 per cent for AC and 37.9 per cent for AG (Varshney *et al.,* 2002). The most infrequent motif was CG in all the species (1.7 to 9.0 %) except in barley, where AT is the least frequent (8.4%). Among the TNRs, motif CCG is the most frequent, ranging from 32 per cent in wheat to 49 per cent in sorghum followed by AGC (13-30 %) in barley, maize, rice and sorghum, and AAC in wheat (27%) and rye (16%). The third most frequent motif was AGG in barley, rice, rye, sorghum, AGC in wheat, and AAC in maize.

The proportion of DNRs, TNRs and Tetra nucleotide repeats (TTNRs) motifs observed varied with the length of the SSRs within and among barley, wheat and rice (Varshney *et al.,* 2005). Yu *et al.* (2004) reported that 74 per cent of the TNRs were found in coding regions, 20 per cent in 5' UTRs and 6 per cent in 3' UTRs. By contrast, only 19 per cent of the DNRs were in coding regions and 42 per cent and 39 per cent were in 5' and 3' UTRs, respectively.

Expansion of trinucleotide repeats in some human genes was reported to be associated with neurological disorders (Sasaki *et al.,* 1996; Sanpei *et al.,* 1996; Pulst *et al.,* 1996; Neri *et al.,* 1996; Pujana *et al.,* 1997). Variation in number of GA/CT repeats in the 5' UTR of the waxy gene is correlated with amylose content in rice (Ayers *et al.,* 1997). In addition, Cho *et al.* (2000) reported 27 rice genes, which had SSR in the exons (8), introns (5), 5'UTR (8) or 3'UTR regions (5). Yet, function of genes that contain SSRs and the role of the SSR motif itself in plant genes are poorly documented.

Table 3: Density of microsatellites in cereal genomes

Crops	*Source of SSRs*	*Density (kb of DNA per SSR)*	*References*
Barley	Genomic DNA	7.40	Cardle *et al.* (2000)
	ESTs	7.50	Varshney *et al.* (2002)
		3.40	Kantety *et al.* (2002)
Maize	Genomic DNA	4.5/5.71	Morgante *et al.* (2002)
	ESTs	8.10	Cardle *et al.* (2000)
		1.63/2.12	Morgante *et al.* (2002)
		1.50	Kantety *et al.* (2002)
		7.50	Varshney *et al.* (2002)
		28.32	Gao *et al.* (2003)
Rice	Genomic DNA	225-240	Wu and Tanksley (1993)
		330-365	Panaud *et al.* (1995)
		7.40	Cardle *et al.* (2000)
		16/1.9	Temnykh *et al.* (2001)
		2.64/3.52	Morgante *et al.* (2002)
	BAC end sequences	40/3.7	Temnykh *et al.* (2001)
	ESTs	3.40	Cardle *et al.* (2000)
		19.00	Temnykh *et al.* (2001)
		0.86/1.06	Morgante *et al.* (2002)
		3.90	Varshney *et al.* (2002)
		4.70	Kantety *et al.* (2002)
		11.81	Gao *et al.* (2003)
Rye	ESTs	5.50	Varshney *et al.* (2002)
Sorghum	ESTs	5.50	Varshney *et al.* (2002)
		3.60	Kantety *et al.* (2002)
Wheat	Genomic DNA	440-704	Roder *et al.* (1995)
		212-292	Ma *et al.* (1996)
		3.35/5.16	Morgante *et al.* (2002)
	ESTs	1.33/1.67	Morgante *et al.* (2002)
		6.20	Varshney *et al.* (2002)
		3.20	Kantety *et al.* (2002)
		17.20	Gao *et al.* (2003)
		9.20	Gupta *et al.* (2003)

2.4 TOOLS FOR DATA MINING

Initially, identification of SSRs from publicly available ESTs and gene sequences was done through 'regular expression matching' or BLASTN in FASTA or BLAST2 formatted sequences (Scott *et al.,* 2000; Temnykh *et al.,* 2000). Later several Perl scripts, search modules or programs were developed for recognition of SSR patterns in the sequence files (Table 4). Among the programs available in public domain, the MIcroSAtellite (MISA) search module in perl script has some useful features for EST quality control and for designing the primer pairs for EST-SSRs in a batch file (Thiel *et al.,* 2003; available at http://pgrc.ipkgatersleben.de/misa/). MISA has been used in several studies (Thiel *et al.,* 2003; Kota *et al.,* 2001; Varshney *et al.,* 2002; Yu *et al.,* 2004; Khlestkina *et al.,* 2004). Another SSR finder, called 'Sputnik', has a feature to enable the user to specify the percent imperfection allowed in the SSR (Morgante *et al.,* 2002; available at C. Abajian; http://abajian.net/sputnik/ index.html), and Perl scripts have been written to facilitate routing the output to a relational database and batch primer design for Primer3 (http://wheat.pw.usda.gov/ITMI/ESTSSR/ LaRota/). A user friendly, windows based programme, 'FastPCR', (http://www.biocenter.helsinki.fi/bi/Programs/fastpcr.htm) identifies repeats of different types in batches of upto 1,00,000 sequences at a time (Kalendar, 2006). However, well known softwares like GCG do not have desireable algorithm for identifying the repeats.

2.5 CLUSTERING ESTs:

A cluster is defined here as a group of overlapping EST sequences. For development of unique genic SSR markers, a non redundant EST dataset should be used.

Several programmes have been developed to obtain a non redundant dataset from publicly available ESTs, and each has used a novel approach to meet specific goals. Current indices such as TIGR Human Gene Index (http://www.tigr.org) and EST cluster databases such as UniGene (Boguski *et al.,* 1993; Schuler *et al*., 1996) discard noisy information and rely on longest informative ESTs, significant transcript matches or joined genomic exons to seed index classes. TIGR Human Gene Index (HGI; http:// www.tigr.org) uses the strict assembly method of TIGR_ASSEMBLER (Sutton *et al.,* 1995), tightly grouping highly related sequences, to produce accurate consensus sequences. The method strictly discards under-represented, divergent or noisy sequences in favor of confidence based on transcript redundancy, but in doing so it generates "short" consensus sequences, which might eliminate related sequences that might arise due to alternative splicing and other valuable forms of sequence diversity (Bouck *et al.,* 1999).

A complementary approach of Uni- Gene (Boguski and Schuler, 1995; Schuler *et al.,* 1996), the Genexpress Index (Houlgatte *et al.,* 1995), and the Merck Gene Index (Williamson *et al.,* 1995) group sequences into clusters based on sequence overlap above a given alignment threshold, accepting only the longest representative of an index class as its consensus. Apart from these methods, clustering can also be done using a good sequence assembler manually, by selecting a good representative sequence or a consensus sequence using specific criterias like similarity percentage and overlap length among similar sequences. The process is time consuming and easy for small datasets. Nevertheless, the quality of unigenes is as good as any other clustering softwares.

2.6 APPLICATIONS OF GENIC SSRs

Molecular markers have proven useful to assess and characterize genetic variation within natural populations and among breeding lines for effective conservation and exploitation of genetic resources in crop improvement programs (Mohammadi and Prasanna, 2003). Evaluation of germplasm with SSRs derived from genes, genic ESTs, might enhance the role of genetic markers by assaying the variation in transcribed genes, especially those with known function.

Genic SSRs have the potential of being functional markers in cases where polymorphism in the repeat motifs affect the function of the gene (Anderson and Lubberstedt

Table 4: Important softwares available for identification of repeats.

Script or Program	*References*
MIcroSAtellite (MISA)	http://pgrc.ipk-gatersleben.de/misa/; (Thiel et al., 2003)
SSRFinder	Gao *et al.,* 2003
BuildSSR	Rungis *et al.,* 2004
SSR Identification Tool (SSRIT)	Kantety *et al.,* 2002
Tandem Repeat Finder (TRF)	Benson, 1999
Tandem Repeat Occurrence Locator (TROLL)	Castelo *et al.,* 2002
CUGIssr	http://www.genome.clemson.edu/projects/ssr/
Sputnik C. Abajian;	http://abajian.net/sputnik/index.html
Modified Sputnik	Morgante *et al.,* 2002
Modified Sputnik II	http://wheat.pw.usda.gov/ITMI/EST-SSR/LaRota/
SSRSEARCH	ftp://ftp.gramene.org/pub/gramene/software/scripts/ssr.pl
FastPCR	Kalendar 2006

2003), permitting 'direct allele selection' if they are shown to be associated or responsible for the target trait (Sorrells and Wilson, 1997). Recently, a 'Dof' homolog (*DAG1* gene) that showed a strong effect on seed germination in *Arabidopsis* (Papi *et al.,* 2000) was mapped on chromosome 1B of wheat by using wheat EST-SSR primers (Gao *et al.,* 2004). Similarly, Yu *et al.* (2004) identified two EST-SSR markers linked to photoperiod response gene *(ppd)* in wheat. Mapping candidate genes and genic SSRs will also facilitate genome alignment across related species (Yu *et al.,* 2004; Varshney *et al.,* 2005).

EST-SSRs have been been integrated and genome-wide genetic maps have been prepared in many crops *viz.,*. wheat (Yu *et al.,* 2004; Nicot *et al.,* 2004; Holton. *et al.,* 2002; Gao *et al.,* 2004), barley (Thiel *et al.,* 2003), rye (Khlestkina *et al.,* 2004) and rye grass (Warnke *et al.,* 2004), cotton (Zhiguo *et al.,* 2006; Han *et al.,* 2004), soybean (Zhang *et al.,* 2004), Potato (Feingold *et al.,* 2005) kiwi fruit (Fraser *et al.,* 2004), raspberry (Graham *et al.,* 2004) *etc.* Genic SSRs show a characteristic distrubution throughout the genome and get concentrated in gene rich regions, unlike genomic SSRs, which are clustered around the centromere (Thiel *et al.,* 2003; Yu *et al.,* 2004; Gao *et al.,* 2004). Distribution of genic SSRs on the genetic map will show the distrubution of genes in the genome.

Another important feature of the genic SSR markers is that, unlike genomic SSRs, they are transferable among related species and genera (Yu *et al.,* 2004; Varshney *et al.,* 2005). Thus, EST-SSR markers could be used in related plant species for which little information is available on SSRs or ESTs. In addition, the genic SSRs are good candidates for the development of conserved orthologous markers for genetic analysis and breeding of different species, for instance, Varshney *et al.* (2005) showed that a set of 12 barley EST-SSR markers had significant homology with the ESTs of four monocotyledonous species (wheat, maize, sorghum and rice) and two dicotyledonous species (*Arabidopsis* and *Medicago*).

2.7 FUNCTIONAL CHARACTERIZATION OF ESTs

The starting point of functional marker development is the sequence of a gene with an assigned function. Accumulation of plant nucleotide sequences in recent years has been exponential, with more than 54 million entries deposited at GenBank (February 2006; ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt). In the model plant *Arabidopsis,* less than 10 per cent of 25,000 genes have been functionally characterized (Breyne and Zabeau, 2001), but in other speciesIn other species, the number is very low. However, based on sequence homology, putative functions can be assigned to 30–50 per cent of the expressed sequences in any species (Ronning *et al.,* 2003). In cotton, a set of 33,665 ESTs was annotated using both the 'Gene Ontology' (Ashburner *et al.,* 2000) and 'Protein families (Pfam) indices' (Bateman *et al.,* 2000). All sequences were used to search for similar protein sequences in the UniProt database (BLASTx). Using the best hits found by BLASTX (<1E-20), a putative GO annotation was found for 64 per cent of the cotton ESTs, and these putative gene functions were categorized into functional categories (Udall *et al.,* 2006). This candidate gene approach and synteny relationships between plant genomes (Barnes 2002; Freeling 2001; Ware *et al.,* 2002) have been successfully exploited to identify agronomically relevant genes (Collins *et al.,* 1998; Quint *et al.,* 2002). In addition, high-throughput assays such as expression profiling have been developed recently (Breyne and Zabeau, 2001) to identify candidate genes on a large scale. Other methods, including RNA interference (Denli and Hannon, 2003), T-DNA and transposon tagging (Walden 2002; May and Martienssen, 2003) and gene expression QTL mapping (Jansen and Nap, 2001) have been used to determine gene function.

Accumulation of genomic information on crop plants at the current pace, data mining to extract useful information is order of the day. The present effort is one such, where an attempt was made to mine available ESTs data for obtaining SSR markers, develop a database for EST-SSRs and to experimentally validate the markers.

III. MATERIAL AND METHODS

The present study was conducted at the Institute of Agri-Biotechnology, University of Agricultural Sciences, Dharwad, India. The details of material used and the methodologies adopted were as follows.

3.1 DOWNLOADING SORGHUM ESTs FROM THE PUBLIC DATABASE

All Sorghum ESTs from dbEST database of the National Centre for Biotechnology Information (NCBI) and Fungen (http://www.fungen.org/) were downloaded (last accessed on July 2005). First, 'sorghum AND EST' was used as a string of keyword to search nucleotide sequences at the NCBI databases (http://www.ncbi.nlm.nih.gov). All matched sequences were downloaded by changing the 'display' dropdown menu to FASTA, and the 'send to' dropdown menu to FILE. After downloading, the file containing all the available 2,32,922 sequences was saved as a text file. The major part of the ESTs was contributed by the Laboratory for Genomics and Bioinformatics (known as 'Fungen'), University of Georgia, Department of Plant Biology, USA.

3.2 CONSTRUCTION OF SORGHUM ESTs DATABASE

A database was established using standalone BLAST, a DOS based programme downloaded from ftp://ftp.ncbi.nih.gov/blast/ to retrieve the sequences, count total number of bases and number of entries in the database. The ESTs were retrieved and grouped based on transcriptome definition (stress, tissue, etc.) and separately saved in FASTA-formatted text files for further applications.

3.3 DETERMINATION OF MICROSATELLITE POSITION AND LENGTH

FastPCR program (Kalendar, 2006; http:// www.biocenter.helsinki.fi/bi/bare-1_html/download.htm) was used to analyze the position and the length of microsatellites within ESTs that were saved in text file after sorting the downloaded ESTs. The sequences were copied and pasted into the window of FastPCR and cleaned using the 'clean sequences' option to remove non-nucleotide letters or charcters. The 'repeat search' function was used to search microsatellites. Only 'simple' was checked under 'type of repeats' and all the other variables were left as default. About 1000 sequences were processed at a time with system having pentium IV 1.7 GHz processor and 256 Mb RAM. After each search, results in text form were saved and microsatellite length and positions within the sequence were recorded, maintaing the identity of each sequence.

3.4 IDENTIFICATION OF UNIGENE EST SETS

Redundancy in downloaded ESTs having microsatellite loci was established through clustering to get a set of unigene sequences. The ESTs were clustered using 'Contig Express', a sequence assembler tool in Vector NTI 9.1 Windows version (Invitrogen Inc.,). The ESTs were clustered in batches of 1000 sequences with criteria of at least 20 bases overlap and 85 per cent identity between one end of a read and another end of the other read. The contigs and singletons generated by 'Contig Express' were saved as unigenes in FASTA-formatted files for primer design (Plate 1).

3.6 PRIMER DESIGN FOR EST-SSRs

The primer pairs were designed to the EST sequences having a microsatellite using online primer designing software, Primer3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi). Each of the SSR source sequence was entered specifying the

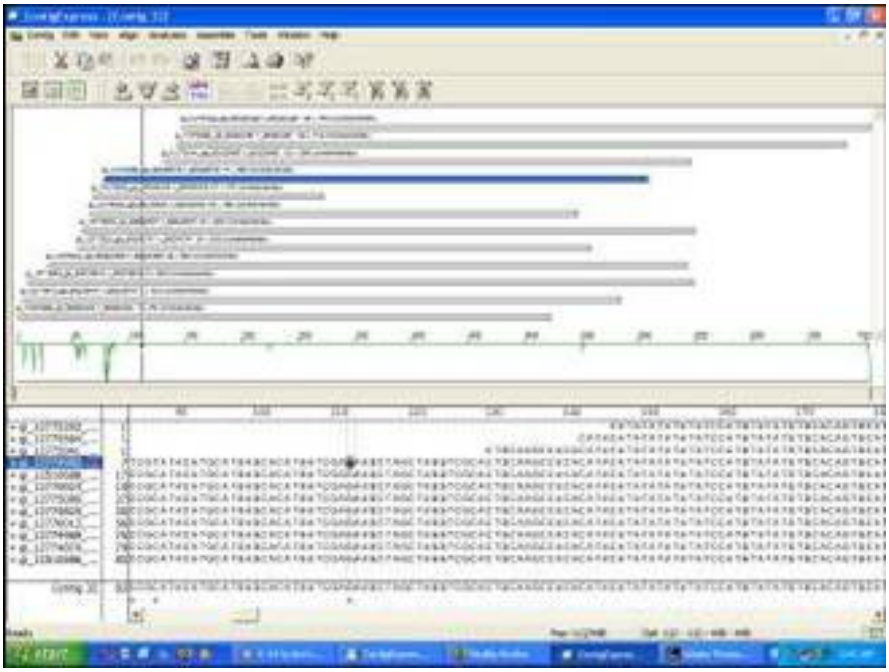

Plate 1: Contig identified using 'Contig Express'. Twelve ESTs from sorghum assembled into a single consensus sequence of 750 bp. Green line below the bars represent 'Contig Weight graph' representing points of nucleotides disagreement with consensus sequence. The highlighted nucleotide in the alignment pane is a disagreeing nucleotide corresponding to the back line in graphic pane

target regions to be amplified by PCR. EST-SSRs were selected for primer design with the following size restrictions:

DNRs or TNRs ≥ 18 bp,
TTNRs ≥ 16 bp,
Edge sequence ≥ 50 bp (surrounding SSRs).

To allow for possible use of primer, the parameters used for the Primer3 program were:

Optimal T_m of 60°C,
T_m range 57-65°C,
GC content 30 to 70%,
PCR product range 100-300 bp,
Low chance of primer dimer and hairpin loops.

3.7 CONSTRUCTION OF REPEATS EST DATABASE FOR SORGHUM

The relational database that catalogues the information about the microsatellite repeats of sorghum is named as 'Jowar GenRepeat Database' was done in Microsoft Access (Office XP version). The database was designed to store three kinds of data: the microsatellite repeats found in sorghum, annotated ESTs information and the primers developed for these microsatellites. Each entry in the sequence information table includes an identifier for the sequence type called 'iabt' ID and the accession number to which it belongs. The accession number is a number, possibly with a few characters in front, that uniquely identifies the sequence stored in the flat-file format given by NCBI. The sequence information includes GenBank id, source organism, sequence end type, tissue/ stress condition and complete EST sequence. The information on the repeats includes; type of repeats (di, tri or tetra), starting position of the repeat, repeat length and repeating unit. The annotation information includes the similarity found with other sequences in the existing 'nr' database of NCBI. The ESTs placed in their respective sequence groupings were classified according to their library type. In addition to the sequence information mentioned above, the database also included a list of primers designed using Primer3. The database forms an integrated platform,

providing the user everything about sorghum microsatellites and microsatellite-based markers.

3.8 DATABASE QUERY

The user-friendly interface for the database was developed using switch board manager option in MS Access, with the comprehensive and integrated Jowar GenRepeat database, the user can query for microsatellites, using the repeat type/motif and the number of repeats. The query results are displayed in a columnar format, showing complete information of the sequence and repeats. User can also query primers or for sequences based on its possible function or any field such as Gene ID, GenBank ID, stress condition, species or repeats. There is an option to convert the results to either Microsoft word or Microsoft Excel for further analysis.

3.9 HOMOLOGY BASED FUNCTIONAL ANNOTATION

BLAST similarity comparison was used to reveal possible functional relationship of each EST containing microsatellite. BLAST search was performed by using BLASTx program against the nonredundant (nr) database. An E-value cut off was used as criteria to asign tentative identity to any sequence. BLASTx search results were visually inspected to ensure that the sequence similarity was contiguious before conclusion was made.

The comparison was also done against cured, highly annotated, proteins in SWISS-PROT database, which contains proteins of demonstrated function. Upon close scrutiny, the EST sequences with function were classified based on their role in any organism.

3.10 VALIDATION OF EST-SSRs

Inorder to bio-validate genic SSRs identified in this study, sorghum RIL population derived from cross IS22380 (P1) and E36-1 (P2) was used. Out of 520 genic SSR primers developed, a random subset of 20 were custom synthesised at Sigma-Aldrich pvt. Ltd., USA, were used to screen the parents and RIL population.

3.10.1 PCR reaction and conditions for the amplification of genic SSRs

The PCR conditions employed were:

Components	*Concentration*	*Quantity (µl)*
Genomic DNA template	5.0 ηg/ml	02.00
dNTP mix – eppendorf.	2.5 mM	01.00
PCR assay buffer - in-house	10 x	02.50
Deionised distilled water	-	17.17
Forward primer	5.0 ρM/µl	01.00
Reverse primer	5.0 ρM/µl	01.00
Taq polymerase	In-house	00.33
	Total	25.00

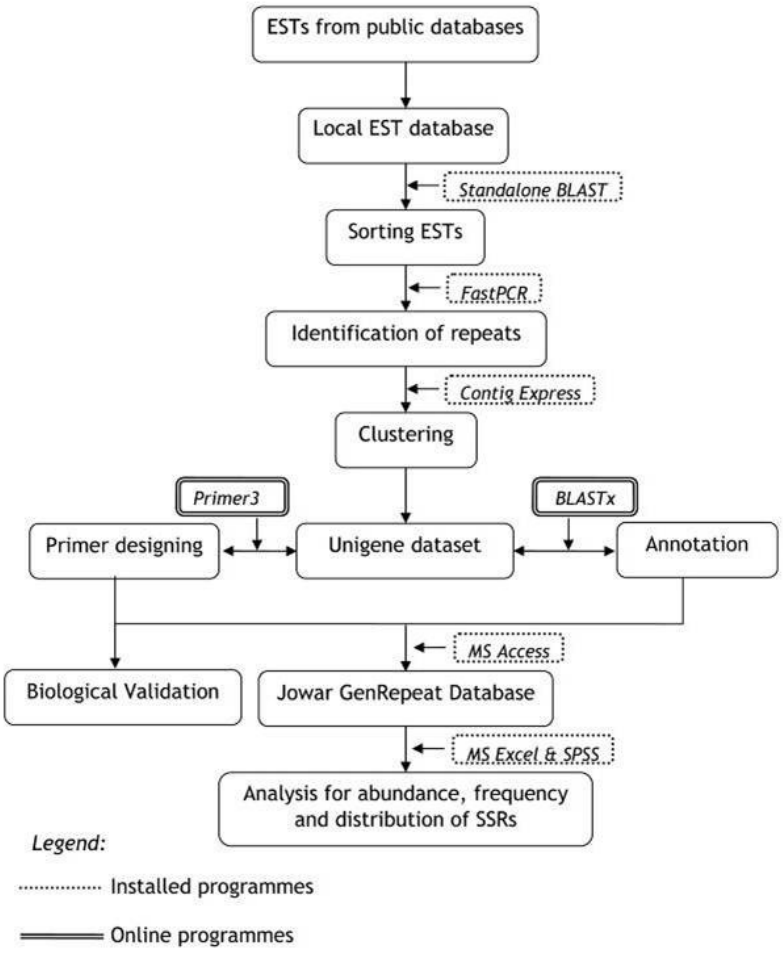

Fig. 1: Scheme for the development of genic SSRs from EST sequences.

