RESEARCH ARTICLE

Characterization of leaf transcriptome, development and utilization of unigenes-derived microsatellite markers in sugarcane (*Saccharum* sp. hybrid)

Mohammad Suhail Khan¹ · Sanjeev Kumar¹ · Ram Kewal Singh^{1,2} · Jyotsnendra Singh¹ · Sanjoy Kumar Duttamajumder¹ · Raman Kapur¹

Received: 7 June 2017 / Revised: 14 May 2018 / Accepted: 22 May 2018 / Published online: 4 June 2018
© Prof. H.S. Srivastava Foundation for Science and Society 2018

Abstract Sugarcane (*Saccharum* species hybrid) is the major source of sugar (> 80% sugar) in the world and is cultivated in more than 115 countries. It has recently gained attention as a source of biofuel (ethanol). Due to genomic complexity, the development of new genomic resources is imperative in understanding the gene regulation and function, and to fine tune the genetic improvement of sugarcane. In this study, a cDNA library was constructed from mature leaves so as to develop ESTs resources which were further compared with nucleotide and protein databases to explore the functional identity of sugarcane genes. The non-redundant ESTs (unigenes) were categorized into 18 metabolic functions. The major categories were bioenergetics and photosynthesis (4%), cell metabolism (5%), development related protein (3%), membrane-related, mobile genetic elements (5%), signal transduction (2%), DNA (1%), RNA (1%) and protein (2%) metabolism, other metabolic processes (3%), transcription factors (1%), transport (4%) and proteins related to stress/defense (4%). From 540 unique ESTs, 212 simple sequence repeats were identified, of which 206 were from 463 singlets and six were mined from 77 contig sequences. A total of 540 unique EST sequences were used for SSR

search of which 97 (17.9%) contained specified SSR motifs, generating 212 unique SSRs. The genes characterized in this study and the EST-derived microsatellite markers identified from the cDNA library will enrich genomic resources for association- and linkage-mapping studies in sugarcane.

Keywords Sugarcane · Transcriptome · Expressed sequence tags · ESTs · Molecular markers · SSR

Introduction

Sugarcane (*Saccharum* species hybrid) is the major source of sugar in the world fulfilling > 80% of world sugar requirement. In the year 2014, globally, sugarcane crop occupied around 27.1 million hectares and produced nearly 1.89 billion tonnes of canes and 178.9 million tonnes of sugar (in 2013) (FAOSTAT 2016). Globally, Brazil ranks first in acreage and has ~ 21% share in world sugar production, followed by India (16%), Europe (9%), China (7%), Thailand (6%) and USA (5%) (FAOSTAT 2016). Additionally, Brazil also produced 86 billion liters of ethanol in the year 2011 from sugarcane. India, with 25.1 million metric tonnes sugar production, is a key player in world sugar economics (FAOSTAT 2016). In recent past, sugarcane research has come to the limelight when ethanol derived from cane molasses was exploited as a renewable biofuel source, turning sugarcane into a global commodity with great economic potential. In fact, much of the sugarcane in Brazil is now diverted for the production of the green fuel, ethanol (www.sugarcane.org).

The present-day cultivated sugarcane is a result of interspecific hybridization involving *Saccharum officinarum*, *S. spontaneum*, and other related genera. Due to

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s12298-018-0563-y>) contains supplementary material, which is available to authorized users.

✉ Mohammad Suhail Khan
drkhansuhail@gmail.com

¹ ICAR-Indian Institute of Sugarcane Research, Raibareli Road, P.O. Dilkusha, Lucknow, U.P. 226002, India

² Present Address: Division of Crop Science, Indian Council of Agricultural Research, Dr. Rajendra Prasad Road, Krishi Bhawan, New Delhi 110 001, India

such multispecies origin, sugarcane possesses one of the most complex plant genomes. Cultivated sugarcane is a complex heteropolyploid (octoploid) and due to the operation of female restitution, chromosome number in hybrids ranges from 108 to 120 (D'Hont et al. 1996) and commensurately have large DNA content (7440 Mbp). Such massive genome organization infers that each single-copy gene is represented at least by around eight alleles, each of which potentially corresponds to a distinct sequence haplotype. Interspecific origin, cytogenetic complexities, large genome size, complex polyploidy and heterozygosity in sugarcane make varietal development a long drawn process of 12–15 years but clonal propagation helps to maintain its fine identity (Stevenson 1965).

