In Silico Biology 6, 0054 (2006); ©2006, Bioinformation Systems e.V
http://www.bioinfo.de/isb/toc_vol_06.html

A database of simple sequence repeats from cereal and legume expressed sequence tags mined *in silico*: survey and evaluation

B. Jayashree[1*], Ramu Punna[2], P. Prasad[1], Kassahun Bantte[2], C. Tom Hash[2], Subhash Chandra[1], David A. Hoisington[2], Rajeev K. Varshney[2]

[1] Bioinformatics and Biometrics Unit, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru 502 324, Andhra Pradesh, India
[2] Applied Genomics Laboratory, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru 502 324, Andhra Pradesh, India

Email: b.jayashree@cgiar.org; r.punna@cgiar.org; p.prasad@cgiar.org; kbantte@yahoo.com; c.hash@cgiar.org; s.chandra@cgiar.org; d.hoisington@cgiar.org; r.k.varshney@cgiar.org

* Corresponding author

Edited by H. Michael; received July 24, 2006; revised and accepted October 17, 2006; published November 14, 2006

Abstract

Simple sequence repeats (SSRs) or microsatellites are an important class of molecular markers for genome analysis and plant breeding applications. In this paper, the SSR distributions within ESTs from the legumes soybean (*Glycine max*, representing 135.86 Mb), medicago (*Medicago truncatula*, 121.1 Mb) and lotus (*Lotus japonicus*, 45.4 Mb) have been studied relative to the distributions in cereals such as sorghum (*Sorghum bicolor*, 98.9 Mb), rice (*Oryza sativa*, 143.9 Mb) and maize (*Zea mays*, 183.7 Mb). The relative abundance, density, composition and putative annotations of di-, tri-, tetra- and penta-nucleotide repeats have been compared and SSR containing ESTs (SSR-ESTs) have been clustered to give a non-redundant set of EST-SSRs, available in a database. Further, a subset of such candidate EST-SSRs from sorghum have been tested for their ability to detect polymorphism between *Striga*-susceptible, stay-green drought tolerant mapping population parent 'E 36-1' and its *Striga*-resistant, non-stay-green counterpart 'N13'. Primer sets for 64% of the EST-SSRs tested produced a clear and specific PCR product band and 34% of these detected scorable polymorphism between the N13 and E 36-1 parental lines. Over half of these markers have been genotyped on 94 RILs from the (N13 × E 36-1)-based mapping population, with 42 markers mapping onto the ten sorghum linkage groups. This establishes the value of this database as a resource of molecular markers for practical applications in cereal and legume genetics and breeding. The primer pairs for non-redundant EST-SSRs have been designed and are freely available through the database (http://intranet.icrisat.org/gt1/ssr/ssrdatabase.html).

Keywords: EST-SSRs, microsatellite markers, conserved orthologous sets, tri-nucleotide repeats, class I repeats, class II repeats, cereals and legumes

Introduction

Simple sequence repeats (SSRs) or microsatellites are ubiquitous in eukaryotic genomes. SSRs are composed of tandemly repeated 1-6 bp long units. Because of their high mutability, SSRs are thought to play an active role in genome evolution by creating and maintaining genetic variability [1]. The presence of SSRs in the transcripts of genes suggests that they may have a role in gene expression or function [2]. For instance in the case of humans, it has been shown that variation in repeat units of SSRs (i) when present in 5'-UTR (untranslated region) - affects gene transcription and/or translation, (ii) when present in coding regions - inactivate or activate genes or truncate protein, and (iii) when present in 3'-UTR may be responsible for gene silencing or

transcription slippage [3]. Further, the number of tandem repeat runs has also been known to change with environmental challenges [4], in bacterial populations stress genes have been shown to contain more repeats than the rest of the genome [5]. Expansion and contraction of SSR repeats in genes of known function, therefore, can be tested for association with phenotypic variation or, more desirably, biological function [6]. Although in earlier studies, SSRs were reported more in non-coding regions than in coding regions of eukaryotes [7], a larger number of tri-nucleotide repeats have been reported in the coding regions of higher genomes in recent studies [3, 8, 9].

In experimental studies, SSRs are amplified by PCR (Polymerase Chain Reaction) with primers that are complimentary to the conserved sequences that flank the SSR loci. SSR polymorphism is the result of variability in the number of repeat units in a defined region of the genome being investigated [10]. The development of SSR primers requires a great deal of time, effort and investment in the construction and screening of genomic libraries and sequencing of clones containing SSR and testing of primers. However, the large number of EST sequences available in public databases provides for an alternative method of microsatellite development, SSRs can be electronically mined from EST databases [9]. SSR markers derived from ESTs are commonly called EST-derived SSRs (EST-SSRs). Since a putative function based on corresponding ESTs can be deduced for the EST-SSRs, they represent a class of 'functional markers' [11]. Frequency and distribution of EST-SSRs have been studied in many plant genomes [12, 13, 14] and have been developed and utilized for a variety of applications in several plant species [9]. Unlike genomic SSRs (derived from genomic libraries in a conventional manner), a considerable proportion of EST-SSR markers developed for a given species have been shown to display transferability in related plant species [15, 16, 17, 18]. Because of the origin of EST-SSR markers from the conserved portion of the genome, it is possible, after clustering them based on cross species homology, to develop conserved orthologous set (COS) markers for a given group of species. The COS markers based on EST-SSRs, in addition to being useful for trait mapping and diversity studies, may be used to better the understanding of genome evolution [19].

The present study was undertaken with the following objectives: (i) Mining the large EST collections of three dicot (soybean, medicago and lotus) and three monocot (sorghum, maize and rice) species for SSRs, (ii) study the frequency and distribution of EST-SSRs within and across the monocot and dicot species, (iii) identify non-redundant EST-SSRs, classify them based on their putative annotations, design primer pairs for the EST-SSRs across all six species, and make them available through a database (iv) validate the EST-SSRs through experimental studies and (v) identify putative COS markers across the monocot and dicot species.