Fig 1: Scheme for the development of genic SSRs from EST sequences

Thermal cycling

PCR reaction was carried out using Master Cycler gradient 5331-Eppendorf version 2.30. 31-09, Germany. The cycler was programmed as under,

Step	*Temperature*	*Duration*	*No. of Cycles*
Initial denaturation	94°C	5 min	1
Denaturation	94°C	1 min	44
Primer annealing	55°C	1 min	
Primer extension	72°C	2 min	
Complete primer Extension	72°C	5 min	1
Hold	4°C		Until removed

After the completion of PCR, the products were stored at 4°C until the gel electrophoresis was done.

3.10.2 Separation of PCR products for SSR

PCR products were separated and visualized both on agarose (2 per cent) as well as polyacrylamide gels (6 per cent). Agarose gels were used only for visualization of amplification and allele sizing of PCR amplified microsatellite products was done in denaturing polyacrylamide gels.

Polyacrylamide gel electrophoresis

Six per cent polyacrylamide gels were used for separation and visualization of PCR amplified microsatellite products. Denaturing gels were cast in Sequi-Gen GT nucleic acid electrophoresis cell (Biorad Ltd) as per the protocol in the manual of MRF, Hyderabad.

Glass plates were prepared before making the gel solution. Both outer (large) glass plate (IPC unit) and inner (small) glass plate were cleaned thoroughly with warm water and detergent and then washed with deionised water.

Fresh binding solution was prepared in fumehood by adding 4 µl of bindsilane to 1 ml of 0.5 per cent acetic acid in 95 per cent ethanol in a 1.5 ml micro-centrifuge tube. Mixture was poured on notched plate (inner glass plate) and spread using tissue paper over the entire surface. Treated side was markated. Similarly, repel silane (250 µl) was added 750 µl 0.5 acetic acid in 95 per cent ethanol in a 1.5 ml microcentrifuge tube. Mixture was poured on large glass plates and spread using tissue paper.

Spacers (0.4 mm) were placed along the side edges of the bind silane treated surface. Large plate was put on small plate so that treated surfaces faced each other and clamped on both sides and the assembly was placed in the precision caster base for sealing both sides and the bottom of the cast.

Polyacrylamide gel (6 per cent; 80 ml) was freshly prepared and just prior to pouring, 60 µl of TEMED (N, N, N'N'-tetramethylethylenediamine) and 600 µl of 10 per cent APS (ammonium per sulphate) was added to initiate the polymerization process. The contents were mixed gently by swirling to avoid bubbles. Before pouring the assembly was kept on the bench top at 45º, the assembly tilted to raise one of the bottom corners and then the solution was carefully poured into the space between the glass plates, starting at the lower corner. As the acrylamide solution filled the space, gel assembly was lowered so that both bottom corners were on the bench, parallel to the bench top.

Shark toothcomb (0.4 mm, 49 wells) was inserted with straight side facing the gel at the top of the gel. If bubbles formed during pouring, they were dislodged by tapping. Gel was left for 20-40 min for complete polymerization.

Electrophoresis

1. After the polymerization process, assembly was detached from the clamp and precision caster base and it was placed in universal base against the back wall. IPC was locked to the base in vertical position by fitting stabilizer bar.
2. TBE (5x) was poured in upper tank of the IPC unit and rest was poured in the bottom chamber (1.8 L of buffer was prepared fresh, each time).
3. Comb was removed and then excess polyacrylamide gel was removed with razor blade. Tissue paper was used to clean the glass plates with buffer.
4. Air bubbles and unpolymerised acrylamide on top of gel were removed by squirting with 5× TBE.
5. Pre-run was given to achieve gel surface temperature of approximately 45 to 50 ^{0}C with following conditions. Temperature 50 ^{0}C, power 2000 V, 50 mA and 75 W run for 1 hr.
6. SSR loading dye (3x STR dye) was added to PCR products to a final of 1× and samples were denatured by heating to 95 ^{0}C for 4 min and immediately cooled on ice.
7. After the pre-run, urea was flushed from the well area using a transfer pipette and the shark tooth comb was inserted into the gel so that the teeth were just touching the surface of the gel. Care was taken to avoid piercing of the gel too deeply.
8. Samples (6 µl) were loaded into the wells for 100 bp marker was also loaded into the first or last well after denaturing.
9. Gel was run using the same conditions as in the pre-run step. Untill the dye reached the bottom of the gel.

Visualization of SSR bands

After electrophoresis, clamps were loosened and buffer was removed. Glass plates were separated using plastic wedge at the right corner. The gel affixed to small glass plate was stained to visualize the DNA fragments with the following staining protocol. The technique was followed with in-house component solutions prepared in separate containers.

Silver staining

1. Gel was rinsed with distilled water for 3-5 min and placed in a shallow plastic tray and it was soaked in 2 L of 2 per cent acetic acid (fix solution) for 20 min.
2. The gel was rinsed with water 2 times, each with 2 min and it was stained 2 L of 1 per cent silver nitrate for 20 min.
3. A quick water wash was given for 10-15 sec.
4. Developer solution was added to the tray and agitated until the bands appeared.
5. Developer was removed and plate was placed in fixer or stop solution for 5 min.
6. Gel was placed in 2 L of impregnate solution for 15 min.
7. Lastly, gel was given water wash for 5 min and kept for drying overnight.

All steps were done with constant shaking conditions. Each solution was prepared afresh and they were used only 4 times, over aperiod of 48 hours. The images on the gels were scanned and documented for further use.

IV EXPERIMENTAL RESULTS

The present study was conducted to identify genic SSRs from sorghum ESTs, determine their biology and functionally annotate them. The result of various analyses and processes leading to EST-SSR database for sorghum and validation of some EST-SSR loci is presented here and the outline of which is presented in Figure 2.

4.1 NATURE OF SORGHUM ESTs IN PUBLIC DATABASE:

All available ESTs of sorghum were downloaded from public databases; NCBI and Fungen to study their narure and structure. The search terms (Sorghum AND EST) resulted in 2,42,656 EST entries, adding up to 140 M bp. The total number reduced to 2,32,921, after the initial search result was subjected to sorting by standalone BLAST programme in MS-DOS interface. The sorghum ESTs were available in 35 different classes, depending on conditions imposed or tissues used to derive them. These sequences were reclassified according to different species of sorghum *viz., S. bicolor* (89.53 %)*, S. proponiquum* (9.57 %) *and S. halepense* (0.87 %). The details on distrubution of ESTs to different groups is given in Table 5

4.2 DETERMINATION OF MICROSATELLITE POSITION AND LENGTH

All thirty five data sets of ESTs from different sources were found to have ESTs containing repeats. A total of 12,235 ESTs recorded different types and length of repeats. The output of the programme gave the repeat unit, length of repeat and their exact location. The ESTs having repeats in them ammounted to 5.25 per cent of the total ESTs in a redundant dataset, but 28.62 per cent of the non redundant ESTs had SSRs. The share of repeats across different cateogories is represented in Table 6.

4.3 CLUSTER ANALYSIS OF ESTs CONTAINING MICROSATELLITES

Among the total ESTs in databases there were 12,235 microsatellite-containing ESTs. 10,436 belonged to 1,482 contigs, and they represented genes whose transcripts were sequenced more than once. The remaining 1,799 ESTs were singletons, which represented genes whose transcripts were sequenced only once. Thus, a total of 3,281 unique genes containing microsatellites were identified, which constitute for 1.41 per cent of the total ESTs, in sorghum EST libraries. Clustering analysis resulted in the identification of 26.82 per cent of unigens with repeats.

Irresepective of source library/ transcriptome, overall redundancy was 73.18 per cent. Transcriptomes varied from 0-83.31 per cent redundancy. Transcriptomes such as pooled green leaves, root tissues, protoplast, leaves, RPH region and greenbug aphid infested recorded zero redundancy. The transcriptomes like aerobic roots (83.31 %), wounded leaves (82.28 %) GA or Brassinoloide treated (81.91 %) and ovary (79.44 %) recorded the highest redundancy across sorghum species (Table 6). The share of repeats in all the three datasets across libraries is presented in Figure 3.

4.4 DISTRIBUTION OF MICROSATELLITE TYPES

TNRs were the most abundant among the sorghum ESTs, which accounted for 50.37 per cent of all ESTs containing SSRs, followed by 42.9 and 6.72 per cent, respectively, for DNRs and TTNRs (Table 7, Figure 4). It was interesting to note that wherever the proportion DNRs was high the proportion of TNRs was low. However no such relationship was observed for TTNRs with either DNRs or TNRs.

Striking differences in the proportion DNRs, TNRs and TTNRs was obvious among the transcriptomes analyzed. Pooled green leaves and root tissues were found to contain only DNRs, while greenbug aphid infested tissue recorded only TNRs. The relative abundance of

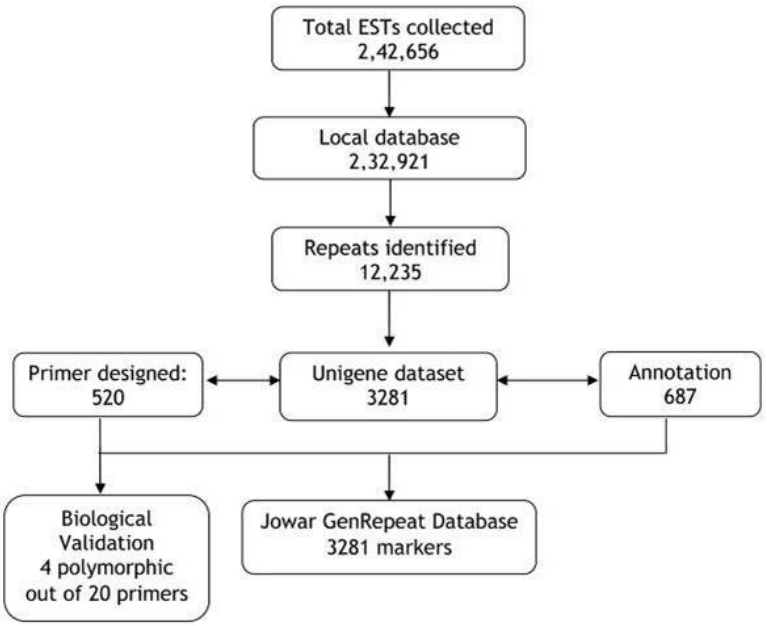

Fig. 2: Schematic representation of results of various steps in the development of genic SSRs from EST sequences.

Fig 2: Schematic representation of results of various steps in the development of genic from EST sequences

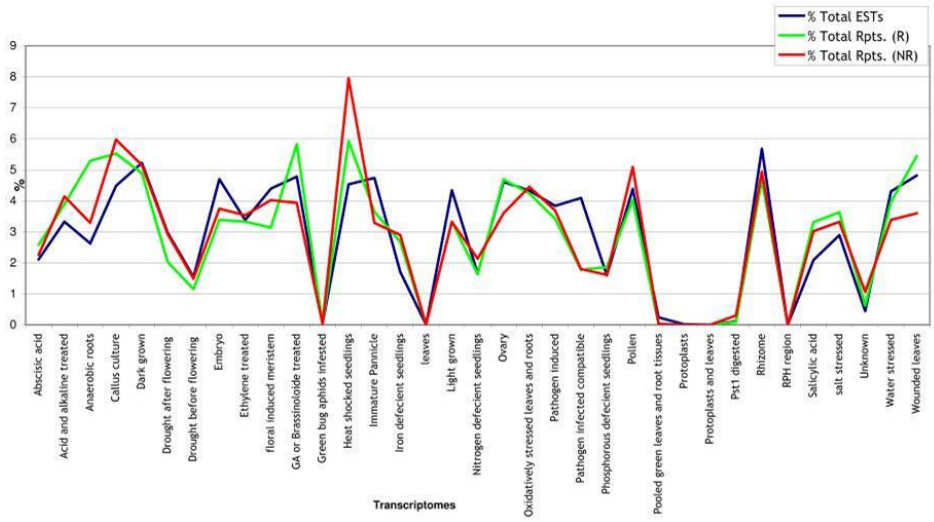

Fig. 3: Percentage of ESTs with repeats in different datasets of sorghum ESTs

Fig 3. Percentage of ESTs with repeats in different datasets of sorghum

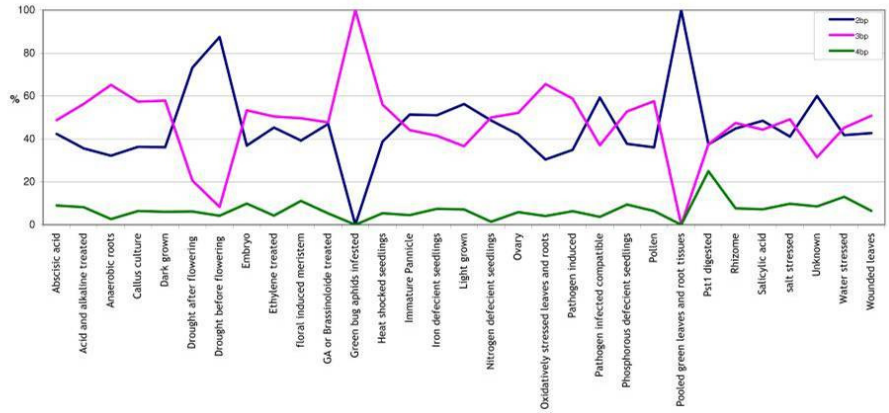

Fig. 4 : Percent share of DNR, TNR and TTNRs in different datasets of sorghum ESTs

Fig 4. Percent share of DNR, TNR and TTNRs in different datasets of sorghum ESTs

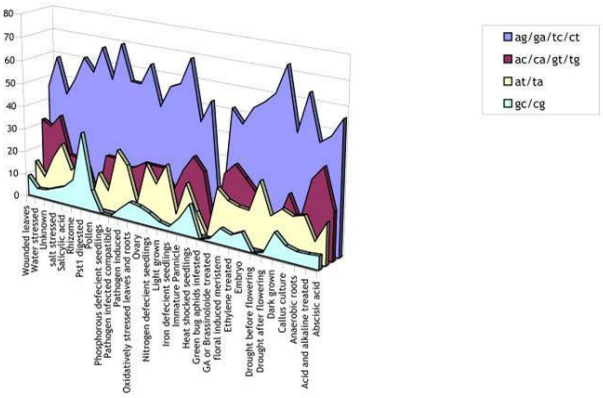

Fig 5. DNR motif distribution across various categories of ESTs

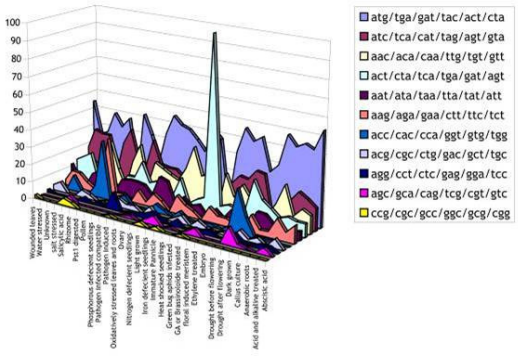

Fig. 6: TNR motif distrubution across various categories of ESTs

Fig 6. TNR motif distribution across various categories of ESTs

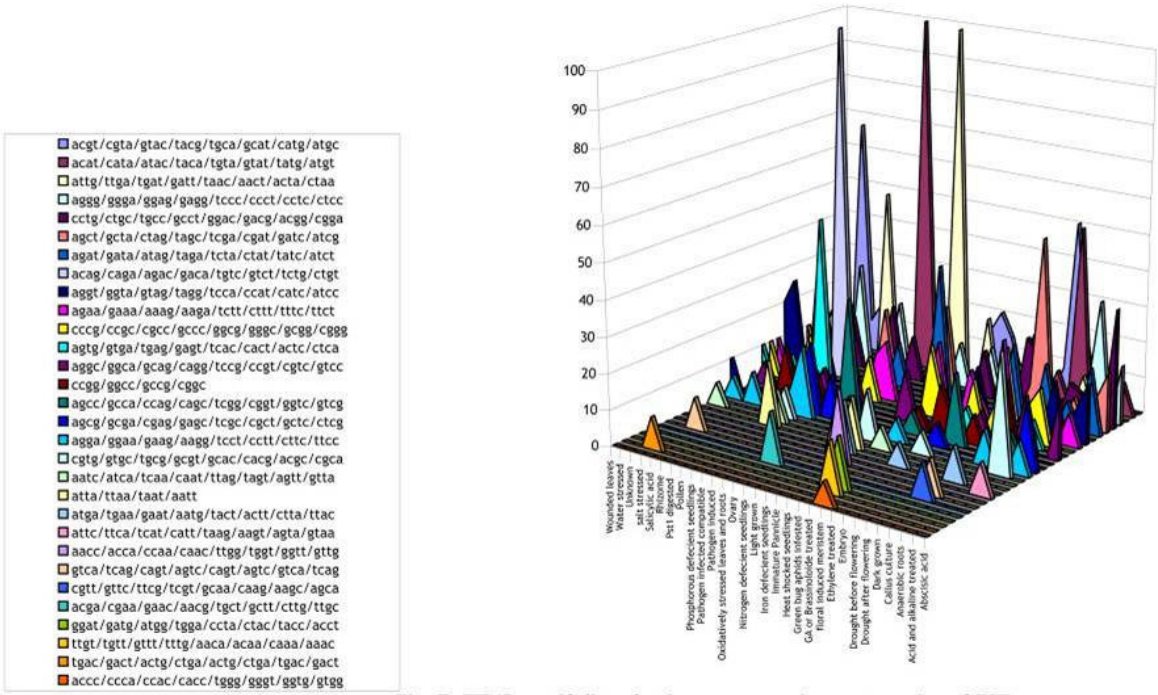

Fig. 7: TTNR motif distrubution across various categories of ESTs

Fig 7. TTNR motif distribution across various categories of ESTs

Table: 5 Distribution of available ESTs among different transcriptomes across three species of sorghum in NCBI database.

Sl. No.	*Transcriptomes*	*S. bicolor*	*S. halepense*	*S. proponiquum*	*Total*
1.	Abscisic acid	4912	0	0	4912
2.	Acid and alkaline treated	7744	0	0	7744
3.	Anaerobic roots	6113	0	0	6113
4.	Callus culture	10449	0	0	10449
5.	Dark grown	12160	0	0	12160
6.	Drought after flowering	6925	0	0	6925
7.	Drought before flowering	3597	0	0	3597
8.	Embryo	10933	0	0	10933
9.	Ethylene treated	7875	0	0	7875
10.	Floral induced meristem	0	0	10238	10238
11.	GA or Brassinoloide treated	11134	0	0	11134
12.	Green bug aphids infested	79	0	0	79
13.	Heat shocked seedlings	10558	0	0	10558
14.	Immature Pannicle	11042	0	0	11042
15.	Iron defecient seedlings	3984	0	0	3984
16.	Leaves	7	0	0	7
17.	Light grown	10116	0	0	10116
18.	Nitrogen defecient seedlings	3849	0	0	3849
19.	Ovary	10735	0	0	10735
20.	Oxidatively stressed leaves and roots	10086	0	0	10086
21.	Pathogen induced	8937	0	0	8937
22.	Pathogen infected compatible	9533	0	0	9533
23.	Phosphorous defecient seedlings	3723	0	0	3723
24.	Pollen	10212	0	0	10212
25.	Pooled green leaves and root tissues	560	0	0	560
26.	Protoplasts	37	0	0	37
27.	Pooled protoplasts and leaves	3	0	0	3
28.	Pst1 digested	267	0	0	267
29.	Rhizome	0	1510	11715	13225
30.	Root specific	1	0	0	1
31.	Salicylic acid	4868	0	0	4868
32.	Salt stressed	6737	0	0	6737
33.	Unknown	119	528	393	1040
34.	Water stressed	10023	0	0	10023
35.	Wounded leaves	11220	0	0	11220
Total		*208537*	*2038*	*22346*	*232921*
		(89.53 %)	*(9.57 %)*	*(0.87 %)*	*(100%)*

Table 6: Frequency of ESTs in different categories.