The complexity of sugarcane genome is a bottleneck in terms of efforts and investments that have been put forth in the development and application of molecular genetic tools for improvement of this crop. The Brazilian effort (SUCEST) has become the milestone in generating 237,954 high-quality ESTs from 26 cDNA libraries derived from different sugarcane tissues (Vettore et al. 2003). This contributed significantly in the understanding of complexity in sugarcane genome. With the advancement of molecular biology, much more relevant novel techniques have become available which can be gainfully utilized to unravel the complexities of sugarcane genome. Expressed sequence tag (EST) is one such tool to access the genes of economically important traits of sugarcane. Interestingly, ESTs have also been proven a cheaper resource for identification of novel microRNAs as a key regulator in post-transcriptional regulation in sugarcane (Khan et al. 2014).

In the case of sugarcane, significant advancement has been made in recent years for identification of molecular markers associated with traits of economic importance, e.g., sucrose content, red rot resistance, rust resistance, number of millable canes and high yield (Aitken et al. 2006; Costet et al. 2012; Da Silva and Bressiani 2005; Jordan et al. 2004; Singh et al. 2013, 2016). In the absence of complete genome sequence information of sugarcane, high-throughput transcriptome sequencing has been employed to understand the genes involved in complex biological processes. Significant progress has been made in the EST data enrichment by different countries especially Brazil and India (Gupta et al. 2010; Vettore et al. 2003). The Brazilian sugarcane EST project (SUCEST) generated 237,954 ESTs from 26 diverse cDNA libraries, which have been utilized to identify different gene functions (Vettore et al. 2003). The identified ESTs relate to photoreceptor genes, DNA repair genes, genes for cell cycle machinery, flowering-specific genes, molecular chaperones, oxidative stress-related genes, etc. (Andrietta et al. 2001; Borges and Ramos 2001; Costa et al. 2001; Dornelas and Rodrigue 2001; Santelli and Siviero 2001; Soares Netto 2001). In

India, our group previously developed ~ 35,000 ESTs from normal, water-stressed and red rot infected tissues of the sub-tropical sugarcane varieties and identified EST clusters that showed differential expression for biotic and abiotic stresses (Gupta et al. 2010). In present endeavor, ESTs were developed from leaf tissues of a subtropical sugarcane variety CoS 767, and the transcriptome was further explored to identify genes expressed in leaf tissues. The ESTs thus generated were mined to identify SSR repeat motifs to enrich the marker resources.

Materials and methods

Plant material

The leaf tissues used for the construction of cDNA library were collected from 90 days old plants of the sugarcane variety CoS 767. It is a well-adapted variety of subtropical India, popular in the states of Punjab, Haryana, U.P., and Bihar. It belongs to the medium maturity group having medium thick erect tall canes with good cane yield (Srivastava et al. 1989). The experimental sugarcane varieties were planted during Feb.–March at IISR research farm, Lucknow located at 26.56°N and 80.52°E and 111 m above mean sea level. The experimental site falls in subtropical zone of India having temperatures between 10 (min.) and 30 °C (max.) during Feb.–March. Recommended package of practices was followed so as to raise a good crop.