Method

SSR detection and classification

The EST sequence data available in the public domain from three cereals namely rice (*Oryza sativa*), maize (*Zea mays*) and sorghum (*Sorghum bicolor*) and three legumes namely soybean (*Glycine max*), medicago (*M. truncatula*) and lotus (*Lotus japonicus*) were downloaded in fasta format from http://www.tigr.org between December 2004-January 2005. EST sequences less than 200 bp, because of poor quality, was not included in the analysis. The identification of microsatellites was carried out using the tool SSRIT [20]. Microsatellites greater than 12 bp were considered, which means there should be six occurrences of a di-nucleotide repeat, four occurrences of a tri-nucleotide repeat, three occurrences of a tetra- and 2.5 occurrences of a penta-nucleotide repeat. Simple scripts in VB (Visual Basic) were written to transfer SSRIT output into a relational database and identify SSR-containing ESTs. Query pages to the database were developed using ASP (Active Server Page).

The grouping of repeats into classes was carried out according to the method of Jurka and Pethiyagoda [21]. For example the tri-nucleotide repeat class TTC repeats [which includes $(AAG)_n$, $(TCT)_n$, $(CTT)_n$, $(AGA)_n$ and $(GAA)_n$] were considered equivalent when read in different reading frames and on the complementary strand.

Thus, di-nucleotide repeats could be grouped into four classes, tri-nucleotide repeats into 10 classes, tetra-nucleotide repeats into 33 classes and penta-nucleotide repeats into 102 classes. All SSR classes were analysed for their frequency of occurrence, density and relative abundance. Density was calculated by dividing the number of base pairs contributed by each SSR by total length analyzed (Mb). Relative abundance was calculated as number of SSRs per kb of sequence.

Clustering and primer design

The CAP3 program was used for the clustering of SSR-ESTs [22] to identify non-redundant SSR-ESTs and scripts were written to parse the output into the database. The resulting EST-SSRs were further clustered based on sequence homology using the standalone version of BLASTn, to give cross-species clusters of EST-SSRs. To be grouped as conserved orthologous sets, reciprocal best hits with a BLASTn cut-off value of 1e-5 and over 70% identity over the length of alignment were retained.

Putative identities for the contigs and singletons were obtained after searching the non-redundant protein sequence database with entries from GenPept, Swiss-Prot, PIR, PDF, PDB, and NCBI RefSeq; using BLASTx from the Paracel BLAST 1.6.1-paracel package on a Paracel 4 node Linux cluster. BLAST search descriptions were retained for the best hit if they met the criteria: E-value < 1e-10, at least 30% of query sequence aligned with >80% identity. The putative annotations were manually grouped into classes based on the biochemical function of the putative proteins. Those sequences classed under hypothetical/proteins of unknown function category were further searched against the respective Repeats database (http://www.tigr.org/tdb/e2k1/plant.repeats/) for annotation to repetitive elements. Primer pairs to the EST-SSRs were developed using the stand-alone version of Primer3 (Primer3.exe version 1.0; http://frodo.wi.mit.edu/primer3/binary-distributions.html). The data is housed in an SQL-2000 database and information can be retrieved through a suite of query pages available at http://intranet.icrisat.org/gt1/ssr/ssrdatabase.html.

Mapping EST-SSRs

DNA samples from the parents and a subset of 94 of 226 sorghum mapping progenies from the cross N13 × E 36-1, previously used to map host plant resistance to the parasitic weed *Striga hermonthica* [23] and to map the stay-green component of terminal drought tolerance [24], were used as template. PCRs were conducted in 5 µl reaction volumes containing 2.5 ng of genomic DNA template, final concentrations of 0.4 pM of both forward and reverse primers, 2 mM of $MgCl_2$, 0.1 mM of dNTPs, 1X buffer, and 0.1 U of AmpliTaq polymerase enzyme, and diluted to volume with doubled distilled water. The cycling conditions for PCR on a GeneAmp® PCR System 9700 (PE-Applied Biosystems) thermal cycler were optimized to initial denaturation of 15 min at 94°C, followed by 10 cycles (touchdown) of 94°C for 15 sec, annealing touchdown temperature reducing from 61 to 51°C for 20 sec over 10 cycles, with extension at 72°C for 30 sec. This was followed by denaturation at 94°C for 10 sec, annealing at 54°C for 20 sec, and extension at 72°C for 30 sec for 34 cycles, followed by final extension of 20 min at 72°C. PCR products were separated electrophoretically on polyacrylamide gels, visualized by silver staining [25], and scored manually. EST-SSR genotype data for 94 RILs in the (N13 × E 36-1)-based mapping population [23, 24, 26] was merged with previously generated RFLP, AFLP and SSR data, and linkage analysis was performed on the combined marker data set using MAPMAKER/EXP 3.0 [27].

Results

SSR frequency, relative abundance and density

The datasets obtained from public resources represented 98.9 Mb of sorghum (from 187282 ESTs), 183.7 Mb of maize (407423 ESTs), 143.9 Mb of rice (272567 ESTs), 45.4 Mb of lotus (109618 ESTs), 135.86 Mb of soybean (330436 ESTs) and 121.1 Mb from *Medicago* (226923 ESTs). On an average 19% of the ESTs from cereals (35-65,000 ESTs) and 10.6% of the ESTs from legumes (10-35,000 ESTs) were found to contain SSRs in the complete redundant set of ESTs analyzed. The frequency of SSRs observed under the conditions in this study amounted to 1 SSR/1.79 kb in sorghum, 1 SSR/2.21 kb in maize, and 1 SSR/1.72 kb in rice while in the three legumes considered in this study the frequency of occurrence was 1 SSR/3.5 kb. Thus the three cereal crops had a higher relative abundance (number of SSRs/kb of sequence) of SSRs compared to the legumes: 0.56 SSRs/kb in sorghum, 0.45 in maize, 0.58 in rice as compared to 0.29 in the legumes (Tab. 1). Over 60% of the predicted SSRs in cereals and legumes had repeat unit sizes that are multiples of three nucleotides. While comparing the two crop groups i. e. cereals and legumes, the cereals showed a higher frequency of tri- (1 SSR/2.9 kb in cereals and 1 SSR/5.6 kb in legumes), tetra- (1 SSR/8kb as compared to 1 SSR/15.2 kb in the legumes), and penta-nucleotide repeats (1 SSR/29.6 kb for 1 SSR/50.8 kb in legumes). The highest density (bp/Mb) of tri-nucleotide repeats was observed in rice (5560 bp/Mb) and sorghum (5027 bp/Mb) (Tab. 1), while the lowest density of tri-nucleotide repeats was found in soybean (2401 bp/Mb). Among the six crop species considered, rice enjoyed the highest combined high densities of di- and tri-nucleotide repeats.