Sl. No.	Classes of EST set	ESTs			% of ESTs						
		Sequences (1)	Repeats (unclustered) (2)	Repeats (clustered) (3)	(1)	(2)	(3)	(2) in (1)	(3) in (1)	(3) in (2)	Redundancy
1	Abscisic acid	4912	315	74	2.11	2.57	2.26	0.14	0.0318	0.60	76.51
2	Acid and alkaline treated	7744	475	136	3.32	3.88	4.15	0.20	0.0584	1.11	71.37
3	Anaerobic roots	6113	647	108	2.62	5.29	3.29	0.28	0.0464	0.88	83.31
4	Callus culture	10449	676	196	4.49	5.53	5.97	0.29	0.0841	1.60	71.01
5	Dark grown	12160	598	169	5.22	4.89	5.15	0.26	0.0726	1.38	71.74
6	Drought after flowering	6925	248	96	2.97	2.03	2.93	0.11	0.0412	0.78	61.29
7	Drought before flowering	3597	141	49	1.54	1.15	1.49	0.06	0.0210	0.40	65.25
8	Embryo	10933	414	123	4.69	3.38	3.75	0.18	0.0528	1.01	70.29
9	Ethylene treated	7875	407	116	3.38	3.33	3.54	0.17	0.0498	0.95	71.50
10	Floral induced meristem	10238	384	132	4.40	3.14	4.02	0.16	0.0567	1.08	65.63
11	GA or Brassinoloide treated	11134	713	129	4.78	5.83	3.93	0.31	0.0554	1.05	81.91
12	Green bug aphids infested	79	1	1	0.03	0.01	0.03	0.00	0.0004	0.01	0.00
13	Heat shocked seedlings	10558	726	261	4.53	5.93	7.95	0.31	0.1121	2.13	64.05
14	Immature Pannicle	11042	446	108	4.74	3.65	3.29	0.19	0.0464	0.88	75.78
15	Iron defecient seedlings	3984	327	95	1.71	2.67	2.90	0.14	0.0408	0.78	70.95
16	Leaves	7	0	0	0.00	0.00	0.00	0.00	0.0000	0.00	0.00
17	Light grown	10116	407	109	4.34	3.33	3.32	0.17	0.0468	0.89	73.22
18	Nitrogen defecient seedlings	3849	199	70	1.65	1.63	2.13	0.09	0.0301	0.57	64.82
19	Ovary	10735	574	118	4.61	4.69	3.60	0.25	0.0507	0.96	79.44
20	Oxidatively stressed leaves and roots	10086	519	146	4.33	4.24	4.45	0.22	0.0627	1.19	71.87
21	Pathogen induced	8937	418	121	3.84	3.42	3.69	0.18	0.0519	0.99	71.05
22	Pathogen infected compatible	9533	217	59	4.09	1.77	1.80	0.09	0.0253	0.48	72.81
23	Phosphorous defecient seedlings	3723	228	53	1.60	1.86	1.62	0.10	0.0228	0.43	76.75
24	Pollen	10212	493	167	4.38	4.03	5.09	0.21	0.0717	1.36	66.13
25	Pooled green leaves and root tissues	560	1	1	0.24	0.01	0.03	0.00	0.0004	0.01	0.00
26	Protoplasts	37	0	0	0.02	0.00	0.00	0.00	0.0000	0.00	0.00
27	Pooled protoplasts and leaves	3	0	0	0.00	0.00	0.00	0.00	0.0000	0.00	0.00
28	*Pst*1 digested	267	12	10	0.11	0.10	0.30	0.01	0.0043	0.08	16.67

29	Rhizome	13225	573	162	5.68	4.68	4.94	0.25	0.0696	1.32	71.73
30	Root specific	1	0	0	0.00	0.00	0.00	0.00	0.0000	0.00	0.00
31	Salicylic acid	4868	405	99	2.09	3.31	3.02	0.17	0.0425	0.81	75.56
32	Salt stressed	6737	444	109	2.89	3.63	3.32	0.19	0.0468	0.89	75.45
33	Unknown	1040	76	35	0.45	0.62	1.07	0.03	0.0150	0.29	53.95
34	Water stressed	10023	485	111	4.30	3.96	3.38	0.21	0.0477	0.91	77.11
35	Wounded leaves	11220	666	118	4.82	5.44	3.60	0.29	0.0507	0.96	82.28
	Total	**232922**	**12235**	**3281**	**100**	**100**	**100**	**5.25**	**1.41**	**26.82**	**73.18**

Table 7 Proportion of repeat types across cDNA sets

	Frequency			*Percentage*		
Transcriptomes	*DNRs*	*TNRs*	*TTNRs*	*DNRs*	*TNRs*	*TTNRs*
Abscisic acid	33	38	7	42.31	48.72	8.97
Acid and alkaline treated	48	76	11	35.56	56.30	8.15
Anaerobic roots	36	73	3	32.14	65.18	2.68
Callus culture	74	117	13	36.27	57.35	6.37
Dark grown	60	96	10	36.14	57.83	6.02
Drought after flowering	71	20	6	73.20	20.62	6.19
Drought before flowering	42	4	2	87.50	8.33	4.17
Embryo	45	65	12	36.89	53.28	9.84
Ethylene treated	53	59	5	45.30	50.43	4.27
Floral induced meristem	53	67	15	39.26	49.63	11.11
GA or Brassinoloide treated	62	63	7	46.97	47.73	5.30
Green bug aphids infested	0	1	0	0.00	100.00	0.00
Heat shocked seedlings	101	146	14	38.70	55.94	5.36
Immature Pannicle	57	49	5	51.35	44.14	4.50
Iron defecient seedlings	48	39	7	51.06	41.49	7.45
Light grown	63	41	8	56.25	36.61	7.14
Nitrogen defecient seedlings	35	36	1	48.61	50.00	1.39
Ovary	50	62	7	42.02	52.10	5.88
Oxidatively stressed leaves and roots	45	97	6	30.41	65.54	4.05
Pathogen induced	44	74	8	34.92	58.73	6.35
Pathogen infected compatible	32	20	2	59.26	37.04	3.70
Phosphorous defecient seedlings	20	28	5	37.74	52.83	9.43
Pollen	62	99	11	36.05	57.56	6.40
Pooled green leaves and root tissues	1	0	0	100.00	0.00	0.00
Pst1 digested	3	3	2	37.50	37.50	25.00
Rhizome	70	74	12	44.87	47.44	7.69
Salicylic acid	47	43	7	48.45	44.33	7.22
Salt stressed	46	55	11	41.07	49.11	9.82
Unknown	21	11	3	60.00	31.43	8.57
Water stressed	48	52	15	41.74	45.22	13.04
Wounded leaves	53	63	8	42.74	50.81	6.45
Total	1423	1671	223	42.90	50.37	6.72

Table 8 Test of significance for segregation of genic SSR markers in RILs of IS22380 x E36-1

Sl. No.	Marker	P_1 allele	P_2 allele	$\chi 2$ (Calculated)
1	labtgs 1	47	46	0.011
2	labtgs 2	37	51	2.227
4	labtgs 9	43	49	0.391
3	labtgs 20	39	54	2.419

Table χ^2 at 1 degrees of freedom = 3.841

Plate 2. IS22380 and E36-1 showing polymorphism for 20- genic SSR primer pairs in 2% agarose gel for 20 primers

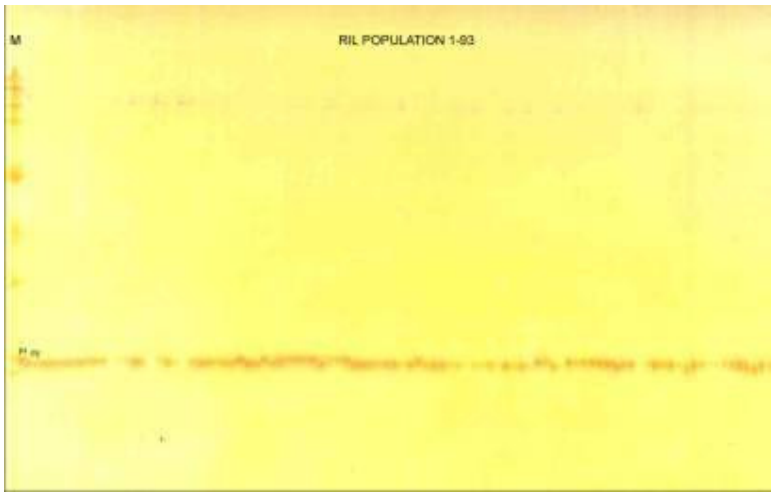

Plate 3. Genic SSR primer (IABTGS2) showing polymorphism across RIL population derived from IS22380 X E36-1

various DNR, TNR and TTNR motifs across different transcriptomes is presented in Appendix II.

Of the DNRs, AG/CT was the most abundant type, accounting for 55.23 per cent of all DNRs across different transcriptomes. AC/GT was the second most abundant DNR type (21.94 per cent). The AT repeats were lower at 16.17 per cent, while the CG repeat was rare (6.63 per cent). Among TNRs the most abundant type was ATG/TAC followed by ATC/TAG. These two types accounted for 52.12 per cent of all TNRs found. AAC/TTG and ACT/TGA were each a little over 14 per cent. All other types of trinucleotide repeats togather were less than 20 per cent and the abundant TTNRs were ACGT/TGCA (9.11%), ACAT/TGTA (8.14%), ATTG/TAAC (7.96%), AGGG/TCCC (6.96%) and CCTG/GGAC (6.35%). The less abundant TTNRs were AGAT/TCTA (5.98%), ACAG/TGTC (4.90%), AGGT/TCCA (4.89%), AGAA/TCTT (4.22%) and CCCG/GGGC (4.10%). The remaining 19 types of TTNRs togather represented only 28.56 per cent of all TTNRs. (Appendix II). The trend of individual DNRs, TNRs and TTNRs across various transcriptome libraries is shown in Figure 5, 7 and 8 respectiviely.

4.5 DESIGN OF PRIMER PAIRS FOR EST-SSRs

To determine the usefulness of the microsatellites, FastPCR program was used to locate the microsatellites within each EST sequence and to determine if primers could be designed to obtain detecteable amplicon. With objective the ESTs having minimum of 150 to 300 bp sequence on either side of the repeat unit was considered as minimum for primer design. Out of the 3,281 ESTs containing SSRs, only 520 had desired length of repeat with enough flanking sequences for primer design, while 1988 ESTs had microsatellite repeats at either the beginning or at the end of EST sequences and remaing did not have suffecient length of repeat motif to consider for primer design. For all the 520 optimal ESTs containing SSRs, primer pairs were designed using primer3 programme.

4.6 VALIDATION OF SSRS

A subset of 20 randomly selected primer pairs for genic SSRs designed as a result of this study were considered for validation. The parental lines, IS22380 and E36-1 were surveyed against these probable genic SSR markers to detect polymorphism and to identify markers by genotyping RILs. All the 20 pairs of primers produced the expected amplicon, of which four showed polymorphism (Plate 2). This four polymorphic primer pairs were then used to genotype individual RILs of a sorghum mapping population (Plate 3). The RIL

population derived from IS22380 x E36-1 is routinely used in our laboratory for mapping studies. The polymorphic markers segregated in the expected 1:1 proportion (Table 8).

4.7 REPEAT EST DATABASE FOR SORGHUM

The database constructed for the EST-SSRs was named as 'Jowar GenRepeat database' where 'GenRepeat' is for genes (ESTs) with repeats. It has information related to EST sequences containing repeats. All the 3,281 ESTs with SSRs were stored in the database with all related features *viz.,* information about the microsatellite repeats of sorghum (type, size, position and length), annotated information (function and E-value score) and primer information (forward and reverse primers). Apart from these, it also contains the reference to source; GenBank ID, source organism, sequence end type, transcriptome and the complete EST sequence. This database is seach enabled using keywords in all fields, updateable and user friendly. Various screen shots of the Jowar GenRepeat database are presented in Plates 4 and 5.

4.8 FUNCTIONAL ANNOTATION

Most of the sorghum ESTs are not assigned to any kind of function. In order to deduce possible function, all the ESTs containing SSRs, were compared with all non-redundant GenBank CDS translations, PDB, SwissProt, PIR and PRF samples (32,97,000 sequences) using BLASTx algorithm (last accessed 07 Aug 2005). Among the 3,281 ESTs containing repeats, putative function could be assigned to 687 ESTs (20.93 per cent) which are listed in Appendix III. Among them, 135 sequences allowed primer design as discribed in subsection 4.5. The class distrubution of EST-SSRs in different transcriptomes and the function itself are presented in the Figure 8 and 9, respectively.

V DISCUSSIONS

Expressed sequence tags (ESTs) are currently the most widely sequenced nucleotide commodity of the plant genomes in terms of the number of sequences and the total nucleotides. EST projects are useful, sequence information; provide a robust sequence resource that can be exploited for gene discovery, genome annotation and comparative genomics. For instance, the availability of complete genome sequence of *Arabidopsis thaliana* revealed that the 105,000 ESTs available at the end of the year 2000 were enough to tag 60% of the 25,500 genes (The Arabidopsis Genome Initiative 2000). As the robot systems increase throughput, cost per read come down. It is now affordable to determine a sequence tag for a large number of genes using random cDNA sequencing approach (Andersen and Lubberstedt, 2003). Combined with breakthroughs in parallel designs for gene expression analysis, large-scale EST projects now offer new perspectives for understanding the molecular basis of important traits in plants of agricultural relevance (Duggan *et al.,* 1999). However, abstract nature of the EST collections, with their high levels of sequence redundancy, low-quality sequence attributes and short sequence lengths have left this enormous sequence collection rather an under-exploited resource (Rudd, 2003).

It is only recently that plant biologists have considered these vast EST datasets to mine the data for novel attributes; *de novo* annotation of the sequences, use of sequences within proteomics-based analysis pipelines and for molecular marker development. Further, there has been incresased interest in the field of expression profiling. By clustering and relating genes on the basis of their expression patterns, they can be assigned to a metabolic pathway, functional or structural complex, or to a co-regulated group. ESTs have a potential here, beyond the basic subtraction methods available earlier.

Several studies have used EST data mining (Temnykh *et al.,* 2001; Kantety *et al.,* 2002; Varshney *et al.,* 2002; Thiel *et al.,* 2003) and traditional approaches (Cagigas *et al.,* 1999; Delghandi *et al.,* 2003; Sekino *et al.,* 2003) for the development of microsatellite markers in different plant species. Earlier reports of datamining in sorghum used less number of ESTs due to lack of their availability at that time (Kantety *et al.,* 2002). The present study, on the development of EST-SSR markers with their tentative function for sorghum using ESTs is the first of its kind. In addition to 3281 ESTs derived microsatellite, a total of 687 gene associated microsatellite associated markers were identified from this study. This accounts for a large number of microsatellites reported from any single study in sorghum. These results suggest that bioinformatic analysis of ESTs is an efficient way of identifying microsatellite markers. Further, the possibility of associating these EST-SSRs with genes allows mapping of genes to physical maps, while microsatellites offer polymorphism and allow them to be positioned on meiotic maps, their nucleotide sequence permits assignment of genes to physical maps. Mapping a common set of markers on both meiotic and physical maps should lay necessary frame work for map integration.

5.1 DATA MINING AND ANALYSIS OF REPEAT PATTERNS

Though of many software packages are availble for bioinformatic analysis, suitability of a specific program for the task of interest have to be tested. In the present study, the Windows-based FastPCR for localization of repeat position and length was used because of ease in transfering output results to spreadsheet for further analysis. Similarly, VectorNTI ContigExpress module for contig assembly was used, which is a module with an intuitive fragment assembly program for medium-size projects of this type. The linear assembly algorithm permitted the analysis of resonably large number of sequences encountered in this study. The advantage of 'Vector- NTI' lies in its user-friendliness and graphical output of contigs, but its capacity is limited to about 5000 sequences for contig assembly. For contig assembly involving larger data sets, Unix-based programs such as 'CAP3' and 'Phrap' should be considered. For primer designing although many programs such as OLIGO, Primer3, PrimerTour, GeneTool, DNASTAR *etc.,* are available, Primer3 provided for designing primers in batchmode with maximum parameter controls.

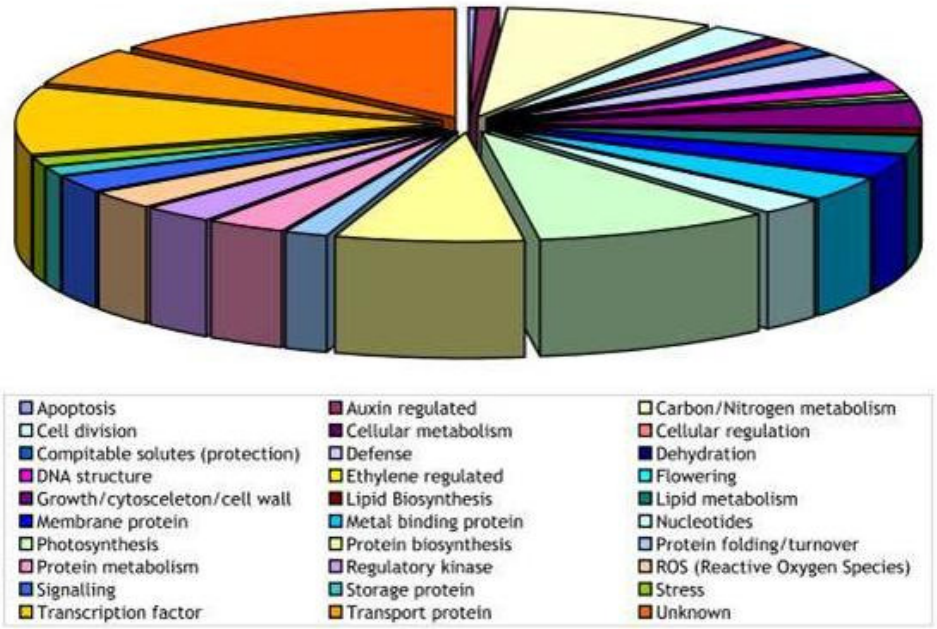

Fig 9. Distribution of Genic SSRs in different transcripomes

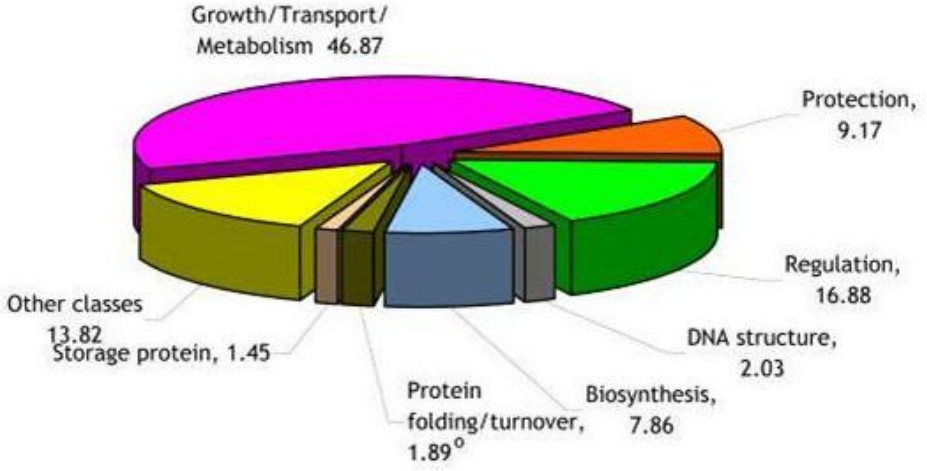

Fig 10. Functional distribution of EST-SSRs

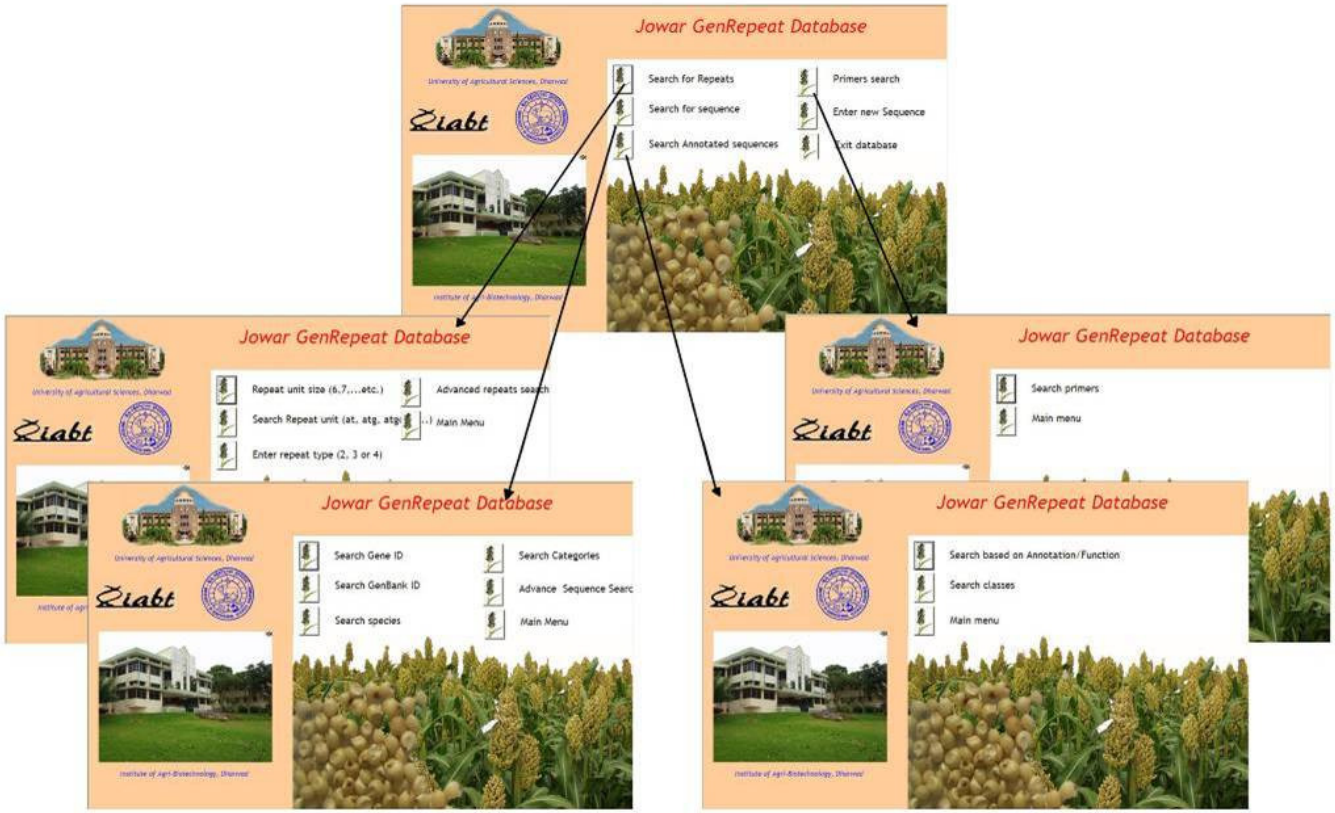

Plate 4. Screen shots of 'Jowar GenRepeat database' showing mainpage and subpages

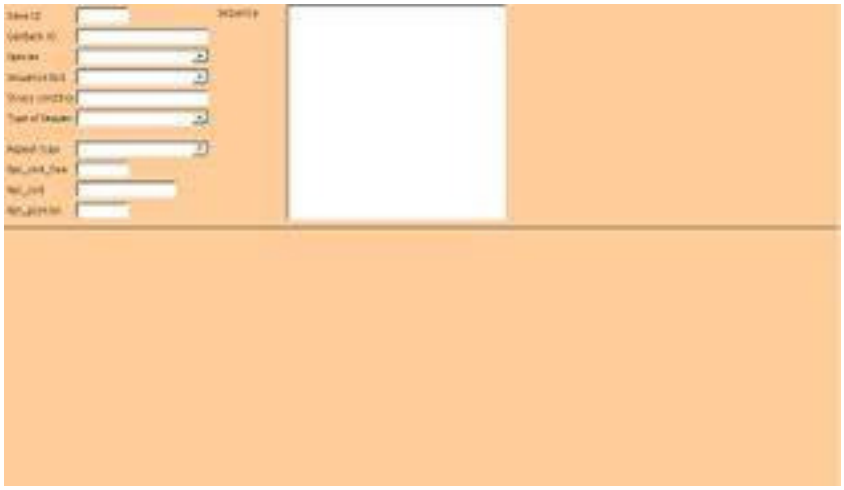

Plate 5a. Updating window for entering newly discovered EST-SSR sequence

Plate 5b. Output format of Jowar GenRepeat database

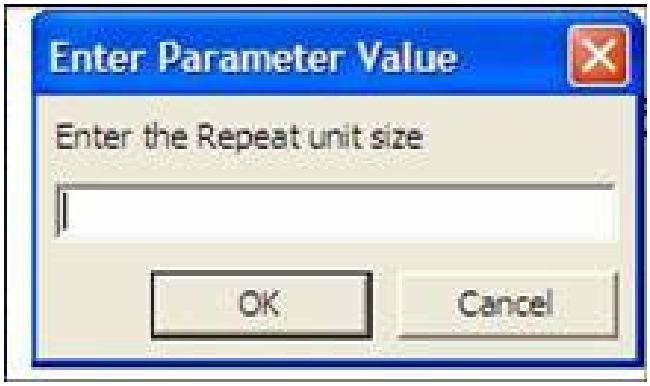

Plate 5c. Search box for entering keyword for querying database

Among the three species, *Sorghum bicolor* had the highest number of repeats owing its importance as a food crop. The proportion of microsatellite ESTs in sorghum was as low as 1.41 per cent which translates into one SSR in 10.4 Kb sequence. The occurence of a microsatellite EST in sorghum is comparable to other cereals like maize 8.1 Kb (Cardle *et al.,* 2000), rice 11.81 Kb (Gao *et al.,* 2003), barley 7.5 Kb, rye 5.5 Kb (Varshney *et al.,* 2002) and wheat 9.2 Kb (Gupta *et al.,* 2003). The proportion of microsatellite-containing ESTs was in range from 1.5-4.7 per cent in these cereals (Kantety *et al.,* 2002). However, it is important to note that the frequency of EST-SSRs detected depends on the criteria used for database mining, level of redundancy or low effeciency of clustering and insuffecient representation of the genome in the database.