RNA purification and preparation of cDNA library

For RNA isolation, leaf samples from three random plants were collected and pooled. The leaf tissues were rinsed with ice-cold DEPC-treated water so as to prevent RNA degradation, then quickly stabilized in liquid nitrogen and transferred to a – 80 °C deep freezer until RNA isolation. The total RNA was isolated following guanidine isothiocyanate-phenol method with minor modifications (Chomczynski and Sacchi 1987). The poly (A)⁺ mRNA purification was then carried out from total RNA by affinity chromatography using oligo (dT)-cellulose and cDNA library was prepared using the ZAP Express cDNA Synthesis Kit (Stratagene) on the ZAP Express vector using manufacturer's protocol. In brief, the mRNA was converted into cDNA using 50-base oligonucleotide linker-primer and reverse transcribed using StrataScript reverse transcriptase and 5-methyl dCTP. The uneven termini of the ds-cDNA were polished using *Pfu* DNA polymerase, and *Eco*RI adapters were ligated to the blunt ends. The size-fractionated and finished cDNA was cloned into the pBK-CMV ZAP Express vector in a sense orientation with reference to the lac Z promoter. The λ-library was

packaged in a high-efficiency Gigapack III Gold packaging extract system (Kretz et al. 1994) and was plated on the *E. coli* XL1-Blue MRF'. Further, the primary leaf cDNA library was amplified to obtain a stable library in large quantity. The phagemids were excised using helper cell and cell mixtures were plated on Luria Agar containing X-gal and IPTG so as to select recombinant colonies for plasmid isolation and sequencing of ESTs.

Sequencing of the cDNA library and annotation of ESTs

The recombinant colonies were grown in Luria-Bertani broth in 96 deep-well plates. DNA sequencing of the cDNA clones was carried out using ABI Prism 3700 Genetic Analyzer. The data were recorded using Data Extractor software and, after the complete run, it was further analyzed with the help of the DNA Sequencing Analysis program. The cDNAs were sequenced from 5' end so as to generate EST sequences. Primary processing of EST sequences, quality trimming and base calling was done using the software PHRED (Ewing and Green 1998), and vector/adaptor sequences were screened using Cross Match (http://www.phrap.org/phrap_documentation.html). EST sequences of > 299 bp with PHRED score > 20 were submitted to the GenBank. The assembly software CAP3 was used to estimate the redundancy in the EST data and make sequence clusters (Huang and Madan 1999). To establish the gene identity, ESTs were used to perform homology search in NCBI non-redundant (NR) database (www.ncbi.nlm.nih.gov/BLAST/) using BLASTn and BLASTx tools (Altschul et al. 1990).

In silico identification of SSR motifs in ESTs

Simple sequence repeats (SSRs) were mined from EST sequences using the program SSR Primer (<http://hornbill.cspp.latrobe.edu.au/cgi-bin/pub/ssrprimer/indexssr.pl>) with default parameters. This SSR Primer software searched for all possible repeat patterns with $n \geq 7$ for di-, $n \geq 5$ for tri-, and $n \geq 3$ for tetra-/penta-nucleotide repeat motifs against the Cluster Cap 3 consensus sequences. Further, this program also provided a list of all the possible primer sequences to amplify SSRs. Later, out of all the designed primers, 50 primer pairs were synthesized (Life Technologies, USA).

Validation of SSR primers

For validation, all the 50 SSR primers were used to amplify DNA of six sugarcane varieties, viz., CoS 767, Co 1148, CoJ 64, CoS 96268, CoLk 94184 and BO 91. In addition, the SSR primer (SUMSSR_38) was also tested to reveal the amplification pattern from 24 different sugarcane

genotypes (Supplementary Table 1). For PCR amplification, total genomic DNA was extracted from lyophilized powdered young leaf tissues (500 mg) as per the method of Doyle and Doyle (1990). The total reaction volume of PCR was 10 μ l consisting of 10X PCR assay buffer (Fermentas, USA), 200 μ M each of dNTPs (Fermentas, USA), 12 ng (1.8 pico Mole) both forward and reverse primers (Life Technologies, USA), 0.5 units of *Taq* DNA polymerase (Fermentas, USA) and 25 ng of genomic DNA. The amplification reactions were performed in a thermal cycler (BioRad, USA) using the following reaction parameters: 94 °C for 5 min.; 33 cycles of 94 °C for 1 min., 55 °C for 1 min., 72 °C for 2 min., and finally, a primer extension cycle of 7 min. at 72 °C. Finally, the PCR amplified products were resolved on 10% polyacrylamide gel and visualized in a gel documentation system (G: Box; Syn-gene, UK). A flow chart representing series of steps for generation of ESTs, SSR identification and their validation has been given in Supplementary Figure 1.