Table 1: Relative abundance and density of SSRs in the EST datasets.

	Sorghum	Maize	Rice	Average across cereals	Soybean	Medicago	Lotus	Average across legumes
Total length analyzed (Mb)	98.9	183.7	143.9	142.1	135.9	121.1	45.4	100.8
Relative abundance (SSRs/kb)	0.56	0.45	0.58	0.53	0.29	0.29	0.28	0.29
Di-nucleotide (SSRs/kb)	0.02	0.02	0.03	0.02	0.03	0.02	0.02	0.02
Tri-nucleotide (SSRs/kb)	0.36	0.28	0.40	0.35	0.17	0.18	0.18	0.18
Tetra-nucleotide (SSRs/kb)	0.13	0.12	0.12	0.12	0.07	0.07	0.06	0.07
Penta-nucleotide (SSRs/kb)	0.04	0.04	0.03	0.04	0.02	0.02	0.02	0.02
Di-nucleotide density (bp/Mb)	431.2	391.3	569.7	464	691.3	371.6	420.5	494.5
Tri-nucleotide density (bp/Mb)	5027.1	3704.6	5560.0	4763.9	2401.8	2532.3	2500.3	2478
Tetra-nucleotide density (bp/Mb)	1748.8	1552.1	1586.9	1629.2	860.6	892.7	784.9	846
Penta-nucleotide density (bp/Mb)	654.8	584.7	439.2	559.6	257.1	364.5	351.6	324.4

Most common and longest SSR motifs

All dimeric repeat combinations excluding homomeric dimers were grouped into four classes $(AC)_n$, $(AG)_n$, $(AT)_n$ and $(CG)_n$. In all six-crop species considered, the largest number of occurrences of di-nucleotide repeats was from the $(AG)_n$ class of dimers (that includes CT, GA and TC repeats) (Tab. 2). In conformance with other reports the CG class of dimeric repeats was extremely rare. The triplet repeats in ESTs were grouped into 10 subclasses (Fig. 1) each representing six overlapping complementary patterns [21]. As expected, the CCG trimer class (consisting of CCG, GGC, CGC, GCC, GCG, CGG) motif was found to be most abundant in cereal ESTs. While in case of legume species, as observed in Arabidopsis earlier, the most frequent trimer class was TTC, with densities ranging from 580 bp/Mb to 770 bp/Mb in the three legume species.

Table 2: Most common and longest motifs in Cereal and Legume ESTs.

Repeats	Sorghum	Maize	Rice	Soybean	Medicago	Lotus
Most common di-nucleotide repeat	AG CT GA TC (1187)	AG CT GA TC (2039)	AG CT GA TC (2853)	AG CT GA TC (2192)	AG CT GA TC (982)	AG CT GA TC (596)
Most common tri-nucleotide repeat	CCG GGC CGC GCC GCG CGG (15810)	CCG GGC CGC GCC GCG CGG (15792)	CCG GGC CGC GCC GCG CGG (22862)	TTC AAG TCT CTT AGA GAA (5694)	TTC AAG TCT CTT AGA GAA (6318)	TTC AAG TCT CTT AGA GAA (2104)
Most common tetra-nucleotide repeat	AGCT GCTA CTAG TAGC TCGA CGAT GATC ATCG (2075)	AGCT GCTA CTAG TAGC TCGA CGAT GATC ATCG (2700)	AGCT GCTA CTAG TAGC TCGA CGAT GATC ATCG (2887)	AAAC AACA ACAA CAAA TTTG TTGT TGTT GTTT (1552)	AAAT AATA ATAA TAAA TTTA TTAT TATT ATTT (1287)	AAAT AATA ATAA TAAA TTTA TTAT TATT ATTT (378)
Most common penta-nucleotide repeat	ACGTG CGTGA GTGAC TGACG GACGT TGCAC GCACT CACTG ACTGC CTGCA (347)	AACAC ACACA CACAA ACAAC CAACA TTGTG TGTGT GTGTT TGTTG GTTGT (1005)	AGAGG GAGGA AGGAG GGAGA GAGAG TCTCC CTCCT TCCTC CCTCT CTCTC (255)	AAAAT AAATA AATAA ATAAA TAAAA TTTTA TTTAT TTATT TATTT ATTTT (356)	AAAAT AAATA AATAA ATAAA TAAAA TTTTA TTTAT TTATT TATTT ATTTT (277)	AAATT AATTA ATTAA TTAAA TAAAT TTTAA TTAAT TAATT AATTT ATTTA (202)
Longest di-nucleotide repeat	TC (69)	AT (266)	GA (75)	GA (106)	GA (84)	GA (51)
Longest tri-nucleotide repeat	GAA (30)	TAA (24)	TTC (29)	TAT (23)	TTC (17)	GAA (23)
Longest tetra-nucleotide repeat	CATA (16)	TATA (23)	TATG (18)	AGAG (24)	CTCT (22)	TCTC (11)
Longest penta-nucleotide repeat	ATGTA (10)	GGAGA (7)	GTGCT (7)	ACAAT (10)	CAACT (7)	CCACA (6)

Amino acid repeats for glycine, alanine, arginine and proline (GARP) were most frequent in the cereal ESTs (Tab. 3). In the legumes, the most common codon repeats were those of leucine and serine in medicago, serine in lotus and leucine in soybean. The longest repeats identified are also indicated in Tab. 2. Expansions of codon repeat coding for leucine (CTT/CTC/CTA/CTG/TTA/TTG) is tolerated well in soybean and lotus (>18 times) while serine (TCT/TCC/TCA/TCG/AGT/AGC) is tolerated up to 14 times in the three legume species studied. In the cereal ESTs, the commonly occurring GC-rich arginine/alanine codon triplets were found to occur with repeats ranging from 8-23.