ESTs are represented from a variety of situations called transcriptomes. In the public database, the proportion of genic SSRs containing ESTs in each kind of transcriptomes remained the same in different datasets (total ESTs, redundandant repeat containing ESTs and non-redundandant repeat containing ESTs).

TNRs are the most common repeats in sorghum, followed by DNRs and TTNRs, which is similair to the report by Varshney *et al.* (2002). But, DNRs and TNRs dominated when individual transcriptomes of sorghum were considered. The abundance of TNRs has been attributed to the lack of selection against length variation in SSRs present in genes, as it does not cause frameshift mutations (Metzgar, *et al.,* 2000). Further, among the TNRs, codon repeats corresponding to small hydrophilic amino acids are thought to be more easily tolerated. Converesly, selection pressure probably eliminates addition or deletion of codon repeats, encoding hydrophobic and basic amino acids that bring about major changes in the resultant protein (Katti, *et al.,* 2001).

Among the DNRs in sorghum, the motif AG is the most frequent (55.3%) followed by AC (21.94%), which are comparable with all the other cereals except rye, where these frequencies are 50 per cent for AC and 37.9 per cent for AG (Varshney *et al.,* 2002). Similarly, the most infrequent motif in sorghum is CG (6.6%), as observed in other species (1.7-9.0 %) except in barley, where AT is the least frequent (8.4%). A di-nucleotide motif can represent multiple codons depending on the reading frame and translate into different amino acids. For example, the AG/CT motif can represent GAG, AGA, UCU and CUC codons in a mRNA population and translate into the amino acids; Arg, Glu, Ala and Leu, respectively. Ala and Leu are present in proteins at high frequencies of 8 and 10 per cent, respectively (Lewin, 2003). This could be one of the reasons why AG/CT motifs are present at such high frequencies in ESTs.

Among the TNRs, the motif CCG (34.67%) is the most frequent, similar to wheat (32 %), barley (49 %), followed by AGC (17.42%) similar to barley, maize and rice (13-30 %). Prevalance of GC rich TNRs is expected as the GC rich regions are known to be associated with genes. Among TTNRs, abundant tetranucleotide repeats were ACGT/TGCA (9.11%) followed by ACAT/TGTA (8.14%). TTNRs abundance varies with the species and no single motif is known to occur frequently in all cases.

5.5 MARKER ESTs DATABASE

Molecular markers are required in any gene screening approach, from gene-mapping within traditional 'forward genetics' approaches through QTL identification studies to genotyping and haplotyping. As we enter the post-genomics era, the need for genetic markers does not diminish, even in the species with fully sequenced genomes. 'Jowar GenRepeat Database' developed here is a comprehensive dataset of all the sorghum ESTs containg varied degrees of repeats in them. The database consists of 3281 ESTs representing 35 transcriptomes, available in public databases. In addition to the regular features available at NCBI, other features such as annotaion information and primer pairs to amplify the repeat motif etc. have been included. Since the database provides functionality, the user can search

for transcriptome specific markers on the by using any specific criteria. The database is a relational database with different fields linked togather and the contents are readily accessable through a user friendly interface designed in Microsoft Access. The information corresponding to the keywords will be retreived and the output format is flexible with options for converting to either Microsoft Word or Excel for further analysis. The Jowar GenRepeat database established here will remain synchronous with EST collections available in public databases with update facility.

5.2 VALIDATION OF SSRS

A random subset of 20 primer pairs of the 520 primer pairs designed for EST-SSRs were tested for their ability to amplify corresponding fragment from genomic DNA of two test genotypes IS22380 and E36-1. All the 20 primer pairs were functional as they produced expected product length. Thus, the designed primer pairs were within the exon/intron splice sites, preventing other genomic DNA to be amplified as it is shown by Cordeire *et al.* (2001). Of the 20 functional primer pairs, four (20 %) were polymorphic between two genotypes. Polymorphism observed is due to the length varriation in the repeat motif present in the gene. Low rate of polymorhism for a set of genic SSRs in any pair of genotypes is expected. The functional part of the genes to which these ESTs belong are likely to be highly conserved during evolution, so would be the case with repeat motifs within exons.

5.3 FUNCTIONAL ANNOTATION

Computational tools are now regularly used to infer function based upon significant sequence similarity to experimentally verified proteins or putative proteins. These analyses implement BLAST comparisons against non-redundant databases as well as Gene Ontology (GO) annotation. The EST sequences were compared against known databases using these tools. Protein homology searches were performed in order to identify the putative function of the ESTs. Of the 3281 unigenes, 687 had significant matches to corresponding proteins/ genes in the database. A comparison against SWISS-PROT was also performed which is a curated, highly annotated database of 153,871 proteins of demonstrated function. The possible reasons for ‘no matches’ for majority of sequences may be due to truncated open reading frames, which is characteristic of EST sequences or that they represent long untranslated regions, structural RNAs or bonafide proteins which are unique to sorghum at this level of comparison. The sorghum unigenes were further annotated by gene ontology assignment based on the single "best hit" match against the SWISS-PROT database. 687 ESTs with hits to SWISS-PROT had matching GO-Terms. They were grouped in 30 categories by their function and then regrouped into main categories based on function, process and component as criteria. The majority (46.87%) of the ESTs were assigned to growth, transport and metabolism, followed by regulation (16.88%). The other categories included protection (9.17%), biosynthesis (7.80%), DNA structure (2.03%), protein folding (1.89%) and storage protein (1.45%). Thus, majority of the ESTs in the database represented house keeping genes.

The EST-SSRs extracted from the public databases, and partly validated biologically, can be used for genetic mapping in relation to other molecular markers. In addition to the great value of the markers, gene-associated microsatellites are also useful for comparative mapping. One way to do this is to map them by genetically and also BLAST search for them in the other crop genome sequences to compare their genomic locations relative to neighboring genes. Despite unknown identities of most genes, presently, the gene sequences are highly conserved through evolution. BLAST searches of some microsatellite-containing ESTs (with unknown gene identity) to protein sequences will also help in knowing their function.

In future it is planned to

1. Update the database as and when new ESTs are added to the public databases.

2. Syntesize primers pairs for all the EST-SSRs and validate them as markers, and use the polymorphic markers in mapping efforts already initiated in IABT.

VI SUMMARY

The recent progress in high throughput assays and genome sequencing projects have lead to accumulation of rich sequence information in major cereal crops like rice, wheat and sorghum. This surge in genomic information has led to the development of novel and more robust approaches to derive biological utility of this resource. In the present study, ESTs from public database were analyzed for repeats, devloped them as markers and established a database for these markers with additional functional annotation data. The summury of the present investigation is presented here;

1. The sorghum EST resource in public databases with 2,32,921 ESTs represented 35 different transcriptomes. Majority of the ESTs are from *Sorghum bicolor* followed by *S. proponiquum* and *S. halepense*.

2. Repeat scan analysis identified 12,235 ESTs with repeats of varied type and length, which accounted for 5.25 per cent of all the ESTs in a redundant dataset.

3. Of the 12,235 microsatellite-containing ESTs, 10,436 clones fell within 1,482 contigs and remaining 1,799 ESTs were singletons, making up a total of 3,281 unique genes containing microsatellites.

4. TNRs (50.37 %) were the most abundant types, followed by 42.9 and 6.72 per cent, respectively, for DNRs and TTNRs.

5. Among the DNRs, AG/CT was the most abundant type, accounting for 55.23 per cent of all DNRs, followed by AC/GT (21.94 %). The AT repeats were low at 16.17 per cent, while the CG repeats were rare at 6.63 per cent.

6. Among TNRs, most abundant type was ATG/TAC (34.67 %) followed by ATC/TAG (17.42 %). AAC/TTG and ACT/TGA represented a little over 14 per cent of TNRs. All other 19 types of TNRs togather were less than 20 per cent.

7. The frequency of individual TTNRs varied from 9.11 per cent for ACGT/TGCA to 0.24 per cent for ACCC/TGGG.

8. Pairs of primers could be designed for 520 of the 3281 ESTs. Others had repeat motif at the ends of the EST or the repeat motifs were not long enough for the primer design.

9. The parental lines, IS22380 and E36-1 were scanned with a random set of 20 genic SSRs. All the 20 primer pairs produced the expected amplicon from the genomic DNA of both genotypes, of which 4 were polymorphic. Thus the process of ESTs based detection of the SSRs and their biological utility was validated.

10. The database constructed for the EST-SSRs was named as 'Jowar GenRepeat database' with 3281 ESTs with SSRs as records. Each record has complete information about the repeats, functional annotation and primers information in addition to information available in respective source database. This database is compatible for seach, using keywords in all fields, updateable and user friendly.

11. Each EST was compared to annotated proteins in the databases with BLASTx algorithm and tentative function was assigned for 687 (20.93 per cent) of the 3281 ESTs containing repeats. Of this lot of 687 ESTs, primer pairs have been designd for 135 annotated ESTs which can be used for further applications in functional genomics.

VII REFFERENCES

ADAMS, M. D., KELLEY, J. M., GOCAYNE, J. D., DUBNICK, M., POLYMEROPOULOS M. H., XIAO H., MERRIL, C. R., WU, A., OLDE, B., MORENO, R., KERLAVAGE, A. R., MCCOMBIE, W. R. AND VENTER, J. C., 1991, Complementary DNA sequencing: expressed sequence tags and the human genome project. *Science,* **252**:1651-1656.

ANDERSEN, jr. AND LUBBERSTEDT, T., 2003, Functional markers in plants. *Trends in Plant Science,* **8**:554-560.

ANONYMOUS, 2004, Production of food and agriculture organization, quartery bulletin of statistics, pp 211-138.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TRAVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. AND SHERLOCK, G., 2000, Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25***:*25-29.

AYERS, N. M., MCCLUNG, A. M., LARKIN, P. D., BLIGH, H. F. J., JONES, C. A. AND PARK, W. D., 1997, Microsatellites and a single nucleotide polymorphism differentiate apparent amylose classes in an extended pedigree of US rice germplasm. *Theoretical and Applied Genetics,* **94**:773.

BARNES, S., 2002, Comparing Arabidopsis to other flowering plants. *Current Opinion in Plant Biology,* **5:**128–134.

BATEMAN, A., BIRNEY, E., DURBIN, R., EDDY, S. R., HOWE, K. L. AND SONNHAMMER, E. L. L., 2000, The Pfam Protein Families Database. *Nucleic Acids Research,* **28**:23–266.

BEDELL, J. A., BUDIMAN, M. A., NUNBERG, A., CITEK, R. W., ROBBINS, D., JONES, J., FLICK, E., ROHLFING, T., FRIES, J., BRADFORD, K., MCMENAMY, J., SMITH, M., HOLEMAN, H., A. ROE, B., WILEY, G., KORF, I. F., RABINOWICZ, P. D., LAKEY, N., MCCOMBIE, W. R., JEDDELOH, J. A. AND MARTIENSSEN, R. A., 2005, Sorghum Genome Sequencing by Methylation Filtration. *Public Library of Science Biology,* **3**:e13.

BENSON, D. A., MIZRACHI, I. K., LIPMAN, D. J., OSTELL, J. AND WHEELER, D. L., 2006, GenBank. *Nucleic Acids Research,* **34**:D16–D20.

BENSON, G., 1999, Tandem repeats finder: a program to analyse DNA sequences. *Nucleic Acids Research,* **27**:573-580.

BHATTRAMAKKI, D., DONG, J., CHHABRA, A. K. AND HART, G. E., 2000, An integrated SSR and RFLP linkage map of *Sorghum bicolor* (L.) Moench. *Genome*, **43**:988-1002.

BOGUSKI, M. S. AND SCHULER, G. D. 1995. Establishment of a transcript map. *Nature Genetics,* **10:**369–371.

BOGUSKI, M. S., LOWE, T. M. AND TOLSTOSHEV, C. M., 1993, dbEST database for “expressed sequence tags.” *Nature Genetics,* **4**:332–333.

BOUCK, J. W., YU, W., GIBBS, R., AND WORLEY, K., 1999, Comparison of gene indexing databases. *Trends in Genetics,* **15:**159–162.

BREYNE, P. AND ZABEAU, M., 2001, Genome-wide expression analysis of plant cell cycle modulated genes. *Current Opinion in Plant Biology,* **4**:136-142.

CAGIGAS, M. E., VAZQUEZ, E., BLANCO, G., AND SANCHEZ, J. A., 1999, Combined assessment of genetic variability in populations of brown trout (*Salmo trutta* L.) based on allozymes, microsatellites, and rapid markers. *Marine Biotechnology,* **1**:286–296.

CALLEN, D. F., THOMPSON, A. D., SHEN, Y., PHILLIPS, H. A., RICHARDS, R,. I., MULLEY, J. C., AND SUTHERLAND, G. R., 1993, Incidence and origin of “null” alleles in the (AC)n microsatellite markers. *American Journal of Human Genetics,* **52**:922–927.

CARDLE, L., RAMSAY, L., MILBOURNE, D., MACAULAY, M., MARSHALL, D. AND WAUGH, R., 2000, Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics,* **156**:847-854.

CASTELO, A., MARTINS, W. AND GAO, G., 2002, TROLL—tandem repeat occurrence locator. *Bioinformatics*, **18:**634–636.

CHABANE, K., ABLETT, G. A., CORDEIRO, G. M., VALKOUN, J. AND HENRY, R. J., 2005, EST versus Genomic Derived Microsatellite Markers for Genotyping Wild and Cultivated Barley. *Genetic Resources and Crop Evolution,* **52:**903-909.

CHAMBERS, G. K. AND MACAVOY, E. S., 2000, Microsatellites: consensus and controversy. *Comparative Biochemistry and Physiology,* **126**:455-476.

CHEMA, Z. A. AND A. KHALEEQ, 2000, Use of sorghum allelopathic properties to control weeds in irrigated wheat in a semiarid region of Punjab. *Agriculture Ecosystems and Environment,* **79:**105-112.

CHO, Y. G., ISHII, S. T., TEMNYKH, X., CHEN, L., LIPOVICH, S. R., MCCOUCH, W. D., AYRES, P. N. AND CARTINHOUR, S., 2000, Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa* L.). *Theortical Applied Genetics*, **100**:713–722.

COLLINS, N. C., WEBB, C. A., SEAH, S., ELLIS, J. G., HULBERT, S. H., AND PRYOR, A., 1998, The isolation and mapping of disease resistance gene analogs in maize. *Molecular Plant-Microbe Interaction,* **11**:968–978.

CORDEIRO, G. M., CASU, R., MCINTYRE, C. L., MANNERS, J. M. AND HENRY, R. J., 2001, Microsatellite markers from sugarcane (*Saccharum* spp) ESTs across transferable to *Erianthus* and *Sorghum*. *Plant Science,* **160**:1115–1123.

DELGHANDI, M., MORTENSEN, A., AND WESTGAARD, J. I., 2003, Simultaneous analysis of six microsatellite markers in Atlantic cod (*Gadus morhua*): a novel multiplex assay system for use in selective breeding studies. *Marine Biotechnology,* **5**:141–148.

DENLI, A. M. AND HANNON, G. J., 2003, RNAi: an ever-growing puzzle. *Trends in Biochemical Sciences,* **28:**196–201.

DUGGAN, D. J., BITTNER, M., CHEN, Y., MELTZER, P, AND TRENT, J. M., 1999. Expression profiling using cDNA microarrays. *Nature Genetics,* **21**:10–14.

EDWARDS, A., CIVITELLO, H., HAMMOND, H. A. AND CASKEY, C., 1991, T. DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *American Journal of Human Genetics,* **49:**746-756.

EUJAYL, I., SORRELLS, M. E., BAUM, M., WOLTERS, P. AND POWELL, W, 2002, Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. *Theoretical and Applied Genetics,* **104**:399–407.

EUJAYL, I., SORRELLS, M., BAUM, M., WOLTERS, P., AND POWELL, W., 2001, Assessment of genotypic variation among cultivated durum wheat based on EST-SSRS and genomic SSRS. *Euphytica,* **119**:39-43.

FEINGOLD, S., LLOYD, J., NORERO, N., BONIERBALE, M. AND LORENZEN, J., 2005, Mapping and characterization of new EST-derived microsatellites for potato (*Solanum tuberosum* L.). *Theoretical and Applied Genetics,* **111**:456-466.

FOLTA, K. M, STATON, M., STEWART, P. J, JUNG, S., H BIES, D., JESDURAI, C. AND MAIN, D., 2005, Expressed sequence tags (ESTs) and simple sequence repeat (SSR) markers from octoploid strawberry (*Fragaria × ananassa*). *BMC Plant Biology,* **5**:12.

FRASER, L. G., HARVEY, C. F., CROWHURST, R. N. AND DE SILVA, H. N., 2003, EST-derived microsatellites from Actinidia species and their potential for mapping. *Theortical Applied Genetics*, **108**:1010-1016.

FREELING, M., 2001, Grasses as a single genetic system: reassessment. *Plant Physiology,* **125:**1191–1197.

GAO, L. F., JING, R. L., HUO, N. X., LI, Y., LI, X. P., ZHOU, R. H., CHANG, X. P., TANG, J. F., MA, Z. Y. AND JIA, J. Z., 2004, One hundred and one new microsatellite loci derived from ESTs (EST-SSRs) in bread wheat. *Theoretical and Applied Genetics*, 108:1392-400.

GAO, L. F., TANG, J. F., LI, H. W. AND JIA, J. Z., 2003, Analysis of microsatellites in major crops assessed by computational and experimental approaches. *Molecular Breeding*, **12**:245-261.

GORDON, D., ABAJIAN, C. AND GREEN, P., 1998, Consed: a graphical tool for sequence finishing. *Genome Research*, **8**:195-202.

GRAHAM, J., SMITH, K., MACKENZIE, K., JORGENSON, L., HACKETT, C. AND POWELL, W., 2004, The construction of a genetic linkage map of red raspberry (*Rubus idaeus* subsp. *idaeus*) based on AFLPs, genomic-SSR and EST-SSR markers. *Theoretical and Applied Genetics*, **109:**740–749.

GUPTA, P. K. AND VARSHNEY, R. K., 2000, The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. *Euphytica,* **113**:163-165.

GUPTA, P. K., BALYAN, H. S., SHARMA, P. C. AND RAMESH, B., 1996, Microsatellites in plants: a new class of molecular markers. *Current Science,* **70**:45-54.

GUPTA, P. K., RUSTGI, S., SHARMA, S., SINGH, R., KUMAR, N. AND BALYAN, H. S., 2003, Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. *Molecular Genetics and Genomics*, **270**:315-23.

HACKAUF, B. AND WEHLING, P., 2002, Identification of microsatellite polymorphisms in an expressed portion of the rye genome. *Plant Breeding*, **121:**17–25.

HAN, Z. G., GUO, W. Z., SONG, X. L. AND ZHANG, T. Z., 2004, Genetic mapping of EST-derived microsatellites from the diploid *Gossypium arboreum* in allotetraploid cotton. *Molecular Genetics and Genomics*, **272**:308–327.

HANCOCK, W. S., 2002, The importance of databases. *Journal of Proteome Research*, **1**:201.

HEUMANN, K. AND MEWES, H. W., 1996, The Hashed Position Tree (HPT): a suffix tree variant for large data sets stored on slow mass storage devices. In Ziviani, N., Baeza-Yates, A., Guimaraes, G. (Eds.). *Proceedings of the Third South American Workshop on String Processing*, Recife, Brazil , pp. 101–115.

HOEVEN V. R., RONNING C. AND TANKSLEY S. D., 2002, Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. *Plant Cell* **14:**1441–1456.

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. *Molecular Breeding*, **9:**63–71.

HOULGATTE, R., MARRIAGE-SAMSON, R., DUPRAT, S., TESSIER, A., BENTOLILAL, S., LAMY, B. AND AUFFRAY, C., 1995, The Genexpress Index: A resource for gene discovery and the genic map of the human genome. *Genome Research,* **5**:272–304.

HUANG, X. AND MADAN, A., 1999, CAP3: a DNA sequence assembly program. *Genome Research*, **9:**868–877.

IGNACIMUTHU, S. J., 2003, *Basic Bioinformatics*. Narosa publishing house, New Delhi.

JANSEN, R. C. AND NAP, J. P., 2001, Genetical genomics: the added value from segregation. *Trends in Genetics*, **17**:388–391.

KALENDAR, R., 2006, FastPCR: PCR primer design, DNA and protein tools, repeats and own database searches program. URL http:// www. biocenter. helsinki. fi/bi/bare-1_html/fastpcr. html.

KANTETY, R. V., LA ROTA, M., MATTHEWS, D. E. AND SORRELLS, M. E., 2002, Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Molecular Biology*, 48:501-10.

KATTI, M. V., RANJEKAR, P. K., AND GUPTA, V. S., 2001, Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Molecular Biology and Evolution,* **18**:1161–1167.

KHLESTKINA, E. K., THAN, M. H. M., PESTSOVA, E. G., RÖDER, M. S., MALYSHEV, S. V., KORZUN, V. AND BÖRNER, A., 2004, Mapping of 99 new microsatellite-derived loci in rye (*Secale cereale* L.) including 39 expressed sequence tags. *Theoretical and Applied Genetics*, 725–732.

KIKUCHI, S., SATOH, K., NAGATA, T., KAWAGASHIRA, N., DOI, K., KISHIMOTO, N., YAZAKI, J., ISHIKAWA, M., YAMADA, H., OOKA, H., HOTTA, I., KOJIMA, K., NAMIKI, T., OHNEDA, E., YAHAGI, W., SUZUKI, K., LI, C. J., OHTSUKI, K., SHISHIKI, T., OTOMO, Y., MURAKAMI, K., LIDA, Y., SUGANO, S. ,FUJIMURA, T., SUZUKI, Y., TSUNODA, Y., KUROSAKI, T., KODAMA, T., MASUDA, H., KOBAYASHI, M., XIE, Q H., LU, M., NARIKAWA, R., SUGIYAMA, A., MIZUNO, K., YOKOMIZO, S., NIIKURA, J., IKEDA, R., ISHIBIKI, J. ,KAWAMATA, M., YOSHIMURA, A., MIURA, J., KUSUMEGI, T., OKA, M., RYU, R. ,UEDA, M., MATSUBARA, K., KAWAI, J., CARNINCI, P., ADACHI, J., AIZAWA, K. ,ARAKAWA, T., FUKUDA, S., HARA, A., HASHIDUME, W., HAYATSU, N., IMOTANI, K., ISHII, Y., ITOH, M., KAGAWA, L., KONDO, S., KONNO, H., MIYAZAKI, A., OSATO, N., OTA, Y., SAITO, R., SASAKI, D., SATO, K., SHIBATA,

K., SHINAGAWA, A., SHIRAKI, T., YOSHINO, M. AND HAYASHIZAKI, Y., 2003, Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science,* **301:**376–379.