Results

cDNA library preparation from sugarcane leaf tissue

With an aim to generate expressed sequence tags (ESTs), representative of the set of mRNAs expressed in mature leaf tissue of sugarcane variety CoS 767, a cDNA library was prepared and cloned into λ -phage. The size of the primary cDNA library was 2.01×10^5 pfu/ml. The primary cDNA library was amplified using XL1-Blue MRF host cells to produce large and stable high-titer stocks (1.62×10^9 pfu/ml). Mass excision was carried out on the ZAP Express cDNA library to convert the library into the phagemid. Approximately 60% of the recombinant clones had the insert of 1–4 kb, and many of the inserts possessed internal EcoRI and/or XhoI sites.

Sequencing of leaf cDNA library

A total of 1440 clones from the leaf cDNA library were randomly selected for sequencing and further generation of ESTs. Out of 1440 sequences, 664 sequences were of < 10 bp and discarded as failed sequences. The remaining 776 sequences (> 10 bp) were used for downstream processing. Of these, 417 (54%) sequences had PHRED score 20 and were more than 299 bp in size (Supplementary Figure 2); such sequences were checked manually and 397 good quality sequences were submitted to GenBank (DV548660–DV549056). All the EST sequences were assembled using the software CAP3 with default parameters in order to remove redundancy (Huang and Madan

1999). The characteristics of mature leaf library are furnished in Table 1. The assembled data was manually verified to remove any misassembled sequence. This resulted in a final assembly containing 368 total reads assembled as 77 contigs, and rest of 1072 sequences remained as singletons (Table 1). Analysis of these singletons revealed that there were 463 ESTs of good quality (Phred score > 20) and the requisite size (> 100 bp). Thus, it is clear that these sequences are unique in the present dataset, and due to lack of any overlapping partner, such sequences were considered as a separate group of the assembly. The sequence clustering resulted in the identification of 540 unique consensus sequences, thus making the level of redundancy at ~ 54%.

Homology search and identification of genes

The BLAST search using the NCBI non-redundant nucleotide as well as protein databases was performed for putative identification of ESTs and to corroborate their plant origin. The BLASTn analysis of 540 mature leaf unigenes revealed that 484 (~ 90%) were homologous to nucleotide sequences, while 56 (~ 10%) did not show any match indicating them to represent new or unreported genes. But BLASTx search from peptide sequence database showed 446 (~ 83%) unigenes having homology, whereas, remaining 94 (~ 17%) unigenes did not exhibit sequence similarity to peptide databases. Therefore, it was concluded that 10–17% unigenes were either new or non-

reported. When subjected to BLASTn, among the total 540 unigenes, only 180 showed homology to known plant sequences, whereas, a large number (304) of ESTs showed sequence similarity to previously identified sequences from non-plant species ranging from bacteria to human. Of the 180 identified clones homologous to plant nucleotide database, 57 were similar to maize, 49 to rice, 37 to sorghum, 15 to sugarcane, 6 to *Arabidopsis*, 3 to *Medicago*, and the remaining 13 unigenes showed homology to other plant species (Supplementary Figure 3, Supplementary Table 2). A few redundant clones were also detected during the sequence analyses; the database search indicated that the nucleotide sequences of such clones were not identical to each other, thus indicating further that these do not represent copies of the clone in cDNA library. A majority of them were homologous to different regions of the gene, however, small regions of sequence overlap between clones were also observed in some cases.

When subjected to BLASTx, 106 (~ 19%) of the 540 unigene (NR) ESTs had an amino acid sequence homology significant to plant genes identified previously (Supplementary Table 3), while 340 (~ 62%) showed homology to various non-plant species. Notably, three unigenes sequences showed homology to proteins reported from sugarcane. Further, the majority (~ 72%) of unigenes exhibited homology to monocots like rice, maize, and sugarcane.