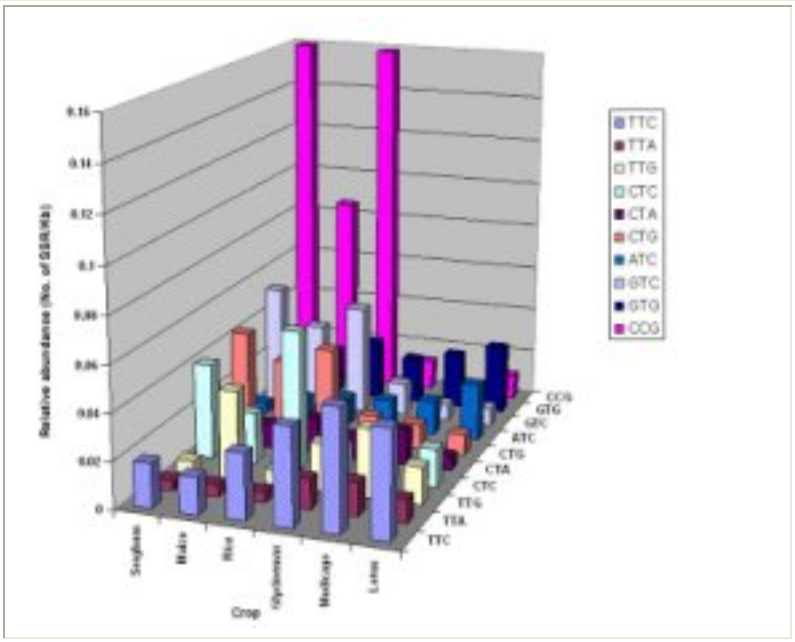

Figure 1: Density of tri-nucleotide repeats classes per million base pairs in the cereal and legume ESTs.

Table 3: Frequency and distribution of codon repeats in the cereals and legumes.

Amino Acid	Codon	Sorghum frequency (%)	Maize frequency (%)	Rice frequency (%)	Soybean frequency (%)	Medicago frequency (%)	Lotus frequency (%)
Alanine	GCT, GCC, GCA, GCG	18.30	17.68	13.95	6.40	2.85	4.10
Arginine	CGT, CGC, CGA, CGG, AGA, AGG	18.94	14.47	21.51	9.32	4.74	6.03
Asparagine	AAT, AAC	0.27	0.78	0.49	3.85	4.85	3.70
Aspartic acid	GAT, GAC	2.99	4.06	2.82	2.84	2.71	1.29
Cysteine	TGT, TGC	1.76	2.46	1.83	2.76	3.81	2.55
Glutamic acid	GAA, GAG	3.23	2.37	4.19	6.15	6.22	4.84
Glutamine	CAA, CAG	3.60	9.33	2.29	8.02	9.10	3.94
Histidine	CAT, CAC	1.87	2.48	2.67	4.86	5.24	7.56
Isoleucine	ATT, ATC, ATA	0.57	1.02	0.86	4.48	5.60	4.43
Leucine	CTT, CTC, CTA, CTG, TTA, TTG	5.35	7.50	7.78	10.41	9.40	13.45
Serine	TCT, TCC, TCA, TCG, AGT, AGC	7.05	8.40	8.53	9.06	9.26	16.73
Threonine	ACT, ACC, ACA, ACG	2.35	4.19	2.22	3.95	4.39	4.94
Tyrosine	TAT, TAC	0.55	0.54	0.41	1.46	1.68	0.41
Valine	GTT, GTC, GTA, GTG	2.76	5.56	2.42	2.54	2.09	1.48
Glycine	GGA, GGC, GGG, GGT	14.27	7.37	10.32	3.01	1.79	3.00
Proline	CCA, CCC, CCG, CCT	11.63	7.54	12.56	8.58	6.23	10.49
Lysine	AAA, AAG	1.67	1.17	1.81	3.19	2.94	1.98

The SSR loci were categorized into two groups based on the length of their SSR tracts [20]: class I SSRs ≥ 20 nucleotides in length and class II containing perfect SSRs >12 but <20 nucleotides in length (Fig. 2). Of the total number of SSRs identified in the cereal group, about 7% (7.9% in sorghum, 6.7% in maize and 6.9% in rice) were identified as class I and class II consisted of 93% of the SSR-containing ESTs. In the legume group about 9.8% of SSRs (9.5% in lotus, 10.4% in soybean, and 9.3% in medicago) were defined as class I and 90.2% were class II microsatellites (Fig. 2). In general, the class I repeats were largely composed of di- and penta-nucleotide SSRs, while class II repeats were higher in tri- and tetra-nucleotide repeats in both cereals and legumes.

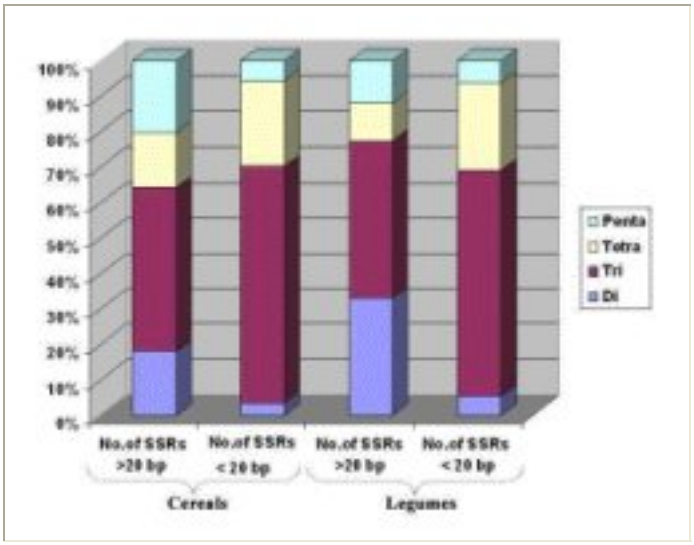

Figure 2: Distribution of class I and class II repeats in cereals and the legumes.

Table 4: Results after clustering of the non-redundant SSR containing dataset, primer design and assignment of putative function.

Crop	Total number of SSR-ESTs (redundant)	Non redundant SSR-ESTs			% Reduction in redundancy	Primer pairs available	Putative function available*
		Contigs	Singletons	Total			
Sorghum	39,106	4799	5245	10,044	74%	9594	3371
Maize	63,889	6728	5722	12,450	80.6%	12031	4053
Rice	58,343	6760	9862	16,622	71.5%	15,619	4541
Soybean	32,508	4515	5200	9715	70%	8912	2891
Medicago	27,308	3485	3412	6897	74.8%	6355	2295
Lotus	10,867	1581	2100	3681	66%	3475	946

* Numbers exclude annotations to hypothetical proteins/proteins of unknown function.