KOTA, R., WOLF, M., MICHALEK, W. AND GRANER, A., 2001, Application of DHPLC for mapping of single nucleotide polymorphisms (SNPs) in barley (*Hordeum vulgare* L.). *Genome,* **44**:523-528.

KUNZLER, P., MATSUO, K. AND SCHAFFNER, W, 1995, Pathological, physiological, and evolutionary aspects of short unstable DNA repeats in the human genome. *Biological Chemistry Hoppe-Seyler,* **4**:201-211.

LAMBLIN, A. F. J., CROW, J. A., JOHNSON, J. E., SILVERSTEIN, K. A. T., KUNAU, T. M., KILIAN, A., BENZ, D., STROMVIK, M., ENDRÉ, G., VANDENBOSCH, K. A., COOK, D. R., YOUNG, N. D. AND RETZEL, E. F., 2003, MtDB: a database for personalized data mining of the model legume Medicago truncatula transcriptome. *Nucleic Acids Research,* **31**:196–201.

LEHMANN, T., HAWLEY, W. A. AND COLLINS, F. H., 1996, An evaluation of evolutionary constraints on microsatellite loci using null alleles. *Genetics*, **144:**1155–1163.

LEWIN, B., 2003, *Genes VIII*. Oxford University Press, New York.

LIU, B. AND WENDEL, J. F., 2000, Retrotransposon activation followed by rapid repression in introgressed rice plants. *Genome,* **43**:874-880.

MA, Z. Q., RODER, M. AND SORRELLS, M. E., 1996, Frequency and sequence characterstics of di-, tri- and tetranucleotide microsatellites in wheat. *Genome,* **39**:123-130.

MAY, B. P. AND MARTIENSSEN, R. A., 2003, Transposon mutagenesis in the study of plant development. *Critical Reviews in Plant Sciences,* **22**:1–35.

MESSIER, W., LI, S. H., STEWART, C. B., 1996, The birth of microsatellites. *Nature*, **381**:483.

METZGAR, D., BYTOF, J., WILLS, C., 2000, Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Research,* **10**:72-80.

MOHAMMADI, S. A. AND PRASANNA, B. M., 2003, Analysis of genetic diversity in crop plants – salient statistical tools and considerations. *Crop Science*, **43:**1235–1248.

MORGANTE, M. AND OLIVERI, A. M., 1993, PCR-amplified microsatellites as markers in plant genetics. *The Plant journal*, **3**:175-182.

MORGANTE, M., HANAFEY, M. AND POWELL, W., 2002, Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nature Genetics,* **30**:194-200.

MOXON, E. R. AND WILLS, C., 1999, DNA microsatellites: agents of evolution? *Scientific American,* **280**:94-99.

MULLET, J. E., KLEIN, R. R. AND KLEIN, P. E., 2002, *Sorghum bicolor* - an important species for comparative grass genomics and a source of beneficial genes for agriculture. *Current Opinion in Plant Biology,* **5**:2 118-121.

NERI, C., ALBANESE, V., LEBRE, A. S., HOLBERT, S., SAADA, C., BOUGUELERET, L., MEIER-EWERT, S., LE-GALL, I., MILLASSEAU, P., BUI, H., GIUDICELLI, C., MASSART, C., GUILLOU, S., GERVY, P., POULLIER, E., RIGAULT, P., WEISSENBACH, J., LENNON, G., CHUMAKOV, I., DAUSSET, J., LEHRACH, H., COHEN, D. AND CANN, H. M., 1996, Survey of CAG/CTG repeats in human cDNAs representing new genes: candidates for inherited neurological disorders. *Human Molecular Genetics,* **5**:1001-1009.

NICOT, N., CHIQUET, V., GANDON, B., AMILHAT, L., LEGEAI, F., LEROY, P., BERNARD, M. AND SOURDILLE, P., 2004, Study of simple sequence repeat (SSR) markers from wheat expressed sequence tags (ESTs). *Theoretical and Applied Genetics,* **109**:800–805.

O'BRIEN, S. J., 1991, Molecular genome mapping lessons and prospects. *Current Opinion in Genetic Development,* **1**:105–111.

PANAUD, O., CHEN, X. AND MCCOUCH, S. R., 1995, Frequency of microsatellite sequences in rice (*Oryza sativa* L.). *Genome*, **38**:1170-1176.

PAPI, M., SABATINI, S., BOUCHEZ, D., CAMILLERI, C., COSTANTINO, P., AND VITTORIOSO, P., 2000, Identification and disruption of an Arabidopsis zinc finger gene controlling seed germination. *Genes and Development*, **14**:28–33.

POWELL, W., MACHRAY, G. C. AND PROVAN, J., 1996, Polymorphism revealed by simple sequence repeats. *Trends in Plant Science,* **1**:215–222.

PRATT, L. H., LIANG, C., SHAH, M., SUN, F., WANG, H., ST. REID, P., GINGLE, A. R., PATERSON, A. H., WING, R., DEAN, R., KLEIN, R., NGUYEN, H. T., MA, H. M., ZHAO, X., MORISHIGE, D. T., MULLET, J. E. AND CORDONNIER-PRATT, M. M., 2005, Sorghum expressed sequence tags identify signature genes for drought, pathogenesis, and skotomorphogenesis from a milestone set of 16,801 unique transcripts. *Plant Physiology*, **139:** 869-884.

PRIMMER, C. R. AND ELLEGREN, H., 1998, Patterns of molecular evolution in avian microsatellites. *Molecular Biology and Evolution*, **15**:997-1008.

PUJANA, M. A., GRATACOS, M., CORRAL, J., BANCHS, I., SANCHEZ, A., GENIS, D., CERVERA, C., VOLPINI, V. AND ESTIVILL, X., 1997, Polymorphisms at 13 expressed human sequences containing CAG/CTG repeats and analysis in autosomal dominant cerebellar ataxia (ADCA) patients. *Human Genetics*, 101:18-21.

PULST, S. M., NECHIPORUK, A., NECHIPORUK, T., GISPERT, S. CHEN, X. N., LOPES-CENDES, I., PEARLMAN, S., STARKMAN, S., DIAZ, O. G., LUNKES, A., DEJONG, P., ROULEAU, G. A., AURBURGER, G., KORENBERG, J. R., FIGUEROA, C. AND SAHBA, S., 1996, Moderate expansion of a normally biallelic trinucleotide repeat in *Spinocerebellar ataxia* type 2. *Nature Genetics,* **13**:269–276.

QUACKENBUSH, J., CHO, J., LEE, D., LIANG, F., HOLT, I., KARAMYCHEVA, S., PARVIZI, B., PERTEA, G., SULTANA, R. AND WHITE, J., 2001, The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research,* **29**:159–164.

QUINT, M., MIHALJEVIC, R., DUSSLE, C., XU, M., MELCHINGER, A. AND LÜBBERSTEDT, T., 2002, Development of RGA–CAPS markers and genetic mapping of candidate genes for sugarcane mosaic virus resistance in maize. *Theoretical and Applied Genetics,* **105:**355–363.

RODER, M. S., KORZUN, V., GILL, B. S. AND GANAL, M. W., 1998, The physical mapping of microsatellite markers in wheat. *Genome*, **41**:278-283.

RONNING, C. M., STEGALKINA, S. S., ASCENZI, R. A., BOUGRI, O., HART, A. L., TERESA, R. UTTERBACH, VANAKEN, S. E., RIEDMULLER, S. B., WHITE, J. A., CHO, J., PERTEA, G. M., LEE, Y., KARAMYCHEVA, S., SULTANA, R., TSAI, J., QUACKENBUSH, J., GRIFFITHS, H. M., RESTREPO, S., SMART, C. D., FRY, W. E., RUTGER, VAN DER HOEVEN, TANKSLEY, S., ZHANG, P., JIN, H., YAMAMOTO, M. L., BAKER, B. J. AND BUELL, C. R., 2003, Comparative analyses of potato expressed sequence tag libraries. *Plant Physiology*, 131:419–429.

RUDD, S., 2003, Expressed sequence tags: alternative or complement to whole genome sequences? *Trends in Plant Science,* **8:**321-329.

RUDD, S., MEWES, H. W. AND MAYER, K. F. X., 2003, Sputnik: a database platform for comparative plant genomics. *Nucleic Acids Research*, **31:**128–132.

RUNGIS, D., BÉRUBÉ, Y., ZHANG, J., RALPH, S., RITLAND, C. E., ELLIS, B. E., DOUGLAS, C., BOHLMANN, J. AND RITLAND, K., 2004, Robust simple sequence repeat markers for spruce (*Picea* spp.) from expressed sequence tags. *Theoretical and Applied Genetics,* **109**:1283–1293.

RUSSELL, J., BOOTH, A., FULLER, J., HARROWER, B., HEDLEY, P., MACHRAY, G. AND POWELL, W., 2004, A comparison of sequence-based polymorphism and haplotype content in transcribed and anonymous regions of the barley. *Genome,* **47**:389–398.

SAHA, M. C., MIAN, R. M. A., EUJAYL, I., ZWONITZER, J. C., WANG, L. J. AND MAY, G. D., 2004, Tall fescue EST-SSR markers with transferability across several grass species. *Theoretical and Applied Genetics,* **109**:783–791.

SANPEI, K., TAKANO, H., IGARASHI, S., SATO, T., OYAKE, M., SASAKI, H., WAKISAKA, A., TASHIRO, T., ISHIDA, Y., IKEUCHI, T., KOIDE, R., SAITO, M., SATO, A., TANAKA, T., HANYU, S., TAKIYAMA, Y., NISHIZAWA, M., SHIMIZU, N., NOMURA, Y., SAGAWA, N., IWABUCHI, K., EGUCHI, T., TANAKA, H., TAKANASHI, H. AND TSUJI, S., 1996, Identification of the spinocerebellar ataxia type 2 gene using a direct identification of repeat expansion and cloning technique. *Nature Genetics,* **14**:277-284.

SASAKI, T., BILLET, E., PETRONIS, A., YING, D., PARSONS, T., MACCI ARDI, F. M., MELTZER, H. Y., LIEBERMAN, J., JOFFE, R. T., ROSS, C. A., MCINNIS, M. G.,

LI, S. H. AND KENNEDY, J. L., 1996, Psychosis and genes with trinucleotide repeat polymorphism. *Human Genetics*, **97**:244–246.

SCHULMAN, A. H., GUPTA, P. K. AND VARSHNEY, R. K., 2004, Organization of retrotransposons and microsatellites in cereal genomes. In *Cereal Genomics* eds. Gupta, P. K. and Varshney, R. K., 2004, Kluwer Academic Publishers, pp. 83–118.

SCOTT, K. D., EGGLER, P., SEATON, G., ROSSETTO, M., ABLETT, E. M., LEE, L. S. AND HENRY, R. J., 2000, Analysis of SSRs derived from grape ESTs. *Theoretical and Applied Genetics,* **100:**723–726.

SEKINO, M., HAMAGUCHI, M., ARANISHI, F., AND OKOSHI, K., 2003, Development of novel microsatellite DNA markers from the Pacific oyster *Crassostrea gigas*. *Marine Biotechnology,* **5**:227–233.

SCHULER, G. D., BOGUSKI, M. S., STEWART, E. A., STEIN, L. D., GYAPAY, G., RICE, K., WHITE, R. E., RODRIGUEZ-TOMÉ, P., AGGARWAL, A., BAJOREK, E., BENTOLILA, S., BIRREN, B. B., BUTLER, A., CASTLE, A. B., CHIANNILKULCHAI, N., CHU, A., CLEE, C., COWLES, S., R. DAY, P. J., DIBLING, T., DROUOT, N., DUNHAM, I., DUPRAT, S., EAST, C., EDWARDS, C., FAN, J. B., FANG, N., FIZAMES, C., GARRETT, C., GREEN, L., HADLEY, D., HARRIS, M., HARRISON, P., BRADY, S., HICKS, A., HOLLOWAY, E., HUI, L., HUSSAIN, S., LOUIS-DIT-SULLY, C., MACGILVERY, A M. A. J., MADER, C., MARATUKULAM, A., MATISE, T. C., MCKUSICK, K. B., MORISSETTE, J., MUNGALL, A., MUSELET, D., NUSBAUM, H. C., PAGE, D. C., PECK, A., PERKINS, S., PIERCY, M., QIN, F., QUACKENBUSH, J., RANBY, S., REIF, T., ROZEN, S., SANDERS, C., SHE, X., SILVA, J., SLONIM, D. K., SODERLUND, C., SUN, P W. L., TABAR, T., THANGARAJAH, VEGA-CZARNY, N., VOLLRATH, D., VOYTICKY, S., WILMER, T., WU X., ADAMS, M. D., AUFFRAY, C. ,. WALTER, N. A. R., BRANDON, R., DEHEJIA, A., GOODFELLOW, P. N., HOULGATTE, R., HUDSON, J. R. jr., IDE, S. E., IORIO, K. R., LEE, W. Y., SEKI, N., NAGASE, T., ISHIKAWA, K., NOMURA, N., PHILLIPS, C., POLYMEROPOULOS, M. H., SANDUSKY, M., SCHMITT, K., BERRY, R., SWANSON, K., TORRES, R., VENTER, J. C., SIKELA, J. M., BECKMANN, J. S., WEISSENBACH, J., MYERS, R. M., COX, D. R., JAMES, M. R., BENTLEY, D., DELOUKAS, P., LANDER, E. S. AND HUDSON, T. J. ,1996, A gene map of the human genome. *Science,* **274**:540–546.

SHRAGER, J., HAUSER, C., CHANG, C. W., HARRIS, E. H., DAVIES, J., MCDERMOTT, J., TAMSE, R., ZHANG, Z. AND GROSSMAN, A. R., 2003, *Chlamydomonas reinhardtii* genome project. A guide to the generation and use of the cDNA information. *Plant Physiology,* **131**:401–408.

SORRELLS, M. E. AND WILSON, W. A., 1997, Direct classification and selection of superior alleles for crop improvement. *Crop Science,* **37**:691–697.

SREENIVASULU, N., KAVI KISHOR, P. B., VARSHNEY, R. K. AND ALTSCHMIED, L., 2002, Mining functional information from cereal genomes – the utility of expressed sequence tags. *Current Science*, **83**:965–973.

STOESSER, G., BAKER, W., VAN DEN BROEK, A., GARCIA-PASTOR, M., KANZ, C., KULIKOVA, T., LEINONEN, R., LIN, Q., LOMBARD, V., LOPEZ, R., MANCUSO, R., NARDONE, F., STOEHR, P., TULI, M. A., TZOUVARA, K. AND VAUGHAN, R., 2003, The EMBL Nucleotide Sequence Database: major new developments. *Nucleic Acids Research*, **31**:17-22.

SUTTON, G., O. WHITE, D. ADAMS, AND A. KERLAVAGE, 1995, TIGR Assembler: A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology,* **1**: 9–18.

TARAMINO, G., TARCHINI, R., FERRARIO, S., LEE, M. AND PE, M. E., 1997, Characterization and mapping of simple sequence repeats (SSRs) in *Sorghum bicolor*. *Theoretical and Applied Genetics*, **95**: 66-72.

TAUTZ, D. AND RENZ, M., 1994, Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Research*, **12**:4127-4138.

TEMNYKH, S., DECLERCK, G., LUKASHOVA, A., LIPOVICH, L., CARTINHOUR, S. AND MCCOUCH, S., 2001, Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Research*, **11**:1441-1452.

TEMNYKH, S., PARK, W. D., AYRES, N., CARTINHOUR, S., HAUCK, N., LIPOVICH, L., CHO, Y. G., ISHII, T. AND MCCOUCH, S. R., 2000, Mapping and genome organization of microsatellite sequences in rice (*Oryza sativa* L.). *Theoretical and Applied Genetics,* **100**:697-712.

THE ARABIDOPSIS GENOME INITIATIVE, 2000, Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature,* **408**:796–815.

THIEL, T., MICHALEK, W., VARSHNEY, R. K. AND GRANER, A., 2003, Exploiting EST databases for the development of cDNA derived microsatellite markers in barley (*Hordeum vulgare* L.). *Theoretical and Applied Genetics,* **106**:411-422.

TOTHET, G., GASPARI, Z. AND JURKA, J., 2000, Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Research,* **10**:967-981.

TROJANOWSKA, M. R. AND BOLIBOK, H., 2004, Characteristics and a comparison of three classes of Microsatellite-based markers and their application in plants. *Cellular and Molecular Biology,* **9**:221-238.

UDALL, J. A., SWANSON, J. M., HALLER, K., RAPP, R. A., SPARKS, M. E., HATFIELD, J., YEISOO, YU, WU, Y., DOWD, C., ARPAT, A. B., SICKLER, B. A., WILKINS, T. A., GUO, J. Y., YA, X., CHEN, SCHEFFLER, J., TALIERCIO, E., TURLEY, R., MCFADDEN, H., PAYTON, P., KLUEVA, N., ALLEN, R., ZHANG, D., HAIGLER, C., WILKERSON, C., SUO, J., SCHULZE, S. R., PIERCE, M. L., ESSENBERG, M., KIM, H., LLEWELLYN, D. J., DENNIS, E. S., KUDRNA, D., WING, R. AND ANDREW, H., 2006, A global assembly of cotton ESTs. *Genome Research,* **16**:441–450.

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., THIEL, T., STEIN, N., LANGRIDGE, P. AND GRANER, A., 2002, *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cell Molecular Biology,* **7**:537-546.

WALDEN, R., 2002, T-DNA tagging in a genomics era. *Critical Reviews in Plant Sciences,* **21:**143–165.

WARE, D. H., JAISWAL, P., NI, J., YAP, I. V., PAN, X., CLARK, K. Y., TEYTELMAN, L., SCHMIDT, S. C., ZHAO, W., CHANG, K., CARTINHOUR, S., STEIN, L. D. AND MCCOUCH, S. R., 2002, Gramene, a tool for grass genomics. *Plant Physiology,* **130**:1606–1613.

WARNKE, S. E., BARKER, R. E., JUNG, G., SUNG-CHUR, S., ROUF-MIAN, M. A., SAHA, M. C., BRILMAN, L. A., DUPAL, M. P. AND FORSTER, J. W., 2004, Genetic linkage mapping of an annual x perennial ryegrass population. *Theoretical and Applied Genetics,* **109:**294–304.

WHEELER, D. L., BARRETT, T., BENSON, D. A., BRYANT, S. H., CANESE, K., CHETVERNIN, V., CHURCH, D. M., DICUCCIO, M., EDGAR, R., FEDERHEN, S., GEER, L. Y., HELMBERG, W., KAPUSTIN, Y., KENTON, D. L., KHOVAYKO, O., LIPMAN, D. J., MADDEN, T. L., MAGLOTT, D. R., OSTELL, J., PRUITT, K. D., SCHULER, G. D., SCHRIML, L. M., SEQUEIRA, E., SHERRY, S. T., SIROTKIN, K., SOUVOROV, A., STARCHENKO, G., SUZEK, T. O., TATUSOV, R. N, TATUSOVA, T. A., WAGNER, L. AND YASCHENKO, E., 2003, Database resources of the National Center for Biotechnology. *Nucleic Acids Research,* **31**:28–33.

WILLIAMSON, ELLISTON, A., K. AND STURCHIO, J., 1995, The Merck Gene Index, a public resource for genomics research. *Journal of National Institute of Health Research*, **7**:61–63.

WOODHEAD, S. AND G. COOPER-DRIVER, 1979, Phenolic acids and resistance it insect attack in *Sorghum bicolor*. *Biochemical Systematics and Ecology,* **7**:309-310.

WU, K. S. AND TANKSLEY, S. D., 1993, Abundance, polymorphism and genetic mapping of microsatellites in rice. *Molecular and General Genetics,* **241**:225-235.

YU, JU-KYUNG, SINGH, S., DAKE, T. M., BENSCHER, D., GILL, B. S., AND SORRELLS, M. E., 2004, Development and mapping of EST-derived simple sequence repeat (SSR) markers for hexaploid wheat. *Genome,* **47:**805–818.

ZHANG, W. K., WANG, Y. J., LUO, G. A., ZHANG, J. S, HE, C. Y., WU, X. L., GAI, J. Y., AND CHEN, S. Y., 2004, QTL mapping of ten agronomic traits on the soybean (*Glycine max* L. Merr.) genetic map and their association with EST markers. *Theoretical and Applied Genetics,* **108**:1131-1139.

ZHIGUO, H., WANG, C., SONG, X., GUO, W., GOU, J., LI, C., CHEN, X. AND TIANZHEN, 2006, Characteristics, development and mapping of *Gossypium hirsutum* derived EST-SSRs in allotetraploid cotton. *Theoretical and Applied Genetics,* **112**:430-439.