Functional classification of the leaf unigene ESTs

All ESTs regardless of their E-value and redundancy were included for functional classification. In a further attempt, proteins were grouped according to functional characteristics/cellular roles, rather on the basis of their biochemical identities. As opposed to specific binding or catalytic function specific roles were allocated on the basis of known or presumed involvement of a particular gene/protein in cellular processes. All identified non-redundant ESTs (unigenes) could be categorized into 18 metabolic functions. The major categories are bioenergetics and photosynthesis (4%), cell metabolism (5%), development-related protein (3%), membrane-related, mobile genetic elements (5%), signal transduction (2%), DNA (1%), RNA (1%) and protein (2%) metabolism, other metabolic processes (3%), transcription factors (1%), transport (4%), stress/defense related (4%) (Fig. 1). The mobile genetic elements included polyproteins, transposon proteins, Ac-like transposases, putative GAG-POL precursor, putative copia-type pol polyprotein, RIRE2, putative arm repeat-containing protein, and retrotransposon protein. Further, a large number of unigene sequences were found to encode proteins with no definite identity/function, which included homology to hypothetical proteins (32%), unnamed

Table 1 Summary of mature leaf EST library obtained from sugarcane genotype CoS 767

Number of cDNA clones sequenced (5' end)	1440
Number of sequences that passed quality check	776
Success index (%)	54
Average size of cDNA insert (bp)	600
Average size of a good sequence (bp)	500
Number of bases with phred quality > 20/read	399
Number of singletons	463
Number of contigs	77
Number of reads assembled in contigs	368
Number of unique putative transcripts (unigenes)	540
Number of assembled ESTs (unigenes) with	
BLASTn match	484
No BLASTn match	56
BLASTn match with plant sequences	180
BLASTx match	446
No BLASTx match	94
Observed redundancy (%) ^a	54

^aObserved redundancy: (EST after quality check—Unigene)/Unigene

proteins (12%), unknown (8%), as well as unclassified proteins (3%).

Development of unigenes-derived microsatellite markers

A total of 540 EST clusters that included 463 singletons and 77 contigs covering 484,631 bp were obtained from the mature leaf cDNA library. A total of 212 SSRs from 540 unique ESTs, among them 206 SSRs were identified in 463 singlets and 6 SSR were mined from 77 contig sequences (Supplementary Table 4). In this way, 97 (17.9%) unique EST sequences contained SSRs of which 40 (41.2%) ESTs contained more than one SSR, while 28 (28.8%) were found to be compound SSRs having repeats more than one repeat type.

The analysis of SSR motifs revealed that the mean length of each unit varied between 16 and 25 bp without a uniform distribution. The SSR motifs in unigenes contained 45 di-, 51 tri-, 84 tetra- and 32 pentanucleotide repeats (Supplementary Figure 4), and the mean SSR length was 18.7 bp for all repeat classes except dinucleotides for which it was a bit higher (25.6 bp), with the maximum of 87 bp for (AG)n/(CT)n. A total of 46 SSR motifs (Supplementary Table 5), represented 4 types of di-repeats, 7 types of tri-, 14 types of tetra-, and 21 types of pentanucleotides repeat motifs (Supplementary Figure 4). As far as frequency of repeats is concerned, (AG)n/(CT)n was the most frequent dinucleotide motif that accounted for 57.8%, followed by (AC)n/(GT)n (37.8%) (Supplementary Figure 5). The most common tri-nucleotide motif was

(ATG)n/(CAT)n with 37.3% frequency, followed by (AAG)n/(CTT)n with 17.6%, and, (ACG)n/(CGT)n and (AGT)n/(ACT)n with 11.8%. The motif (GCC)n/(GGC)n was the least (2%) abundant motif among tri-repeats (Supplementary Figure 6). In the case of tetra-nucleotides, the most abundant (47%) motif was (AGTG)n/(CACT)n, followed by (AGAT)n/(ATCT)n and (ATGT)n/(ACAT)n with 9.5% (Supplementary Figure 6). The motif (ATGAG)n/(CTCAT)n with 12.5% was the most abundant penta-repeat followed by (AGAGA)n/(TCTCT)n (9.4%) (Supplementary Figure 6). In totality, the most abundant (18.9%) repeat motif among the 212 motifs was (AGTG)n/(CACT)n, followed by (AG)n/(CT)n motif (12.3%).