Functional EST-SSR markers

The SSR-ESTs were masked with RepeatMasker (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker) prior to clustering with CAP3 (Contig Assembly Program) [22]. Overlapping EST sequences were obtained in a cluster and each cluster had a consensus sequence. The CAP3 algorithm computes overlaps between sequences and then joins reads in decreasing order of overlap scores to form contigs. This clustering of SSR-ESTs eliminates the redundancy in the dataset. As a result, a collection of non-redundant EST-SSRs was obtained for each species. For instance in case of sorghum, 39,106 SSR-ESTs (redundant set) was reduced to 10,044 non-redundant or

unigene EST-SSRs (Tab. 4). At an overlap of 90%, the non-redundant EST datasets from each crop species could be reduced to a few contigs and singletons with an average of 75% elimination in the cereals and 70% elimination in the legumes.

To convert *in silico* identified EST-SSRs to molecular markers, primer pairs were designed for non-redundant EST-SSRs. Primer pairs were designed for 95% of cereals and 92% of legumes non-redundant EST-SSRs (Tab. 4). Approximately 32-33% of the EST-SSRs with primers had a putative function assigned other than hypothetical proteins/proteins with unknown function.

Further, putative COS EST-SSR markers were identified for cereal and legume crops separately as well as across all the species. Cross-species clusters were identified based on sequence homology using BLASTn reciprocal best hits. (See "Methods" for details). Thus there were 18,464 COS consisting of sequences from two species, 3466 COS containing sequences from 3 species, 406 COS clusters with sequences from 4 species, 36 COS markers consisting of sequences from 5 species, and only one COS marker across all the six crop species considered in this study. Of these, 4063 COS markers consisted of only legume sequences while 10,989 conserved ortholog sets were made up of sequences from cereals. A total of 22,373 COS sets consisted of both cereal and legume sequences. Primer pairs are available for the entire COS marker set, and the markers and associated information on the clustering can be retrieved through the query pages either by specifying the cluster size or the crop species.

This large-scale analysis has helped identify many transcripts with plausible putative annotations. These putative annotations could be grouped under the following categories: metabolic enzymes, ribosomal proteins, transcription factors, factors involved in translation, cell cycle regulators, proteins with hypothetical/unknown function, general structural proteins, chaperones, stress transcripts, protein degradation, ligand binding function, transcripts involved in signal transduction etc. About 60% of cereal EST-SSRs and 52% of legume EST-SSRs have a putative annotation (see Methods for criteria used). Fig. 3 indicates the number of EST-SSRs under each functional category. ESTs with long tri-nucleotide repeat coding for alanine/arginine amino acids in the cereal group had putative annotations to structural proteins, transcription factors and proteins involved in disease and defense. In the legume dataset, ESTs containing tri-nucleotide stretches coding for serine/leucine had annotations to transcription factors and ribosomal proteins.

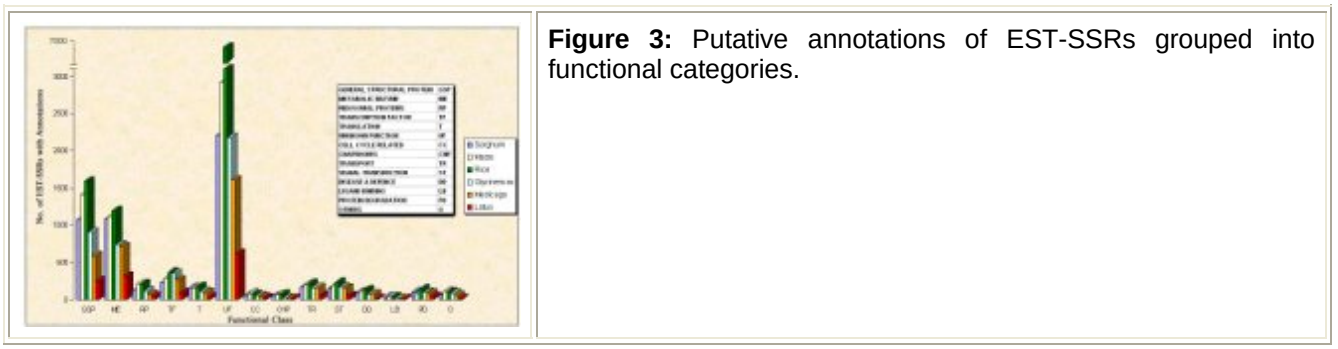

Figure 3: Putative annotations of EST-SSRs grouped into functional categories.

Validation of a subset of EST-SSRs

In order to demonstrate the utility of *in silico* identified EST-SSRs in sorghum, a set of 600 EST-SSRs with their primer pairs (all unpublished) were selected that had rice homologs spanning the entire rice genome. Approximately 13% of the SSRs were class I repeats. These primer pairs were tested for their ability to detect polymorphism between *Striga*-susceptible stay-green drought tolerant mapping population parent 'E 36-1' and its *Striga*-resistant, non-stay-green counterpart 'N13'. Out of 600 primer pairs tested, 457 (76.1%) gave an

amplification product. Among these, 386 primer pairs produced specific amplification products. The remaining 71 markers produced amplicons but did not show a clear banding pattern and were difficult to score (Fig. 4a).

Of the 386 EST-SSR markers tested, 133 markers (34.45%) detected polymorphism between parental lines N13 and E 36-1. About 26% of these markers were class I di-nucleotide repeats and 14% were class I tri-nucleotide repeats. Primer pairs detecting polymorphism were genotyped for mapping on a subset of 94 RILs from the (N13 × E 36-1)-based mapping population. The screening and mapping of these polymorphic markers is in progress. So far 71 markers have been screened on this population. The data produced with 71 EST-SSR markers were merged with previously generated RFLP, AFLP, SSR and RAPD marker data for these RILs and linkage analysis performed with MAPMAKER to position the EST-SSR markers on the existing skeleton map of this sorghum mapping population. Out of 71, forty-two EST-SSR markers mapped onto the ten sorghum linkage groups. Fig. 4b shows the location of newly mapped EST-SSR markers that are expected to show linkage. These markers are expected to show linkage with the indicated stay-green QTLs based on map position, which is expected to be the same in the cross in which these EST-SSRs were mapped and several other crosses-involving stay-green trait donor B35-- in which the two stay-green QTLs have been mapped; we have subsequently demonstrated that these expectations hold true by tracking segregation of these EST-SSR markers in backcross progenies segregating for these B35 stay-green QTLs with stay-green QTLs (*stg*B and *stg*3) from donor parent B35 on sorghum linkage group SBI-02.