APPENDIX I

LIST OF ABBREVATIONS USED

mt	: million tons
mha	: million hectares
min	: minute
sec	: second
℃	: degree centigrade
L	: litre
ml	: millilitre
µl	: microlitre
g	: gram
mg	: milligram
µg	: microgram
ηg	: nanogram
mM	: millimolar
ρM	: picomolar
bp	: basepairs
Mbp	: megabasepairs
Cbp	: complementry basepairs
cDNA	: complementry DNA
EST	: Expressed Sequence Tags
in-del	: Insertions and Deletions
SNP	: Single Nucleotide Polymorphism
SSR	: Simple Sequence Repeats
gSSR	: Genic SSRs
STR	: Simple Tandem Repeats
STS	: Sequence Tagged Sites
DNR	: dineucleotide repeats
TNR	: trineucleotide repeats
TTNR	: tetraneucleotide repeats
UTR	: Untranslated regions
RIL	: Recombinant Inbred Lines
EMBL	: European Molecular Biology Laboratory
DDBJ	: DNA Database of Japan
PDB	: Protein Data Bank
NCBI	: National Centere for Biotechnology Information
GCG	: Genetics Computer Group
MISA	: Microsatellite identification tool
Pfam	: Protein family database
CDS	: Coding Sequence
nr	: Non redundant
dbEST	: EST database
BLAST	: Basic Local Alignment Search Tool
FASTA	: FAST-All (comparision tool)
Mb	: Mega Bytes
kb	: Kilo Bytes
RAM	: Random Access Memory
MS-DOS	: Microsoft Disk Operating System

APPENDIX II

Proportion of different repeat motifs across different transcriptomes

Motif	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*
ac/ca/gt/tg	20.50	37.40	31.50	13.00	23.50	10.90	10.80	20.80	25.80	30.60	25.80	-	24.50	30.00
ag/ga/tc/ct	55.90	48.20	44.80	63.70	46.90	73.00	60.80	56.60	53.40	45.80	51.60	-	52.70	43.30
at/ta	17.70	8.90	18.40	16.90	18.80	14.90	28.20	15.00	15.50	17.00	21.20	-	9.10	20.00
gc/cg	5.90	5.40	5.30	6.50	10.90	1.40	-	7.50	5.20	6.80	1.50	-	13.60	6.70
aac/aca/caa/ttg/tgt/gtt	-	2.60	2.60	0.80	1.00	5.00	-	-	3.00	1.50	1.40	-	3.80	-
aag/aga/gaa/ctt/ttc/tct	10.20	5.20	11.70	3.20	2.90	10.00	-	2.80	4.50	4.50	8.40	-	3.70	6.00
aat/ata/taa/tta/tat/att	-	-	1.30	0.80	1.00	10.00	25.00	2.80	-	1.50	1.40	-	0.60	4.00
acc/cac/cca/ggt/gtg/tgg	5.20	3.90	2.60	4.80	9.60	10.00	-	4.30	7.70	1.50	5.60	-	5.70	4.00
acg/cgc/ctg/gac/gct/tgc	7.70	7.70	9.10	7.20	11.40	15.00	25.00	7.20	6.10	7.50	5.60	100.00	8.30	6.00
act/cta/tca/tga/gat/agt	-	5.20	-	3.20	1.00	5.00	-	1.40	-	1.50	4.20	-	4.50	2.00
agc/gca/cag/tcg/cgt/gtc	12.80	16.70	19.70	13.60	11.50	-	25.00	26.00	20.00	21.00	15.50	-	13.90	16.00
agg/cct/ctc/gag/gga/tcc	10.40	12.90	7.80	24.00	12.50	5.00	-	11.50	18.40	10.50	22.40	-	20.40	18.00
atc/tca/cat/tag/agt/gta	-	3.90	-	0.80	2.00	10.00	-	-	-	4.50	1.40	-	1.20	-
atg/tga/gat/tac/act/cta	-	-	-	-	-	-	-	-	-	-	1.40	-	-	2.00
ccg/cgc/gcc/ggc/gcg/cgg	53.80	42.40	44.70	42.00	47.70	30.00	25.00	43.40	40.00	46.50	32.30	-	37.30	42.00
aaat/aata/ataa/taaa/ttta/ttat/tatt/attt	-	-	-	-	-	-	-	-	-	6.70	-	-	-	-
aacc/acca/ccaa/caac/ttgg/tggt/ggtt/gttg	-	-	-	-	-	-	-	-	-	-	-	-	-	20.00
aatc/atca/tcaa/caat/ttag/tagt/agtt/gtta	-	-	-	-	-	-	-	-	-	-	-	-	6.70	-
acag/caga/agac/gaca/tgtc/gtct/tctg/ctgt	-	-	-	-	-	-	-	-	20.00	-	-	-	-	-
acat/cata/atac/taca/tgta/gtat/tatg/atgt	-	9.10	-	-	-	50.10	-	8.30	-	6.70	-	-	6.70	-
accc/ccca/ccac/cacc/tggg/gggt/ggtg/gtgg	-	-	-	-	-	-	-	-	-	6.70	-	-	-	-
acga/cgaa/gaac/aacg/tgct/gctt/cttg/ttgc	-	-	-	-	-	-	-	-	-	-	-	-	-	-
acgt/cgta/gtac/tacg/tgca/gcat/catg/atgc	-	-	-	7.70	-	16.70	50.00	8.30	-	6.70	-	-	13.40	20.00
agaa/gaaa/aaag/aaga/tctt/cttt/tttc/ttct	-	9.10	-	-	-	-	-	24.90	-	6.70	14.30	-	-	20.00
agat/gata/atag/taga/tcta/ctat/tatc/atct	-	18.20	-	7.70	-	16.70	-	-	20.00	-	-	-	6.70	-
agcc/gcca/ccag/cagc/tcgg/cggt/ggtc/gtcg	-	-	-	7.70	10.00	-	-	-	20.00	-	-	-	6.70	-
agcg/gcga/cgag/gagc/tcgc/cgct/gctc/ctcg	-	18.20	-	-	-	-	-	-	-	6.70	-	-	-	-
agct/gcta/ctag/tagc/tcga/cgat/gatc/atcg	16.70	9.10	-	7.70	-	-	50.00	8.30	-	6.70	-	-	-	-
agga/ggaa/gaag/aagg/tcct/cctt/cttc/ttcc	-	9.10	-	-	10.00	-	-	-	-	-	-	-	6.70	-
aggc/ggca/gcag/cagg/tccg/ccgt/cgtc/gtcc	16.70	-	-	15.40	-	-	-	8.30	-	-	-	-	-	-
aggg/ggga/ggag/gagg/tccc/ccct/cctc/ctcc	16.70	-	33.30	7.70	-	-	-	-	20.00	6.70	14.30	-	13.30	-
aggt/ggta/gtag/tagg/tcca/ccat/catc/atcc	16.70	9.10	-	15.40	-	-	-	-	-	-	14.30	-	6.70	-
agtg/gtga/tgag/gagt/tcac/cact/actc/ctca	-	-	-	7.70	-	-	-	-	-	6.70	-	-	-	-
atga/tgaa/gaat/aatg/tact/actt/ctta/ttac	-	-	-	-	10.00	-	-	-	-	6.70	-	-	-	-
atta/ttaa/taat/aatt	-	-	-	-	-	-	-	-	-	-	-	-	-	-
attc/ttca/tcat/catt/taag/aagt/agta/gtaa	-	9.10	-	-	-	-	-	-	-	-	-	-	-	-
attg/ttga/tgat/gatt/taac/aact/acta/ctaa	-	-	-	-	10.00	-	-	-	-	6.70	-	-	-	20.00
cccg/ccgc/cgcc/gccc/ggcg/gggc/gcgg/cggg	-	-	-	15.40	10.00	-	-	8.30	-	13.40	-	-	6.70	20.00
ccgg/ggcc/gccg/cggc	-	-	-	-	20.00	-	-	8.30	-	6.70	14.30	-	6.70	-
cctg/ctgc/tgcc/gcct/ggac/gacg/acgg/cgga	33.40	-	-	7.70	10.00	-	-	16.60	20.00	-	14.30	-	13.40	-
cgtg/gtgc/tgcg/gcgt/gcac/cacg/acgc/cgca	-	-	33.30	-	-	-	-	-	-	6.70	-	-	-	-
cgtt/gttc/ttcg/tcgt/gcaa/caag/aagc/agca	-	-	-	-	10.00	-	-	-	-	-	-	-	6.70	-
ggat/gatg/atgg/tgga/ccta/ctac/tacc/acct	-	-	-	-	-	-	-	-	-	-	14.30	-	-	-
gtca/tcag/cagt/agtc/cagt/agtc/gtca/tcag	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*
tgac/gact/actg/ctga/actg/ctga/tgac/gact	20.50	37.40	31.50	13.00	23.50	10.90	10.80	20.80	25.80	30.60	25.80	-	24.50	30.00
ttgt/tgtt/gttt/tttg/aaca/acaa/caaa/aaac	55.90	48.20	44.80	63.70	46.90	73.00	60.80	56.60	53.40	45.80	51.60	-	52.70	43.30

Proportion of different repeat motifs across different transcriptomes (contd.)

Motif	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*25*	*26*	*27*	*28*	*29*	*30*
ac/ca/gt/tg	23.40	13.60	22.30	22.30	22.40	19.50	18.20	21.70	21.60	-	14.50	17.70	18.40	34.70	29.30	31.60
ag/ga/tc/ct	68.70	57.50	55.50	46.30	63.20	54.30	54.50	69.50	53.80	66.60	55.30	60.80	51.00	43.40	58.80	45.50
at/ta	5.90	25.80	16.70	24.10	6.10	21.70	27.30	8.70	15.40	-	18.40	13.70	24.40	17.40	7.80	14.00
gc/cg	2.00	3.00	5.60	7.40	8.20	4.30	-	-	9.20	33.30	11.80	7.80	6.10	4.30	3.90	8.80
aac/aca/caa/ttg/tgt/gtt	2.60	4.60	2.40	1.60	1.00	-	10.00	-	2.00	-	2.60	-	-	-	-	1.60
aag/aga/gaa/ctt/ttc/tct	10.40	-	4.80	8.00	2.00	7.80	-	3.40	3.00	33.30	5.20	2.10	5.10	9.10	9.60	3.20
aat/ata/taa/tta/tat/att	-	2.30	4.80	-	-	-	5.00	3.40	1.00	33.30	6.50	-	1.70	9.10	-	-
acc/cac/cca/ggt/gtg/tgg	7.80	2.30	14.60	16.10	4.00	9.10	10.00	10.20	9.00	-	3.90	10.60	8.50	-	1.90	11.00
acg/cgc/ctg/gac/gct/tgc	2.60	15.90	4.90	1.60	11.00	9.10	20.00	10.20	10.00	-	3.90	10.60	1.70	18.20	15.30	12.60
act/cta/tca/tga/gat/agt	-	-	4.80	-	3.00	2.60	5.00	-	1.00	-	2.60	4.30	-	-	5.70	3.10
agc/gca/cag/tcg/cgt/gtc	10.30	15.90	12.10	16.00	16.00	16.90	30.00	13.70	15.00	33.30	24.70	6.30	27.20	27.30	28.90	17.30
agg/cct/ctc/gag/gga/tcc	33.50	20.40	12.20	14.40	17.00	19.50	5.00	27.40	17.00	-	10.40	25.60	17.00	9.10	13.30	6.30
atc/tca/cat/tag/agt/gta	5.20	2.30	-	1.60	1.00	-	-	-	-	-	-	4.30	-	-	-	-
atg/tga/gat/tac/act/cta	-	2.30	-	-	-	1.30	-	-	-	-	2.60	4.20	-	-	-	1.60
ccg/cgc/gcc/ggc/gcg/cgg	28.20	34.00	38.90	40.50	45.00	33.80	15.00	30.80	42.00	-	37.70	31.90	39.10	27.30	24.90	43.80
aaat/aata/ataa/taaa/ttta/ttat/tatt/attt	14.30	-	-	28.60	33.40	-	-	-	9.10	-	-	-	9.10	-	-	-
aacc/acca/ccaa/caac/ttgg/tggt/ggtt/gttg	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
aatc/atca/tcaa/caat/ttag/tagt/agtt/gtta	-	12.50	-	-	-	-	-	-	9.10	-	-	-	-	-	6.70	-
acag/caga/agac/gaca/tgtc/gtct/tctg/ctgt	-	12.50	-	-	-	-	-	-	-	-	100.00	-	-	-	-	-
acat/cata/atac/taca/tgta/gtat/tatg/atgt	-	-	-	14.30	-	12.50	100.00	-	-	-	-	-	-	-	20.00	10.00
accc/ccca/ccac/cacc/tggg/gggt/ggtg/gtgg	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
acga/cgaa/gaac/aacg/tgct/gctt/cttg/ttgc	-	-	-	14.30	-	-	-	-	-	-	-	-	-	-	-	-
acgt/cgta/gtac/tacg/tgca/gcat/catg/atgc	14.30	-	-	-	-	12.50	-	-	-	-	-	14.30	9.10	66.60	6.70	-
agaa/gaaa/aaag/aaga/tctt/cttt/tttc/ttct	-	-	-	-	16.70	12.50	-	-	-	-	-	-	-	-	-	10.00
agat/gata/atag/taga/tcta/ctat/tatc/atct	-	37.50	-	-	-	12.50	-	-	-	-	-	-	9.10	33.30	-	-
agcc/gcca/ccag/cagc/tcgg/cggt/ggtc/gtcg	-	-	-	-	33.40	-	-	-	-	-	-	-	-	-	-	-
agcg/gcga/cgag/gagc/tcgc/cgct/gctc/ctcg	-	-	-	-	-	12.50	-	20.00	9.10	-	-	-	-	-	-	10.00
agct/gcta/ctag/tagc/tcga/cgat/gatc/atcg	-	25.00	-	-	-	-	-	20.00	-	-	-	-	18.20	-	-	10.00
agga/ggaa/gaag/aagg/tcct/cctt/cttc/ttcc	-	-	-	-	-	-	-	20.00	9.10	-	-	-	9.10	-	6.70	-
aggc/ggca/gcag/cagg/tccg/ccgt/cgtc/gtcc	14.30	-	-	-	-	12.50	-	-	9.10	-	-	-	-	-	6.70	-
aggg/ggga/ggag/gagg/tccc/ccct/cctc/ctcc	-	12.50	-	28.60	-	-	-	20.00	-	-	-	28.60	-	-	13.30	-
aggt/ggta/gtag/tagg/tcca/ccat/catc/atcc	-	-	-	-	-	-	-	-	-	-	-	14.30	9.10	-	26.70	20.00
agtg/gtga/tgag/gagt/tcac/cact/actc/ctca	-	-	-	-	-	-	-	-	-	50.00	-	-	9.10	-	-	10.00
atga/tgaa/gaat/aatg/tact/actt/ctta/ttac	14.30	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
atta/ttaa/taat/aatt	14.30	-	-	-	-	-	-	-	18.20	-	-	-	-	-	-	-
attc/ttca/tcat/catt/taag/aagt/agta/gtaa	14.30	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
attg/ttga/tgat/gatt/taac/aact/acta/ctaa	-	-	100.00	-	16.70	-	-	-	-	50.00	-	-	-	-	6.70	10.00
cccg/ccgc/cgcc/gccc/ggcg/gggc/gcgg/cggg	-	-	-	-	-	12.50	-	-	-	-	-	14.30	-	-	-	10.00
ccgg/ggcc/gccg/cggc	-	-	-	-	-	-	-	-	9.10	-	-	14.30	-	-	-	-
cctg/ctgc/tgcc/gcct/ggac/gacg/acgg/cgga	-	-	-	-	-	12.50	-	20.00	-	-	-	14.30	9.10	-	6.70	10.00
cgtg/gtgc/tgcg/gcgt/gcac/cacg/acgc/cgca	14.30	-	-	-	-	-	-	-	9.10	-	-	-	-	-	-	-
cgtt/gttc/ttcg/tcgt/gcaa/caag/aagc/agca	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ggat/gatg/atgg/tgga/ccta/ctac/tacc/acct	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
gtca/tcag/cagt/agtc/cagt/agtc/gtca/tcag	-	-	-	-	-	-	-	-	-	-	-	-	9.10	-	-	-
tgac/gact/actg/ctga/actg/ctga/tgac/gact	-	-	-	-	-	-	-	-	-	-	-	-	9.10	-	-	-
ttgt/tgtt/gttt/tttg/aaca/acaa/caaa/aaac	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Proportion of different repeat motifs across different transcriptomes.

Motif	*Total*	*Per cent*
ac/ca/gt/tg	636.30	21.94819
ag/ga/tc/ct	1601.40	55.23783
at/ta	469.00	16.17743
gc/cg	192.40	6.636542
aac/aca/caa/ttg/tgt/gtt	50.10	1.670445
aag/aga/gaa/ctt/ttc/tct	180.10	6.004935
aat/ata/taa/tta/tat/att	115.50	3.851027
acc/cac/cca/ggt/gtg/tgg	183.90	6.131635
acg/cgc/ctg/gac/gct/tgc	371.40	12.3833
act/cta/tca/tga/gat/agt	60.10	2.003868
agc/gca/cag/tcg/cgt/gtc	522.60	17.42465
agg/cct/ctc/gag/gga/tcc	421.90	14.06708
atc/tca/cat/tag/agt/gta	38.20	1.273673
atg/tga/gat/tac/act/cta	15.40	0.51347
ccg/cgc/gcc/ggc/gcg/cgg	1040.00	34.67591
aaat/aata/ataa/taaa/ttta/ttat/tatt/attt	101.20	3.611706
aacc/acca/ccaa/caac/ttgg/tggt/ggtt/gttg	20.00	0.713776
aatc/atca/tcaa/caat/ttag/tagt/agtt/gtta	35.00	1.249108
acag/caga/agac/gaca/tgtc/gtct/tctg/ctgt	132.50	4.728765
acat/cata/atac/taca/tgta/gtat/tatg/atgt	237.70	8.483226
accc/ccca/ccac/cacc/tggg/gggt/ggtg/gtgg	6.70	0.239115
acga/cgaa/gaac/aacg/tgct/gctt/cttg/ttgc	14.30	0.51035
acgt/cgta/gtac/tacg/tgca/gcat/catg/atgc	246.30	8.79015
agaa/gaaa/aaag/aaga/tctt/cttt/tttc/ttct	114.20	4.07566
agat/gata/atag/taga/tcta/ctat/tatc/atct	161.70	5.770878
agcc/gcca/ccag/cagc/tcgg/cggt/ggtc/gtcg	77.80	2.776588
agcg/gcga/cgag/gagc/tcgc/cgct/gctc/ctcg	76.50	2.730193
agct/gcta/ctag/tagc/tcga/cgat/gatc/atcg	171.70	6.127766
agga/ggaa/gaag/aagg/tcct/cctt/cttc/ttcc	70.70	2.523198
aggc/ggca/gcag/cagg/tccg/ccgt/cgtc/gtcc	83.00	2.96217
aggg/ggga/ggag/gagg/tccc/ccct/cctc/ctcc	215.00	7.673091
aggt/ggta/gtag/tagg/tcca/ccat/catc/atcc	132.30	4.721627
agtg/gtga/tgag/gagt/tcac/cact/actc/ctca	83.50	2.980014
atga/tgaa/gaat/aatg/tact/actt/ctta/ttac	31.00	1.106353
atta/ttaa/taat/aatt	32.50	1.159886
attc/ttca/tcat/catt/taag/aagt/agta/gtaa	23.40	0.835118
attg/ttga/tgat/gatt/taac/aact/acta/ctaa	220.10	7.855103
cccg/ccgc/cgcc/gccc/ggcg/gggc/gcgg/cggg	110.60	3.947181
ccgg/ggcc/gccg/cggc	79.40	2.83369
cctg/ctgc/tgcc/gcct/ggac/gacg/acgg/cgga	188.00	6.709493
cgtg/gtgc/tgcg/gcgt/gcac/cacg/acgc/cgca	63.40	2.26267
cgtt/gttc/ttcg/tcgt/gcaa/caag/aagc/agca	16.70	0.596003
ggat/gatg/atgg/tgga/ccta/ctac/tacc/acct	14.30	0.51035
gtca/tcag/cagt/agtc/cagt/agtc/gtca/tcag	19.10	0.681656
tgac/gact/actg/ctga/actg/ctga/tgac/gact	9.10	0.324768
ttgt/tgtt/gttt/tttg/aaca/acaa/caaa/aaac	14.30	0.51035

APPENDIX III

List of sorghum EST-SSR alleles corresponding to functionally discribed genes

GI	*Putative Functions*	*E-Value*
Apoptosis		
30161644	Bax inhibitor-1	2.00E-42
18064484	Seven transmembrane protein Mlo8	7.00E-04
Auxin regulated		
33108151	Auxin response factor 8	4.00E-62
30970161	Putative auxin transporter PIN1	1.00E-58
14823899	Auxin response factor 6b	7.00E-39
9305677	Indoleacetic acid-inducible protein homologue	1.00E-24
31330911	Putative auxin response factor 10	1.00E-23
Carbon/Nitrogen metabolism		
45968964	Putative chloroplast-targeted beta-amylase	7.00E-119
45958604	Nitrate reductase	3.00E-118
57810422	Pyruvate decarboxylase	2.00E-103
45996968	Carbohydrate kinase -like	8.00E-103
30979498	Fructose-1,6-bisphosphatase, cytosolic	2.00E-98
37754249	Citrate synthase	4.00E-90
30979203	Fructokinase 2	3.00E-86
57821524	Dtdp-glucose 4,6-dehydratase	3.00E-85
34439660	Isocitrate dehydrogenase	9.00E-84
9298140	Putative potassium channel beta subunit	1.00E-83
9298140	Putative potassium channel beta subunit	1.00E-83
30970468	1-aminocyclopropane-1-carboxylate oxid	1.00E-80
57815730	Glucan endo-1,3-beta-glucosidase precursor	2.00E-77
57822031	Putative acetyl-coa C-acyltransferase	1.00E-76
30964079	Pyruvate decarboxylase	7.00E-76
57811585	Galactinol synthase 3	7.00E-75
30972333	Glucose-6-phosphate isomerase	1.00E-74
45958473	Nitrate reductase	5.00E-72
30947299	Putative Aconitate hydratase	3.00E-71
30947299	Putative Aconitate hydratase	3.00E-71
45988935	Beta-glucosidase aggregating factor precursor	3.00E-64
34445645	Putative chitin-inducible gibberellin-responsive protein	3.00E-60
57815449	Putative acetyl transferase	4.00E-59
57817493	Fructose-1,6-bisphosphatase	6.00E-57
45990009	Beta-1,3-glucanase	2.00E-56
30165030	Glutamine-dependent asparagine synthetase 1	5.00E-56
9296214	Xyloglucan endo-transglycosylase/hydrolase	2.00E-51

8089064	Endo-1,3-beta-glucanase	6.00E-51
30940453	Beta-galactosidase	2.00E-50
7536147	Putative xyloglucan endotransglycosylase	2.00E-49
16960247	Glossy1 protein	1.00E-47
57806675	Sucrose synthase-2	2.00E-43
9303289	Nitrate-induced NOI protein	8.00E-41
12692087	GADPH	2.00E-38
12618760	Glutamine synthetase	2.00E-37
11922867	Phosphoglycerate kinase	1.00E-36
14689250	Sucrose synthase 2	4.00E-34
37705565	Sucrose synthase 3	5.00E-34
30162086	Putative glucose-6-phosphate/phosphate translocator	2.00E-33
30164695	Endo-1,3-beta-glucanase	2.00E-31
34446051	Beta3-glucuronyltransferase	8.00E-31
11996789	Putative endo-1,4-beta-glucanase	3.00E-27
57816905	Putative starch synthase III	1.00E-23
14569915	Pyruvate dehydrogenase kinase isoform 1	5.00E-23
57815251	Endoxyloglucan transferase	1.00E-22
33110564	Protein-O-fucosyltransferase 1	6.00E-21
13784416	Phytochelatin synthetase-like protein	1.00E-19
6858577	Pyruvate dehydrogenase E1 beta subunit isoform 1	1.00E-19
18052677	Putative adomet synthase 3	2.00E-15
18052677	Putative adomet synthase 3	2.00E-15
18052677	Putative adomet synthase 3	2.00E-15
9851507	Acetyl-coenzyme A carboxylase ACC2B	9.00E-12
13238600	Cytosolic 3-phosphoglycerate kinase	7.00E-08
45987734	Putative hexokinase I	8.00E-05
9299315	Enoyl coa hydratase-like protein	2.00E-02
13394142	Hexokinase	4.90E-01