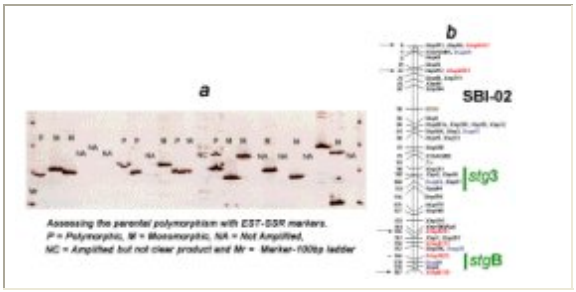

Figure 4: EST-SSRs as markers. (a) Observed polymorphism among four inbred lines (BTx623, IS 18551, N13, and E 36-1). (b) Map of sorghum linkage group SBI-02 for RILs of the cross of *Striga*-resistant, drought-sensitive line 'N13' with its *Striga*-susceptible, drought tolerant counterpart 'E 36-1' (EST-SSRs indicated in red).

Discussion

Microsatellite markers represent an important class of molecular markers for studying genome structure, evolution and applied aspects. Experimental studies on microsatellite frequency or distribution in a genome are generally feasible only for a few SSR types. In the past few years, due to the availability of large number of ESTs and development of several SSR-search tools, it has become possible to carry out *in silico* frequency and distribution studies of SSRs [12, 13, 14]. In the present study, the comparative assessment of frequency and distribution of SSRs between monocot and dicot species reveal a higher proportion of ESTs containing SSRs (19% in cereals and 11% in legumes) as compared to earlier reports [12, 13, 14]. However these figures entirely depend on the SSR search criteria, dataset and searching algorithm [9]. The relative abundance of SSRs in ESTs was found to be very similar across the three cereals and even more so between the three taxonomically related legumes. These estimates indicate the taxonomic proximity of the three investigated cereals [28] and legumes [29]. It has been reported that the relative abundance of different repeats depends on the species examined and is non random [30]. In previous comparative studies on EST-SSRs in monocot and dicot species, it was shown that SSRs are more abundant in genomes of monocots [8, 31], as also reported in this paper.

As expected, tri-nucleotide repeats were more common than others across cereals and legumes, followed by tetra-nucleotide and penta-nucleotide repeats except for soybean and rice where there were nearly equal occurrences of di- and penta-nucleotide repeats. Microsatellite sequences with di-nucleotide repeat are usually known to reside outside the coding regions of genes, which probably explains their reduced numbers relative to tri-/tetra-nucleotide repeats [20]. Multiples of tri-nucleotide repeats are tolerated better in ESTs because they do

not disturb the reading frame. Amino acid repeats are relatively common in eukaryotes. The most common are those of uncharged polar amino acids (such as glutamine, asparagine, serine, proline and threonine), acidic amino acids (glutamic acid, aspartic acid) or small amino acids (such as glycine and alanine) [32]. There is some evidence that homopeptide stretches mediate or modulate protein-protein interactions, modulate disease risks and effect changes in the protein products of genes leading to disease [2, 33, 34]. Studies in the yeast indicate strong reading frame preferences and differences in repeat motif frequencies in coding and non-coding regions [35]. These findings indicate the direct influence of selection in regulating the emergence of amino acid repeats by slippage. Densities of tri-nucleotide repeats in the coding regions could be partially limited by selection at the protein level. Besides, species-specific cellular factors may interact with trimeric repeats, and are likely to play an important role in the genesis of repeats [30].

In the present study, rice had the highest abundance of tri-nucleotide repeats amongst all the EST datasets, largely CCG/GGC/CGC/GCC/GCG/CGG repeats from the GC-rich codons of amino acids glycine, alanine, arginine, and proline. A comprehensive analysis involving 22 completely sequenced genomes shows that nucleotide bias causes a genome wide bias in the amino acid composition of proteins; GC-rich genomes encode proteins rich in the amino acids GARP [36]. Tri-nucleotide repeats formed the largest portion of class I and class II repeats. In a previously reported study across three sequence data sets from rice, BAC (bacterial artificial chromosome)-end sequences, ESTs and completely sequenced BAC and PAC (P1 derived artificial chromosomes) sequences, class I and class II microsatellites were found to be most frequent in the gene rich regions represented by ESTs, BAC and PAC sequences, with a higher frequency of class II than class I repeats [20]. In the present study, the lengths of tri-nucleotide and tetra-nucleotide repeat in cereals ranged from 24-30 and up to 23 repeats in legumes, which could probably have an effect on protein function. Many of these ESTs had putative annotations to transcription factors and proteins involved in disease and defense.

As compared to legume species, there also appears to be an excess of diversity of EST-tetra-nucleotide and penta-nucleotide repeats in the cereals, with almost all classes of tetra- and penta-nucleotide repeats being represented. Repeat-mediated variation has been proposed as a molecular basis for the rapid adaptation of both prokaryotes and eukaryotes to environmental changes. Similar kinds of repeat-mediated variation may be responsible for adaptation of monocot and dicot species to different environments, such as proline accumulation in response to drought [37]. In the pathogen *Neisseria* it was hypothesized that an excess diversity of coding tandem repeats contributed to antigenic variation within the pathogen [38].

A very high degree of redundancy has been observed in the cereal as well as legume EST datasets used in this study, redundancy being an inherent feature with EST datasets generated by random or shotgun sequencing within cDNA libraries [39, 40]. The advantage of having clustered datasets lies in the fact that contigs are longer in length, which facilitates the design of primers for those EST-SSRs where the SSR is near the end of the EST sequence. The database consisting of the EST-SSRs, primer pairs with putative annotations available for 32% of EST-SSR markers in cereals and 33% EST-SSR markers in legumes became a starting point for the development of genic molecular markers in the examined species. Only 22% of the EST-SSRs tested in sorghum, detected polymorphism across two pairs of parental genotypes of mapping populations. Compared to genomic SSRs, the EST-SSR markers detected a lower level of polymorphism in these genotypes (unpublished results) because of their origin from more conserved portions of the genome. Approximately 87% of the 600 sorghum EST-SSRs tested were class II repeats. More than 65% of the amplified markers contained tri-nucleotide repeats of the CCG (proline encoding) class. Polymorphic di-nucleotide repeats were largely GT repeats and comprised approximately 14% of all polymorphic markers. These markers mapped across all ten sorghum linkage groups. Based on their mapped positions on the sorghum genome, two of these EST-SSR markers were selected for large-scale diversity analysis of a core collection consisting of 3000 accessions (unpublished, still in progress). Mapping of these EST-SSRs allowed identification of rice genomic regions syntenic to sorghum QTLs for *Striga* resistance and the stay-green component of terminal drought tolerance on linkage group SBI-02 of sorghum. This has facilitated the development of additional sorghum EST-SSR markers

specifically targeting these QTL regions for future use in marker-assisted selection. This establishes the practical value of this database as a resource for molecular markers.