Cell division		
12502377	Tousled-like kinase 2	2.00E-111
34509749	P34cdc2	4.00E-102
30971058	Beta3 tubulin	5.00E-85
61115535	Putative DIM-like protein	6.00E-73
12618778	Beta tubulin	1.00E-72
57808883	Putative microtubule-associated protein	3.00E-66
57804515	Putative microtubial binding protein	1.00E-58
7219375	Putative spastin protein orthologue	5.00E-48
6858805	Putative tubulin folding cofactor A	2.00E-47
34512433	Profilin A	2.00E-47
30952963	D-type cyclin	2.00E-46

34511627	Alpha-tubulin	4.00E-44
11679522	Putative microtubule bundling polypeptide TMBP200	4.00E-38
45979376	Profilin A	2.00E-24
11680215	TPA: putative phytosulfokine peptide precursor	1.00E-12
16960637	TPA: putative phytosulfokine peptide precursor	1.00E-10
30939303	Alpha-tubulin	4.00E-08
11678504	Meiosis 5	2.00E-06
Cellular metabolism		
45952805	Putative ATP synthase	1.00E-91
30942928	Putative endoglucanase 1 precursor	1.00E-69
33109421	Exoglucanase precursor	3.00E-52
9849411	Putative zeta-carotene desaturase precursor	2.00E-50
57819382	Short-chain alcohol dehydrogenase	2.00E-38
Cellular regulation		
45983050	Putative methyltransferase	7.00E-114
45954604	O-methyltransferase	2.00E-111
45988651	Putative heme A:farnesyltransferase	7.00E-85
30970276	Salt-induced MAP kinase 1	5.00E-74
13152814	Putative gibberelin 20-oxidase	2.00E-19
13152814	Putative gibberelin 20-oxidase	2.00E-19
31331209	Phytoene synthase radicle isoform	4.00E-18
Compatible solutes (protection)		
45960425	Mitochondrial aldehyde dehydrogenase	6.00E-142
30967045	Alcohol dehydrogenase 1	1.00E-74
31385253	Glutathione S-transferase GST 33	3.00E-73
6859304	Alcohol dehydrogenase 2	9.00E-46
30934016	Mitochondrial aldehyde dehydrogenase	4.00E-42
12617813	Delta 1-pyrroline-5-carboxylate synthetase	7.00E-33
61115273	Alcohol dehydrogenase	4.00E-26
12507566	Aldehyde dehydrogenase family 7 member A1	1.90E+00
Defense		
30942646	Glutathione transferase III	1.00E-86
13316299	Pathogenesis-related protein 10c	8.00E-82
31344905	Aspartic proteinase	1.00E-76
30976649	Acidic class III chitinase oschib3a	4.00E-66
45967289	Germin-like protein 7 ,putative Cupin family protein	9.00E-63
9303092	Glutathione transferase	1.00E-62

37709341	Chitinase	4.00E-61
30971846	Zeamatin-like protein	1.00E-56
57806643	NBS-LRR type resistance protein - barley	2.00E-53
45947083	Putative germin protein	8.00E-50
45948922	TPA: putative cystatin	3.00E-45
31348056	Glutathione transferase III	1.00E-44
34518586	D-mannose binding lectin, putative	2.00E-44
34445779	Chitin-inducible gibberellin-responsive protein	2.00E-40
7550926	Germin-like protein 1	9.00E-37
8089187	Putative type-1 pathogenesis-related protein	4.00E-31
7660358	Chitinase	3.00E-26
9301655	Glutathione transferase	5.00E-25
46931958	Thaumatin-like protein TLP8	5.00E-25
7554453	Putative wound-induced protease inhibitor	2.90E-01
7554453	Putative wound-induced protease inhibitor	2.90E-01

Dehydration		
57808808	Putative late embryogenesis abundant protein	3.00E-68
37706449	Dehydrin	6.00E-25
37706449	Dehydrin	6.00E-25
37753893	Dehydrin	5.00E-07
34444181	22 kda drought-inducible protein	1.00E-04

DNA structure		
9854629	Histone deacetylase HDA101	2.00E-59
30973285	Putative oligouridylate binding protein	3.00E-56
30942314	Putative methyl-binding domain protein MBD115	5.00E-54
45977343	Histone H3	8.00E-53
30162185	Single myb histone 1	2.00E-50
34442580	4-coumarate coenzyme A ligase	2.00E-47
11409147	Teosinte glume architecture 1	6.00E-44
45953029	HD2 type histone deacetylase HDA106	2.00E-42
30972653	Putative nucleotide binding protein	6.00E-42
30951882	Putative histone H2B	5.00E-40
34440765	INDETERMINATE-related protein 10	2.00E-31
30938414	INDETERMINATE-related protein 1	8.00E-08
34446275	Putative KH domain protein	5.20E-01
34511267	COG0272: NAD-dependent DNA ligase	8.40E+00

Ethylene regulated		
31334551	Ethylene-responsive small GTP-binding protein	2.00E-52
31334551	Ethylene-responsive small GTP-binding protein	2.00E-52

31334551	Ethylene-responsive small GTP-binding protein	2.00E-52
11678391	24-methylenesterol C-methyltransferase 2	1.00E-38
Flowering		
34442907	MADS-box protein FDRMADS3	4.00E-61
31344730	MADS-box protein RMADS214	2.00E-33
61115530	YABBY protein	2.00E-09
Growth/cytosceleton/cell wall		
45963247	Cell wall invertase; beta-fructosidase	2.00E-110
45969283	WD40 repeat protein	3.00E-106
57807017	Cellulose synthase catalytic subunit 10	3.00E-100
45957003	Beta-expansin 7	1.00E-98
13784637	Reversibly glycosylated polypeptide	2.00E-90
13784637	Reversibly glycosylated polypeptide	2.00E-90
61099173	Alpha-expansin 9 precursor	4.00E-82
31376905	Putative laccase	8.00E-82
30973033	Putative callose synthase 1 catalytic subunit	6.00E-74
57810337	Putative cellulase	1.00E-71
34518356	Alpha-1,4-glucan-protein synthase	3.00E-70
34518356	Alpha-1,4-glucan-protein synthase	3.00E-70
57806265	Cellulose synthase bocesa1b	1.00E-64
34441486	Putative actin	5.00E-62
31346700	Putative actin-depolymerizing factor 1	3.00E-59
30951820	Putative SF21C1 protein	5.00E-53
33110858	Putative cellulose synthase	4.00E-51
57810642	Putative actin related protein 2	7.00E-50
12506694	Cinnamyl alcohol dehydrogenase	4.00E-48
12506694	Cinnamyl alcohol dehydrogenase	4.00E-48
45953285	Beta-expansin 8	2.00E-47
45986604	Beta-expansin 8	3.00E-46
37705479	GA 2-oxidase 5	5.00E-43
30974808	Putative LEUNIG	1.00E-42
12499342	Putative beta-expansin	4.00E-42
34517690	Beta-expansin 1 protein	6.00E-40
45986116	Putative beta-expansin	2.00E-25
34518585	Beta-expansin 7	3.00E-05
Lipid biosynthesis		
57815161	Putative phosphoethanolamine N-methyltransferase	3.00E-105
30975228	Aminoalcoholphosphotransferase	7.00E-77
13469180	Allene oxide synthase	4.00E-58

34517721	Myo-inositol 1-phosphate synthase	1.00E-30
30974481	ASR protein	4.00E-03
Lipid metabolism		
57814489	Putative lipase	1.00E-88
45986907	GDSL-motif lipase/hydrolase-like	1.00E-87
45962881	Putative stearoyl-acyl-carrier protein desaturase	5.00E-87
31330672	Lipoxygenase	2.00E-86
57806704	Phospholipase C	2.00E-75
57810049	Putative phospholipase D	4.00E-61
37754053	Putative latex protein allergen	2.00E-55
13239168	Putative lysophospholipase 2	8.00E-55
45976443	Putative acyl-coa oxidase	5.00E-52
45994116	Fatty acid desaturase	1.00E-45
30969082	Putative early nodulin	3.00E-42
9853895	Putative acyl-coa oxidase	4.00E-41
33108116	Putative phosphoinositide-specific phospholipase C	1.00E-38
57804762	Putative lipid transfer protein	2.00E-34
13469650	Phospholipase C	1.00E-29
13238847	Putative lipid transfer protein	9.00E-24
14570456	Putative fatty acyl coa reductase	2.00E-20
30164702	Putative fatty acid desaturase	6.00E-18
30164702	Putative fatty acid desaturase	6.00E-18
30972266	Fatty acyl coa reductase	5.00E-17
Membrane protein		
45948467	Stomatin-like protein	2.00E-108
57821376	Putative auxin efflux carrier protein	5.00E-99
12505358	Cytochrome P450 88A1	3.00E-87
45965553	Cytochrome P450 71E1	6.00E-86
45969207	Cytochrome P450 79A1	3.00E-81
45986595	Cytochrome P450	1.00E-79
9303162	Putative glycosyl transferase	2.00E-78
45955866	Senescence-associated protein DH	1.00E-72
57807799	Putative cytochrome P450	4.00E-66
31331400	Putative SAC domain protein 9	1.00E-65
61099191	Cytochrome b5	1.00E-64
45946851	Putative elicitor-inducible cytochrome P450	4.00E-61
45985120	Putative cytochrome P450 monooxygenase CYP92A1	2.00E-53
7550367	Putative calmodulin	7.00E-51
45992006	Cytochrome P450 71E1	1.00E-50
45980372	Ubiquinol-cytochrome c reductase iron-sulfur subunit, mitochondrial	9.00E-48

13586881	ORMDL family protein-like	2.00E-47
37708185	Cytochrome c	3.00E-46
12499780	Putative syntaxin 6	8.00E-36
15724686	Cytochrome P450 78A1	7.00E-35
45989416	Probable cytochrome P450 monooxygenase	3.00E-26
12509402	Cytochrome P450 CYP99A1	4.00E-26

Metal binding protein		
45955755	Putative GTP-binding protein	7.00E-87
30934718	Putative inorganic pyrophosphatase	2.00E-83
45986466	Anthocyanidin synthase	5.00E-81
30950647	VEF family protein	1.00E-78
30977823	Flavonoid 3'-hydroxylase	1.00E-74
34442828	Putative inorganic pyrophosphatase	8.00E-74
13152547	Serine/threonine protein phosphatase PP2A-4 catalytic subunit	6.00E-73
33109600	Putative GTP-binding protein	9.00E-61
30163260	Molybdenum cofactor biosynthesis protein	5.00E-59
31385042	Putative GTP-binding protein	1.00E-55
30977988	Anthocyanidin synthase	8.00E-55
30977988	Anthocyanidin synthase	8.00E-55
11552796	Serine/threonine protein phosphatase PP2A-1 catalytic subunit	4.00E-39
8088267	Mitochondrial uncoupling protein 4	1.00E-37
13392407	Putative cis-zeatin O-glucosyltransferase	6.00E-34
9305649	Blue copper-binding protein -like	1.00E-29
9302259	Putative branched-chain amino acid aminotransferase	7.00E-28
12775332	Putative constans	2.00E-27
9297892	Annexin ANXC3.2	2.00E-21
13394832	Metallothionein-like protein	2.00E-19
11680091	Protein binding / ubiquitin-protein ligase/ zinc ion binding	3.00E-19
57810400	GA 3beta-hydroxylase	3.00E-16
18062190	Serine/threonine protein phosphatase PP1	1.00E-14
13152089	RING-H2 finger protein	3.00E-13
61115337	Metallothionein-like protein	3.00E-11

Nucleotides		
31345759	Phosphoribosyl pyrophosphate synthetase	2.00E-118
34446113	26S proteasome regulatory particle triple-A atpase subunit4	4.00E-100
30933138	UDP-glucose-4-epimerase	2.00E-87
30963920	Phosphoribosyl pyrophosphate synthetase	2.00E-77
30973885	Adenosine 5'-phosphosulfate reductase 6	8.00E-77
31346738	Cinnamoyl-coa reductase	4.00E-76
31346738	Cinnamoyl-coa reductase	4.00E-76

31346738	Cinnamoyl-coa reductase	4.00E-76
18062608	UDP-glucuronic acid decarboxylase	6.00E-66
13152417	UMP synthase	3.00E-62
13152583	ATP sulfurylase	3.00E-38
11678332	UDP-glucose pyrophosphorylase	3.00E-31
30974006	Putative CTP synthase	3.00E-26
12503626	Putative UDP-glucose 4-epimerase	1.00E-23
34517205	Putative nucleoside diphosphate kinase	3.00E-20
34514210	Putative UDP-glucosyltransferase	9.00E-18
30973894	Adenosine 5'-phosphosulfate reductase	1.00E-09

Photosynthesis		
45996858	Putative chlorophyll a/b-binding protein precursor	2.00E-104
30933007	Malate dehydrogenase, cytoplasmic	3.00E-98
57817285	Malate dehydrogenase (NADP(+))	3.00E-89
31346021	Putative ubiquitin-conjugating enzyme	5.00E-89
45965641	LHCI-680, photosystem I antenna protein	1.00E-87
45965641	LHCI-680, photosystem I antenna protein	1.00E-87
45989137	Ribulose 1,5-bisphosphate carboxylase/oxygenase small subunit	3.00E-86
45968710	Chlorophyll a/b-binding protein precursor	2.00E-84
9849422	Aerobic Mg-protoporphyrin IX monomethyl ester cyclase	2.00E-84
9849422	Aerobic Mg-protoporphyrin IX monomethyl ester cyclase	2.00E-84
31346514	Chlorophyll a-b binding protein 1, chloroplast precursor	6.00E-84
57817110	PSI type III chlorophyll a/b-binding protein	8.00E-84
9848657	Photosystem II thylakoid membrane protein	4.00E-82
57811809	Uroporphyrinogen decarboxylase, chloroplast precursor	9.00E-82
14089180	Phosphoenolpyruvate carboxylase	1.00E-81
30967782	Adenylosuccinate synthetase, chloroplast precursor	4.00E-80
9854168	Ubiquitin-conjugating enzyme	2.00E-74
31345902	Chaperonin precursor	5.00E-74
30945913	Putative 33kda oxygen evolvingprotein of photosystem II	6.00E-72
57806255	Putative ubiquitin-conjugating enzyme family protein	2.00E-71
7535675	Photosystem II type II chlorophyll a/b binding protein	2.00E-65
18063337	Ribulose-5-phosphate-3-epimerase	2.00E-64
45982447	Phosphoribulokinase precursor	1.00E-63
31331626	NADP-dependent malic enzyme	4.00E-62
12497812	Ubiquitin-conjugating enzyme osubc5a	4.00E-55
12497812	Ubiquitin-conjugating enzyme osubc5a	4.00E-55
31334597	Ferredoxin-NADP+ reductase	3.00E-53
34513532	Putative thioredoxin	1.00E-52
57810023	Ubiquitin-conjugating enzyme -like	1.00E-49
37755406	Putative chaperonin 21 precursor	2.00E-49

31329920	Putative Ubiquitin ligase SINAT5	1.00E-48
33110663	Putative chaperonin 60 beta precursor	6.00E-47
33110663	Putative chaperonin 60 beta precursor	6.00E-47
13318423	Triose phosphate/phosphate translocator	3.00E-46
13065305	Chlorophyll a/b binding protein	7.00E-46
13239099	Ubiquitin conjugating enzyme	8.00E-46
33107206	Chlorophyll a/b-binding protein CP29	1.00E-43
45969374	Photosystem i reaction centre subunit n, chloroplast precursor	7.00E-43
9851435	Photosystem I protein-like protein	1.00E-40
31384839	Chlorophyll a/b-binding protein CP29 precursor	2.00E-40
45988399	Peroxiredoxin Q	1.00E-39
30165520	Photosystem I reaction center subunit XI, putative	7.00E-36
13588229	Photosystem I reaction center subunit III, chloroplast precursor	4.00E-34
33110889	Chlorophyll a/b-binding protein CP29 precursor	7.00E-32
9851546	Ferredoxin-1, chloroplast precursor	2.00E-30
57814556	Chlorophyll a-b binding protein CP24 10B, chloroplast precursor	9.00E-29
57820523	NADP-dependent malic enzyme	2.00E-28
7551117	Isopentenyl pyrophosphate isomerase	4.00E-27
6673839	Polyphenol oxidase	1.00E-26
17886733	Ferredoxin-NADP(H) oxidoreductase	4.00E-23
30162043	Glutamyl-trna reductase, chloroplast precursor	9.00E-23
11064539	Peroxiredoxin	2.00E-22
11064539	Peroxiredoxin	2.00E-22
18068872	Putative chlorophyll synthase	3.00E-22
18068872	Putative chlorophyll synthase	3.00E-22
45975671	Plastocyanin	4.00E-19
37759063	T complex polypeptide 1	7.00E-17
34439792	Type II light-harvesting chlorophyll a /b-binding protein	2.00E-16
33109119	NADP-dependent malic enzyme	4.00E-15
57821918	Chloroplast phytoene synthase 1	8.00E-15
34515133	Ubiquitin-conjugating enzyme -like	3.00E-13
7551441	Phosphoenolpyruvate carboxylase kinase	4.00E-04
1000762	Ferredoxin I (Fd)	

Protein biosynthesis		
45951188	Putative threonyl-trna synthetase	9.00E-139
57803465	Putative glutamate decarboxylase isozyme	4.00E-124
57805559	Putative ribosomal protein L5	7.00E-106
45959935	Cytoplasmic ribosomal protein L18	5.00E-94
57805224	Translational elongation factor Tu	2.00E-92
34446563	Putative splicing factor Prp8	3.00E-89
12617631	Ribosomal protein L3	1.00E-82

30965856	Cytoplasmatic ribosomal protein S13	9.00E-79
45964968	Putative 40S ribosomal protein S15	2.00E-78
31348206	Putative 60S ribosomal protein L11	2.00E-73
45948168	Putative 60S ribosomal protein	3.00E-71
30948055	Ribosomal protein S11	2.00E-69
30162221	No apical meristem protein, putative	3.00E-67
45967335	Ribosomal protein s6 RPS6-2	3.00E-61
11679489	Putative 60S ribosomal protein L13E	1.00E-60
31347160	Translation initiation factor 5A	3.00E-59
34509754	Ribosomal protein L2	9.00E-57
10420990	60S ribosomal protein L21	4.00E-51
13587276	3-phosphoshikimate 1-carboxyvinyltransferase; EPSP-synthase	7.00E-51
13587276	3-phosphoshikimate 1-carboxyvinyltransferase; EPSP-synthase	7.00E-51
30964507	Eukaryotic translation initiation factor 1A	3.00E-50
12617935	Putative phenylalanine ammonia-lyase	1.00E-48
31330372	40S ribosomal protein S10	1.00E-48
9851295	Putative 60S ribosomal protein L35	1.00E-39
30944253	Eukaryotic initiation factor	2.00E-39
30977141	Putative homeotic protein	1.00E-38
30971555	Myb protein	8.00E-35
34518823	Cytoplasmatic ribosomal protein S13	5.00E-34
30160750	Mlip15	4.00E-33
34440676	S-like rnase	4.00E-32
13316561	Eukaryotic translation initiation factor 3 large subunit	8.00E-32
57807246	Putative ribosomal protein S29	5.00E-28
30966275	Acidic ribosomal protein p2a-2	4.00E-27
30162928	Putative gamma-lyase	2.00E-26
30162928	Putative gamma-lyase	2.00E-26
30936943	Plastid RNA polymerase sigma factor	4.00E-23
30932407	OCL4 protein	3.00E-22
9305658	Putative 60S ribosomal protein L39	2.00E-18
30969305	Putative MCB2 protein	2.00E-16
7218552	Putative aminotransferase	4.00E-16
31345885	Translation initiation factor 5A	8.00E-14
14514840	Ovule development protein aintegumenta (ANT)-like	1.00E-13
37752711	ATP-binding cassette, sub-family F	6.00E-13
37706906	Ribosomal protein s6 RPS6-2	1.00E-07
33109192	Putative homeodomain protein	8.00E-06
45978299	Putative nuclear RNA binding protein A	9.00E-04
7552454	Maize Em binding protein-1a	5.00E-03
31330196	Putative U2 snrnp auxiliary factor	5.20E-01
57809484	Knotted1-like homeodomain protein liguleless4a	1.60E+00
Protein folding/turnover		
9852612	Heat shock protein 70	2.00E-92
30964709	Dnaj-related protein ZMDJ1	5.00E-85

12618780	Putative cyclophilin	1.00E-57
13587836	Putative oligosaccharyl transferase STT3	4.00E-56
13318097	Polyubiquitin	2.00E-49
31376875	Heat shock protein 17.2	4.00E-45
34443526	Putative cytosolic chaperonin delta-subunit	9.00E-40
9854138	Dnaj-related protein ZMDJ1	3.00E-38
34511603	Cyclophilin 1; cyp1	2.00E-33
7218380	Polyubiquitin	5.00E-33
13469569	Protein disulfide isomerase	3.00E-06
57806181	Protein tyrosine phosphatase a	4.80E-01
37710385	Protein disulfide isomerase 4	3.30E+00
Protein metabolism		
34508913	Osjnba0038o10.7	4.00E-87
45991373	Serine-type carboxypeptidase	2.00E-79
45991373	Serine-type carboxypeptidase	2.00E-79
13152717	Aspartic endopeptidase Pep2	9.00E-73
13152717	Aspartic endopeptidase Pep2	9.00E-73
13152717	Aspartic endopeptidase Pep2	9.00E-73
30946863	Phosphatidylinositol 3,5-kinase-like	7.00E-62
45997018	Serine-type carboxypeptidase	1.00E-59
45997018	Serine-type carboxypeptidase	1.00E-59
30977272	Serine-type carboxypeptidase	2.00E-57
13469378	Osjnba0091d06.13	5.00E-54
13469378	Osjnba0091d06.13	5.00E-54
34517149	Serine carboxypeptidase III, CP-MIII	1.00E-50
30974889	Nucellin-like aspartic protease	5.00E-43
30979418	Putative insulin degrading enzyme	4.00E-38
21788307	Serine carboxylase II-3	4.00E-30
18063613	Serine carboxypeptidase i precursor	4.00E-28
12775306	Putative prolylcarboxypeptidase, isoform 1	8.00E-26
34516063	Putative glutamate decarboxylase	2.00E-23
34511906	Osjnbb0103i08.19	3.10E+00
Regulatory kinase		
45960649	Putative protein kinase	3.00E-91
14570934	Calcium-dependent protein kinase	2.00E-50
11921509	Putative calcium-dependent protein kinase	7.00E-48
30975057	Mitogen-activated protein kinase	6.00E-47