While there are more reports on SSR markers that can be transferred across species within a genus [19, 20, and several others], reports on transferability of genomic SSR markers across genera are few [41, 42]. EST-SSRs being derived from the transcribed regions of the DNA are generally more conserved, and transferability has been shown to be limited to species within the genus. Conserved orthologous sets or COS markers have been considered an attractive resource for use across genera [43] and also for genome structure and evolution studies. Since a large number of SSR-ESTs were identified for cereal and legume species in this study, different sets of COS EST-SSR markers for cereals and for legume species and orthologous sets comprising both were developed based on the reciprocal best-hit method. Primer pairs are available for all the COS EST-SSR markers. Developed COS EST-SSR markers are a potentially useful resource for related but less studied cereals (e. g. rye, pearl millet) and legumes (e.g. chickpea, pigeon pea, cowpea). Those COS EST-SSR that do not show SSR poymorphism in the genetic material of related species, would still have potential to be developed and used as SNP (single nucleotide polymorphism) markers. Use of such COS markers would further enhance our knowledge on syntenic relationships among those legume species where there is very little information available on this aspect as compared to cereal species.

Our large-scale analysis has identified many transcripts that contain SSR motifs and an analysis of their similarity to known genes indicates that they have a range of plausible functions. A very large number of SSR-ESTs (49% of cereals and 42% of legumes) were annotated hypothetical proteins or proteins with unknown function by virtue of these sequences finding homologs that have been annotated as hypothetical/unknown proteins. Given the opinion raised by Bennetzen *et al.*, 2004, that a large number of ESTs annotated to unknown function may in fact represent proteins encoded by transposable elements [44], the ESTs annotated to unknown function in this dataset of ESTs were searched against the repeats database [45]. Only 0.8% of maize ESTs, 1.52% of rice ESTs and 0.4% of sorghum ESTs had hits to sequences from the repeats database (with an E-value cut off <1e-5); indicating that these ESTs classified as hypothetical proteins had very small representation of transposable elements. Such additional information increases the potential usefulness of these EST-SSR markers in associating phenotypic diversity with genotypic diversity.

Conclusions

In this paper we have presented the results of an analysis of public EST data from three legumes and three cereals. The ESTs have been analysed for SSR content and distribution. The study identified more repeats in cereals than legumes. There was considerable conservation in the distribution of microsatellites in EST sequences from legumes. Tri-nucleotide repeats were the largest class of repeats in both cereals and legumes with alanine, arginine, proline repeats more common in cereals and leucine, serine coding repeats more common in the legumes. The SSR-containing ESTs were clustered resulting in non-redundant EST-SSRs, which were further used to construct conserved orthologous sets. The sequences along with primer pairs, putative annotations and BLAST results are available in a database that can be accessed at http://intranet.icrisat.org/gt1/ssr/ssrdatabase.html. A subset of sorghum EST-SSRs available in the database was used to test for polymorphism between a *Striga*-susceptible stay-green drought tolerant mapping population parents 'E 36-1' and its *Striga*-resistant, drought-sensitive counterpart 'N13'. A small proportion of these EST-SSR markers were found to be polymorphic between the N13 and E 36-1 parental lines, which after genotyping on RILs from the mapping population, were mapped across all ten sorghum linkage groups; establishing the value of this database as a resource for development of molecular markers for practical applications in cereal and legume genetics and breeding.

Acknowledgements

The authors are grateful for partial support from the Generation Challenge Program to JB, PVNSP and Council of Scientific and Industrial Research to RP.

References

1. Tautz, D., Trick, M. and Dover, G. A. (1986). Cryptic simplicity in DNA is a major source of genetic variation. Nature **322**, 652-656.

2. Cummings, C. J. and Zoghbi, H. Y. (2000). Trinucleotide repeats: Mechanisms and pathophysiology. Ann. Rev. Genomics Hum. Genet. **1**, 281-328.

3. Li, Y.-C., Korol, A. B., Fahima, T. and Nevo, E. (2004). Microsatellites within genes: structure, function and evolution. Mol. Biol. Evol. **21**, 991-1007.

4. Trifonov, E. N. (2003). Tuning function of tandemly repeating sequences: a molecular device for fast adaptation. *In:* Evolutionary theory and processes: modern horizons, papers in honor of Eviatar Nevo, Wasser, S. P. (ed.), Kluwer Academic Publishers, Amsterdam, pp. 1-24.

5. Rocha, E. P. C., Matic, I. and Taddei, F. (2002). Over-representation of repeats in stress response genes: a strategy to increase versatility under stressful conditions? Nucleic Acids Res. **30**, 1886-1894.

6. Ayres, 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. Theor. Appl. Genet. **94**, 773-781.

7. Hancock, J. M. (1995). The contribution of slippage-like processes to genome evolution. J. Mol. Evol. **41**, 1038-1047.

8. Morgante, M., Hanafey, M. and Powell, W. (2002). Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nature Genet. **30**, 194-200.

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

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

11. Andersen, J. R. and Lubberstedt, T. (2003). Functional markers in plants. Trends Plant Sci. **8**, 554-560.

12. 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 Mol. Biol. **48**, 501-510.

13. Kumpatla, S. P. and Mukhopadhyay, S. (2005). Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. Genome **48**, 985-998.

14. La Rota, M., Kantety, R. V., Yu, J.-K. and Sorrells, M. E. (2005). Non random distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat and barley. BMC Genomics **6**, 23.

15. Cordeiro, G. M., Casu, R., McIntyre, C. L., Manners, J. M. and Henry, R. J. (2001). Microsatellite markers from sugarcane (Saccharum spp.) ESTs cross transferable to erianthus and sorghum. Plant Sci. **160**, 1115-1123.