31330241	Putative calcium-dependent protein kinase 2	1.00E-46
57810676	Putative protein kinase	3.00E-37
45963265	Putative protein kinase	3.00E-36
30165240	Putative protein kinase	3.00E-31
30165122	Calcium-dependent protein kinase zmcp	6.00E-24
9301954	Putative protein kinase	8.00E-23
9301954	Putative protein kinase	8.00E-23
8088891	NPK1-related protein kinase-like protein	4.00E-21
37708146	Putative serine/threonine kinase	2.00E-20
57813648	Putative protein kinase	6.00E-20
14569551	Putative calcium-dependent protein kinase	4.00E-19
11410103	Serine/threonine kinase	2.00E-13
18068462	Putative protein kinase	2.00E-13
9848706	Protein kinase CK2 regulatory subunit	4.00E-12
7535895	Mitogen activated protein kinase 6	1.00E-06
6677282	Putative protein kinase	5.00E-04
ROS (Reactive Oxygen Species)		
57807798	Peroxidase	8.00E-107
14328619	Mn-superoxide dismutase	5.00E-94
30161896	Putative NADPH-thioredoxin reductase	2.00E-88
45974534	TPA: class III peroxidase 58 precursor	2.00E-75
14593539	Putative peroxidase	1.00E-58
45959942	Putative peroxidase	2.00E-57
7218821	Putative glutathione peroxidase	2.00E-53
30976380	Putative peroxidase	3.00E-52
30969157	Peroxidase	5.00E-50
9300637	TPA: class III peroxidase 25 precursor	6.00E-47
9304639	Putative peroxidase	1.00E-44
9304639	Putative peroxidase	1.00E-44
34518316	GPX12Hv, glutathione peroxidase-like protein	4.00E-39
5043525	Putative peroxidase	6.00E-37
21788474	Putative peroxidase	1.00E-34
13238234	Catalase	3.00E-33
31334706	TPA: class III peroxidase 64 precursor	1.00E-27
30941293	TPA: class III peroxidase 90 precursor	2.00E-03
12510039	Peroxidase	2.60E-02
Signalling		
34510262	ADP-ribosylation factor 1	7.00E-98
30964991	Putative receptor kinase Lecrk	5.00E-88
30964991	Putative receptor kinase Lecrk	5.00E-88

45997071	Putative shaggy-like kinase	2.00E-76
31329462	Cytokinin oxidase 2	2.00E-68
14089228	Ethylene receptor	1.00E-61
57808789	Putative receptor protein kinase CRINKLY4 precursor	9.00E-60
31334672	RPT2-like protein	4.00E-54
14824072	Putative ROP family gtpase ROP2	4.00E-48
30967463	Putative Rop family gtpase, ROP7	4.00E-47
18061217	Nonphototropic hypocotyl 1	2.00E-43
30963824	Protein binding / signal transducer	1.00E-33
31331194	Putative Ser/Thr specific protein phosphatase 2A B regulatory	1.00E-27
57804829	Putative ras-related GTP-binding protein	2.00E-20
34511021	Putative Calcineurin B subunit	5.00E-15
31334778	RPT2-like protein	3.00E-12

Storage protein		
45952034	Lumenal binding protein cbipe3	2.00E-98
30947321	Lumenal binding protein cbipe3	2.00E-97
31385017	Endosperm lumenal binding protein	2.00E-65
45953562	Putative germin A	5.00E-41
14366836	Lumenal binding protein cbipe3	1.00E-34
13238899	Beta-kafirin	2.00E-33
13065532	Putative 24 kda seed maturation protein	4.00E-30
30964563	Vicilin-like embryo storage protein	7.00E-24
30972814	Vicilin-like embryo storage protein	2.00E-18
13237550	Embryo-specific protein	6.00E-14

Stress		
30968419	Heat shock protein 82	2.00E-96
12617891	18kda heat shock protein	2.00E-51
30933027	Putative heat shock factor	3.00E-43
30936750	17 kda class I small heat shock protein	1.00E-33
11065083	Putative cytosolic class II low molecular weight heat shock protein	7.00E-29
5043465	Putative heat shock protein 82	3.00E-24
7553051	18kda heat shock protein	3.00E-24
6859224	ER Hsp70 chaperone bip, putative	4.00E-23
9921299	Glycine-rich RNA-binding protein	6.00E-05
34444181	Abscisic acid- and stress-induced protein - rice	7.00E-05

Transcription factor		
31347189	Putative basic leucine zipper protein	3.00E-89
45994861	Transcription factor MYC7E	3.00E-88
30945290	Putative Myb-related protein Zm38	1.00E-79

45996008	Putative osnac1 protein	2.00E-75
57809186	Putative myb-related protein	1.00E-74
14688999	Putative finger transcription factor	8.00E-70
31334286	NAC domain-containing protein 67	4.00E-66
9302986	DNA-binding protein RAV2, putative	1.00E-65
45951209	Putative RAV-like B3 domain DNA binding protein	3.00E-64
57813459	Putative transcription factor	2.00E-62
57816210	Typical P-type R2R3 Myb protein	1.00E-61
13239208	Putative MADS-domain transcription factor	5.00E-61
11266470	Putative MADS-domain transcription factor	1.00E-60
30965500	Putative NAC domain protein NAC1	7.00E-56
30967473	Putative RING zinc finger protein	4.00E-55
14365957	Putative MADS-domain transcription factor	7.00E-55
14569891	Putative MADS-domain transcription factor	2.00E-54
45968979	EREBP-like protein	1.00E-53
45988690	RISBZ4	2.00E-52
9299618	Putative GATA-1 zinc finger protein	6.00E-50
45953970	Putative bzip protein	1.00E-49
11680392	RAPB protein	2.00E-48
14366631	Putative dehydration-responsive element binding protein 3	2.00E-46
14366631	Putative dehydration-responsive element binding protein 3	2.00E-46
11680392	CCAAT-box transcription factor complex WHAP5	6.00E-46
9304207	TPA: WRKY transcription factor 16	8.00E-45
34513916	Putative PTI1-like kinase	2.00E-44
7534857	Golden2-like transcription factor	1.00E-43
34513616	Putative zinc finger protein	3.00E-41
45965696	Zinc finger protein-like protein	5.00E-39
31344744	Transcription factor Myb3	1.00E-38
6858452	Putative MYB29 protein	4.00E-37
45947225	Homeodomain leucine zipper protein	5.00E-37
45973493	Putative ethylene-responsive element binding factor	5.00E-35
14366334	Putative MADS-domain transcription factor	2.00E-34
8090941	NAC protein	2.00E-34
13239643	WRKY transcription factor 44	6.00E-34
31333621	Helix-loop-helix protein	1.00E-32
57804614	Putative transcription factor (contains Myb-like DNA-binding	2.00E-31
13239024	Putative WOX2 protein	3.00E-30
33111086	Putative RNA polymerase I transcription factor RRN3	9.00E-30
18061163	Putative vascular plant one zinc finger protein	4.00E-28
30966237	Homeodomain leucine zipper protein hox1	1.00E-27
30948047	Putative MYB29 protein	5.00E-27
13784257	TPA: WRKY transcription factor 71	1.00E-25

13784257	TPA: WRKY transcription factor 71	1.00E-25
45953956	VIP1 protein	4.00E-25
30969418	Dehydration responsive element binding protein	3.00E-24
9302651	TPA: WRKY transcription factor 28	9.00E-24
31347843	Homeodomain leucine zipper protein	7.00E-21
34440738	ANAC083; transcription factor	5.00E-20
37755670	Putative NAM protein	2.00E-19
34446191	TPA: WRKY transcription factor 68	2.00E-17
7217600	CCAAT-box transcription factor complex WHAP13	6.00E-16
45996636	Homeodomain leucine zipper protein	3.00E-15
45948932	NAC domain transcription factor	6.00E-15
13392429	Putative myb-related protein	1.00E-14
9848022	Putative nucleic acid-binding protein	4.00E-14
7536201	Putative AP2 domain containing protein RAP2.1	9.00E-14
7536201	Putative AP2 domain containing protein RAP2.1	9.00E-14
57821014	Putative homeodomain leucine zipper protein	2.00E-11
31346326	Homeodomain protein JUBEL1	1.00E-10
30160762	Homeodomain leucine zipper protein	3.00E-08
37709309	TPA: WRKY transcription factor 47	2.00E-03
10421634	Transcription factor MYC7E	1.30E+00
37707773	Transcriptional regulator, gntr family	1.50E+00
37755206	Similar to Myc box dependent interacting protein 1	2.10E+00
11920742	Runt-related transcription factor 1-like protein	3.20E+00

Transport protein		
57820839	Flavin containing polyamine oxidase	5.00E-94
33110681	Phosphate transporter	2.00E-93
45951408	Putative chaperonin	1.00E-91
57813793	Putative high affinity nitrate transporter	1.00E-91
34509905	Putative transmembrane protein Tmp21 precursor	1.00E-81
57811544	MDR-like ABC transporter	1.00E-68
45953242	Monosaccharide transport protein 1	2.00E-62
33109600	Ras-related protein RAB8-3	2.00E-60
30978622	Putative high affinity nitrate transporter	4.00E-60
12617413	Monosaccharide transporter 1	2.00E-58
14365762	Vacuolar atpase B subunit	1.00E-55
13469378	Vacuolar processing enzyme	5.00E-54
37705452	Zinc transporter	6.00E-53
37705452	Zinc transporter	6.00E-53
57807578	Inorganic phosphate transporter 2	2.00E-52
30972806	Iron transport protein 2	2.00E-52
14513257	Vacuolar ATP synthase subunit B isoform 1	2.00E-50

30940411	Putative Sec61	9.00E-50
13239529	Vacuolar H+-translocating inorganic pyrophosphatase	1.00E-49
18061116	Proline transporter, putative	2.00E-46
12776176	NOD26-like membrane integral protein zmnip1-1	1.00E-45
13389580	Putative Na+/H+ antiporter	4.00E-45
45988222	Flavin containing polyamine oxidase	9.00E-45
30941897	Putative Na+-dependent neutral amino acid transporter	1.00E-40
30941897	Putative Na+-dependent neutral amino acid transporter	1.00E-40
18066945	Major facilitator superfamily antiporter	6.00E-40
57813700	High affinity nitrate transporter	2.00E-38
34509294	Putative Sec61 alpha form 2	3.00E-38
12617840	Coatomer delta subunit	1.00E-35
57818888	Putative mitochondrial carrier protein	4.00E-35
34517120	Transmembrane protein, putative	2.00E-34
34447016	Putative potassium transporter KUP3p	1.00E-33
45946764	Gamma-TIP-like protein	4.00E-31
30933884	Aquaporin	9.00E-30
61115239	NOD26-like membrane integral protein	5.00E-21
34518290	Vacuolar atpase subunit c isoform	6.00E-21
13587223	Putative vacuolar ATP synthase subunit H	3.00E-19
13587909	PDR-like ABC transporter	6.00E-18
11921741	Plasma membrane H+ atpase	5.00E-09

Other classes		
31329924	Hypothetical protein FG09178.1	5.00E-108
45979328	Putative beta-ketoacyl-coa synthase	5.00E-102
57811026	2 coiled coil domains of eukaryotic origin (31.3 kd)-like protein	3.00E-95
45965641	PREDICTED OJ1065_B06.19-1 gene product	5.00E-91
34517259	Mob1-like protein	1.00E-90
45990071	Diadenosine 5',5'''-P1,P4-tetraphosphate hydrolase	7.00E-89
45965641	PREDICTED OJ1065_B06.19-1 gene product	1.00E-87
31346850	Prefoldin subunit 3, putative	2.00E-85
31329646	Hypothetical protein FG06473.1	5.00E-79
57810817	Putative MATE efflux family protein	3.00E-78
31329398	Hypothetical protein	3.00E-78
57813135	Mob1-like protein	8.00E-75
9296077	Glycosyltransferase	3.00E-73
34441344	Putative gtpase activating protein	9.00E-73
9301798	Glycosyltransferase	6.00E-71
45947007	Response regulator 6	2.00E-70
31346890	Gb protein	4.00E-69
45958524	D-protein	3.00E-66

31384915	Hypothetical protein	1.00E-64
30966816	Lactate dehydrogenase	3.00E-62
31345707	Co-chaperone Hsc20, putative	4.00E-58
33109805	Hypothetical protein FG08579.1	3.00E-54
31383727	Hypothetical protein FG09718.1	3.00E-51
30973775	Cystathione gamma lyase, putative	2.00E-49
9297650	Chain E, Crystal Structure Of The Sorghum Bicolor Dhurrinase	7.00E-49
9297650	Chain E, Crystal Structure Of The Sorghum Bicolor Dhurrinase	7.00E-49
10421966	Cold acclimation protein COR413-TM1	1.00E-45
30967850	Putative purple acid phosphatase	2.00E-45
33107856	Hypothetical protein FG04969.1 [Gibberella zeae PH-1]	2.00E-45
30162257	Glycine dehydrogenase P protein	5.00E-45
7550392	Putative senescence-associated protein	2.00E-44
45948549	Cyclin, N-terminal domain, putative	2.00E-43
34514038	SNAP-28	3.00E-43
7658929	Putative dihydrolipoamide dehydrogenase precursor	3.00E-40
57809755	Putative protein	1.00E-39
11679325	Putative apical-basal pattern formation protein	1.00E-39
45950931	DNA-binding protein	1.00E-38
45976774	Putative nascent polypeptide associated complex alpha chain	2.00E-38
13239393	Lipid transfer protein	6.00E-38
45959826	Hypothetical protein dgeodraft_0651	7.00E-38
33109767	Respiratory burst oxidase homolog	2.00E-37
31346012	Carbonic anhydrase	5.00E-37
8089713	Dolichol-phosphate mannosyltransferase-like	2.00E-36
18066826	NB-ARC domain, putative	7.00E-36
13317617	Transglutaminase	1.00E-34
57811893	Cold acclimation protein WCOR518	2.00E-33
11679191	Aux/IAA protein	9.00E-33
57803878	Putative pumilio/Mpt5 family RNA-binding protein	8.00E-31
11552374	Silencing group B protein	1.00E-30
30977329	Hypothetical protein	2.00E-29
37705153	Cuticle protein	6.00E-28
5043004	Bacterial flagellar motor protein	1.00E-27
18068479	Hypothetical protein	4.00E-26
11678260	Shugoshin-like protein	2.00E-25
18070440	Putative WD-repeat protein	1.00E-24
31330865	Putative pentatricopeptide	7.00E-23
11552588	Glycosyltransferase	9.00E-22
57812318	Cysteine proteinase	9.00E-22
34512675	Methylthioadenosine/S-adenosyl homocysteine nucleosidase	5.00E-21
30974675	Hypothetical protein	1.00E-20

30979067	Ids3	6.00E-20
30977595	Expressed protein	1.00E-18
34512482	Muscleblind-like 1 isoform d	2.00E-18
14689372	Arachidonic acid-induced DEA1	3.00E-18
6858522	Heme oxygenase 2	4.00E-18
14570985	Putative beta-ketoacyl-coa synthase	6.00E-18
33109062	Arabinoxylan arabinofuranohydrolase isoenzyme AXAH-II	2.00E-17
37705655	Zinc-induced protein-like	4.00E-17
30978072	Hypothetical protein	1.00E-16
37710605	Putative Sip1 protein	2.00E-15
2674238	Carbonic anhydrase	5.00E-15
45963218	Hypothetical protein	1.00E-14
30937893	NADPH HC toxin reductase-like	8.00E-14
33109829	Hypothetical protein rakah01001082	2.00E-12
13655854	DNA-binding protein	9.00E-12
31345592	DNA binding protein PF1	1.00E-11
7551234	AT1	5.00E-11
7551618	Hypothetical protein PY04653	4.00E-10
30974711	Lipid transfer protein	1.00E-09
31333104	Hypothetical protein	3.00E-09
10420220	Cold shock protein-1	1.00E-07
33107406	CG11430-PB, isoform B	5.20E-02
34445586	Fibroin heavy chain precursor	6.80E-02
34511267	Hypothetical protein	2.00E-01
14328678	MT-like protein	5.00E-01
30951753	Hypothetical protein	5.70E-01
34510491	Hypothetical protein	5.90E-01
13065299	Hypothetical protein LOC324010	1.10E+00
30971355	IAA1 protein	2.20E+00
45965429	Similar to Neuromedin U-25 precursor	2.40E+00
14592382	CG7918-PA	2.50E+00
30165447	Peplomer protein	2.90E+00
6673494	Mg chelatase-related protein	4.30E+00
9855886	Conserved hypothetical protein mppy	9.30E+00
57815324	Conserved hypothetical protein	9.60E+00

APPENDIX IV

PREPARATION OF STOCK SOLUTIONS

Extraction buffer	Tris HCl, pH 8.0	100 mM
	EDTA, pH 8.0	50 mM
	NaCl	500 mM
20% SDS	20 g of sodium doedycyl sulphate (sodium lauryl sulphate) dissolved in 100 ml of dH_2O. Should not be autoclaved	
5 M potassium acetate	5M potassium acetate	60.0 ml
	Glacial acetic acid	11.5 ml
	H_2O	28.5 ml
Isopropanol Ammonium Acetate mixture	Three volumes of 10 M Iso-propanol and 1 volume of 10 M ammonium acetate	
T_{10} E_1	Tris 10 mM containing 1 mM EDTA	
RNase (10 mg/ml)	Dissolve RNase in water, place in a tube in a boiling water bath for 10 minutes. Allow this to cool on a bench and store at -20^0C	
Chloroform isoamyl alcohol (24:1)	Chloroform 240 ml Isoamyl alcohol 10 ml. Store in a dark room temperature. Make up and dispense the solution in a fumed cupboard.	
Ethanol (70%)	Absolute alcohol	70 ml
	Distill water	30 ml
NaCl 5 M	Dissolve 292.2 g NaCl in 750 ml water. Make up to 1 litre with water, filter and autoclave.	
Phenol chloroform (24:1)	Mix equal volumes of the buffered phenol and chloroform isoamyl alcohol (24:1). Store at 4^0C	
Sodium acetate (2.5 M, pH 5.2)	Dissolve 340.2 g sodium acetate in 500 ml of water adjust pH to 5.2 with glacial acetic acid and make up the volume to 1 litre and autoclave.	
Tris HCl (1 M, pH 8.0)	Dissolve 121.1 g Tris in 800 ml of water. Adjust pH 8.0 with conc. HCl. Make up the volume to 1 L and autoclave	

Ethidium bromide (10 mg/ml)	Dissolve 100 mg ethidium bromide in 10 ml of distil water, warp tube in aluminium foil and store at 4^0C.	
4% acrylamide solution (1000 ml)	double distilled water	450 ml
	5x TBE	100 ml
	Urea	420 g
	40% Acrylamide/ bisacrylamide (19:1) (w/ w)	100 ml
	Combine water and TBE buffer in beaker, heat using a microwave, add urea and stir until dissolved. Adjust the volume to 900 ml with water and filter to remove any large particles. Then acrylamide solution is combined with other ingredients in a storage container.	
5x TBE	Tris base	540 g
	EDTA	46 g
	boric acid	276 g
	dH_2O to 10 L (store at room temperature, if left for long-periods of time, some of the salts will precipitate. It may be advisable to discard this buffer and make fresh buffer	
0.5% acetic acid in 95% ethanol	glacial acetic acid	1 ml
	95% ethanol	199 ml
	Aliquot and store at room temperature	
10% APS (ammonium persulfate)	ammonium persulfate	1 g
	dH_2O	10 ml
	Should be stored at -20^0C. If the APS will be used within the few weeks, store in light-tight bottle at 4^0C	
TEMED (N, N, N', N' – tetra methyl ethylene diamine)	Store at 4^0C	
Binding solution	Bind silane (γ - methacryloxy profile trimethoxy silane) M-6514 sigma industries	
	4 µl of bind silane + 1000 µl of 0.5% acetic acid in 95% ethanol	

Repel solution	Repel silane (dichloro dimethyl silane 99%) M-440272 sigma industries	
	250 µl repel silane + 750 µl of 0.5% acetic acid in 95% ethanol (should be done in fumehood)	
Loading dye/tracking dye (10x)	Sucrose	167 mg
	Bromophenol blue	4.2 mg
	Water	1 ml
3x SSR dye (3x STR loading solution)	5 M NaOH	0.2 ml
	95% formamide	95 ml
	Bromophenol blue	50 mg
	Xylene cyonol	50 mg
	Sd water make upto	100 ml
100 bp DNA ladder	100 bp marker (Genetix)	10 µl of
	3x SSR dye	95 µl
	sterile dH_2O	95 µl
Fix/stop solution (10% acetic acid)	glacil acetic acid in 1800 ml dH_2O	200 ml
Impregnate solution	glycerol	10%
	glacial acetic acid	10%
Staining solution	silver nitrate	2 g
	37% formaldehyde	3 ml
	ddH_2O	2000 ml
	Should be stored at room temperature in a cabinet or other dark storage space	
Developer solution	sodium carbonate	60 g
	37% formaldehyde	3 ml
	sodium thiosulphate (10 mg/ml)	400 µl
	dH_2O	2000 ml
	sodium carbonate	60 g
	This solution must be prepared fresh for each use. Prepare the solution by dissolving sodiumcarbonate in water and chilling to about 10^0C by placing on ice or in freezer. Just before use add the formaldehyde and sodium thiosulphate	

Loading dye (6x)	Bromophenol blue	0.25 g
	Sucrose in water	40% (W/V)
	Storted at 4°C	
50x TAE (Tris-Acetate EDTA)	Tris base	242 g/l
	Glacial acetic acid	57.1 ml
	0.5 M EDTA (pH 8.0)	100 ml
	Distilled water	1000 ml

IN SILICO EST DATAMINING FOR ELUCIDATION OF REPEATS BIOLOGY AND FUNCTIONAL ANNOTATION IN SORGHUM [*Sorghum bicolor* (L.) Moench.]

Arun S. S.

2006

Major Advisor
B. Fakrudin

ABSTRACT

A database of available sorghum ESTs was constructed and used to identify microsatellites or SSRs. A total of 2,32,921 ESTs from public databases, representing 35 different transcriptomes were downloaded. Repeat scan analysis identified 12,235 ESTs with repeats of varied type and length, which accounted for 5.25 per cent of all the ESTs in a redundant dataset. Clustering of this dataset removed redundancy, leaving a total of 3,281 unique genes containing microsatellites. TNRs (50.37%) were the most abundant types, followed by DNRs (42.9%) and TTNRs (6.72%). Among the DNRs, AG/CT was the most abundant type, accounting for 55.23 per cent of all DNRs, followed by AC/GT (21.94%). The CG repeats were rare at 6.63 per cent. Among TNRs, most abundant type was ATG/TAC (34.67%) followed by ATC/TAG (17.42%). The frequency of individual TTNRs varied from 9.11 per cent for ACGT/TGCA to 0.24% for ACCC/TGGG. Pairs of primers could be designed for 520 of the 3281 ESTs. The parental lines, IS22380 and E36-1 were scanned with a random set of 20 genic SSRs. All the 20 primer pairs produced the expected amplicon from the genomic DNA of both genotypes, of which 4 were polymorphic. The database constructed for the EST-SSRs was named as 'Jowar GenRepeat database' with 3281 ESTs with SSRs as records. Each record has complete information about the repeats, functional annotation and primers information in addition to information available in respective source database. This database is compatible for seach, using keywords in all fields, updateable and user friendly. Each EST was compared to annotated proteins in the databases with BLASTx algorithm and tentative function was assigned for 687 (20.93%) of the 3281 ESTs containing repeats. Of this lot of 687 ESTs, primer pairs have been designd for 135 annotated ESTs, which can be used for further applications in functional genomics.