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

17. Eujayl, I., Sledge, M., Wang, L., May, G. D., Chekhovskiy, K., Zwonitzer, J. C. and Mian, M. A. R. (2004). *Medicago truncatula* EST-SSRs reveal cross-species genetic markers for Medicago spp. Theor. Appl. Genet. **108**, 414-422.

18. Gaitán-Solís, E., Duque, M. C., Edwards, K. J. and Tohme, J. (2002). Microsatellite repeats in common bean (*Phaseolus vulgaris*): isolation, characterization and cross-species amplification in *Phaseolus* spp. Crop Sci. **42**, 2128-2136.

19. Fulton, T. M., van der Hoeven, R., Eannetta, N. T. and Tanksley, S. D. (2002). Identification, analysis and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell **14**, 1457-1467.

20. 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 Res. **11**, 1441-1452.

21. Jurka, J. and Pethiyagoda, C. (1995). Simple repetitive DNA sequences from primates: compilation and analysis. J. Mol. Evol. **40**, 120-126.

22. Huang, X. and Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Res. **9**, 868-877.

23. Haussmann, B. I. G., Hess, D. E., Omanya, G. O., Folkertsma, R. T., Reddy, B. V. S., Kayentao, M., Welz, H. G. and Geiger, H. H. (2004). Genomic regions influencing resistance to the parasitic weed Striga hermonthica in two recombinant inbred populations of sorghum. Theor. Appl. Genet. **109**, 1005-1016.

24. Haussmann, B. I. G., Mahalakshmi, V., Reddy, B. V. S., Seetharama, N., Hash, C. T. and Geiger, H. H. (2002). QTL mapping of stay-green in two sorghum recombinant inbred populations. Theor. Appl. Genet. **106**, 133-142.

25. Tegelstrom, H. (1992) Detection of mitochondrial DNA fragments. *In:* Molecular genetic analysis of populations: A practical approach, Hoelzel, A. R. (ed.). IRL Press, Oxford, pp. 89-114.

26. Haussmann, B. I. G., Hess, D. E., Seetharama, N., Welz, H. G. and Geiger, H. H. (2002). Construction of a combined sorghum linkage map from two recombinant inbred populations using AFLP, SSR, RFLP, and RAPD markers, and comparison with other sorghum maps. Theor. Appl. Genet. **105**, 629-637.

27. Lincoln, S., Daly, M. and Lander, E. (1992). Constructing genetic maps with MAPMAKER/EXP 3.0. Whitehead Institute Technical Report. 3rd edn. Whitehead Institute, Cambridge, MA.

28. Devos, K. M. and Gale, M. D. (2000). Genome relationships: The grass model in current research. Plant Cell **12**, 637-646.

29. Wojciechowski, M. F. (2003). Reconstructing he phylogeny of legumes (Legumniosae): an early 21st century perspective. *In:* Advances in Legume Systematics, Part 10, Higher Level Systematics, Klitgaard, B. B., Bruneau, A. (eds.). Kew: Royal Botanic Gardens, 5-35.

30. Tóth, G., Gáspári, Z. and Jurka, J. (2000). Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. **10**, 967-981.

31. Cardle, L., Ramsay, L., Milbourne, D., Macaulay, M., Marshall, D. and Waugh, R. (2000). Computational and experimental characteristics of physically clustered simple sequence repeats in plants. Genetics **156**, 847-854.

32. Green, H. and Wang, N. (1994). Codon reiterations and the evolution of proteins. Proc. Natl. Acad. Sci. USA **91**, 4298-4302.

33. Perutz, M. F., Johnson, T., Suzuki, M. and Finch, J. T. (1994). Glutamine repeats as polar zippers: Their possible role in inherited neurodegenerative diseases. Proc. Natl. Acad. Sci. USA **9**1, 5355-5358.

34. Kazemi-Esfarjani, P., Trifiro, M. A. and Pinsky, L. (1995). Evidence for a repressive function of the long polyglutamine tract in the human androgen receptor: possible pathogenetic relevance for the (CAG) n-expanded neuronopathies. Hum. Mol. Genet. **4**, 523-527.

35. Mar Albà, M., Santibáñez-Koref, M. F. and Hancock, J. M. (1999). Amino acid reiterations in yeast are over represented in particular classes of proteins and show evidence of a slippage-like mutational process. J. Mol. Evol. **49**, 789-797.

36. Singer, G. A. C. and Hickey, D. A. (2000). Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol. Biol. Evol. **17**, 1581-1588.

37. Yamada, M., Morishita, H., Urano, K., Shiozaki, N., Yamaguchi-Shinozaki K., Shinozaki K. and Yoshiba Y. (2005). Effects of free proline accumulation in petunias under drought stress. J. Exp. Bot. **56**, 1975-1981.

38. Jordan, P., Snyder, L. A. and Saunders, N. J. (2003). Diversity in coding tandem repeats in related *Neisseria* spp. BMC Microbiol. **3**, 23.

39. Weber, J. L. and Myers, E. W. (1997). Human whole-genome shotgun sequencing. Genome Res. **7**, 401-409.

40. Bouck, J., Miller, W., Gorrell, J. H., Muzny, D. and Gibbs, R. A. (1998). Analysis of the quality and utility of random shotgun sequencing at low redundancies. Genome Res. **8**, 1074-1084.

41. Decroocq, V., Favé, M. G., Hagen, L., Bordenave, L. and Decroocq, S. (2003). Development and transferability of apricot and grape EST microsatellite markers across taxa. Theor. Appl. Genet. **106**, 912-922.

42. Zhang, L. Y., Bernard, M., Leroy, P., Feuillet, C. and Sourdille, P. (2005). High transferability of bread wheat EST-derived SSRs to other cereals. Theor. Appl. Genet. **111**, 677-687.

43. Parida, S. K., Anand Raj Kumar, K., Dalal, V., Singh, N. K. and Mohapatra, T. (2006). Unigene derived microsatellite markers for the cereal genomes. Theor. Appl. Genet. **112**, 808-817.

44. Bennetzen, J. L., Coleman, C., Liu, R., Ma, J. and Ramakrishna, W. (2004). Consistent over-estimation of gene number in complex plant genomes. Curr. Opin. Plant Biol. **7**, 732-736.

45. Ouyang, S. and Buell, R. C. (2004). The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res. **32**, D360-D363.