Functional identification of the EST clusters containing SSRs

In order to identify the putative function(s) of ESTs from which SSRs were derived, the corresponding EST was compared to NR protein database in GenBank using the BLASTx (Altschul et al. 1990). The clusters containing microsatellites showed homology to a wide range of proteins including enzymes, RNA-/DNA-binding proteins, regulatory proteins, membrane/cell wall-associated proteins, and stress-/defence-associated proteins (Supplementary Table 6). However, a majority (44.3%) of the clusters were associated with hypothetical proteins and also proteins with unknown functions. Few discrepancies were also noticed for some clusters (10.3%) that showed significant similarity in relation to an unnamed protein derived from human/virus gene product.

Fig. 1 Functional classifications of sugarcane leaf unigenes according to biochemical and metabolic processes. Regardless of similarity score, search was included in the analysis for all cDNA clones for which an identity was assigned through non redundant protein database

Validation of EST-derived SSR primers

A total of 212 primers pairs were designed out of which 50 primers (Table 2) that flank di-, tri-, tetra- or pentameric regions of SSRs were exploited to detect polymorphism across a set of six sugarcane varieties, viz., CoS 767, Co 1148, CoJ 64, CoS 96268, CoLk 94184 and BO 91. Of the 50 primer pairs, 36 (72%) yielded amplification products of different sizes; five of these primers produced monomorphic amplification products, while, the remaining 31 produced polymorphic products in at least one of the six genotypes analyzed. The remaining 14 (28%) EST-SSR primers failed to give any amplification. On basis of good amplification pattern, the primer SUMSSR_38 was used to further validate its amplification in another 24 sugarcane genotypes (Supplementary Table 1) producing more than one amplicon including the expected 330 bp DNA amplification product (Fig. 2).

Discussion

Sugarcane is one the most important commercial crops that fulfills the demand of sugar and biofuel-ethanol. Even the best plant breeding techniques could not be exploited to their full potential for breeding sugarcane varieties mainly due to high polyploidy, heterozygosity, and chromosome mosaicism. Further, the interactions of introduced gene(s) in a complex genomic system pose a great challenge for transgenic technology. In previous studies, the sugarcane transcriptome yielding ESTs has allowed the identification of the gene(s) associated with biotic/abiotic stress response, disease resistance and sucrose accumulation as well. In plant species like *Arabidopsis thaliana* and *Oryza sativa*, the huge EST databases serve as a resource for unraveling genetic diversity and identifying gene transcripts and their regulation. In the present study, 20% of the leaf cDNA clones did find similarity to known plant-related gene sequences. A similar level of homology has been reported in leaf cDNA library from maize (Keith et al. 1993). In contrast, Carson and Botha (2000) reported a much higher (26%) homology of immature leaf cDNA library from sugarcane. In the case of rice, a higher (25%) frequency was obtained (Yamamoto and Sasaki 1997), while it was 32% for cDNA clones from seedlings, roots, leaves, and inflorescences of *Arabidopsis* (Newman et al. 1994). It has been reported that a significantly higher degree of identity is obtained if the experimental tissue was subjected to processes comprising of well-characterized proteins (Van De Loo et al. 1995). A much higher (63%) proportion of clones showed a significant homology to sequences associated with genes not annotated from plant system. There might be a possibility that such gene(s) are

not well-known in plants species. Results from the present study indicate that routine resubmission of clones without sequence similarity result in more identifications mainly due to new interim additions to the databases.

A random sequencing of clones may result in the identification of genes belonging to the superabundant/abundant classes. Thus, in order to identify rare genes, it is imperative to sequence complete transcriptome, or alternatively, a normalized library could also be prepared. However, both these approaches are expensive due to the involvement of manpower and economic resources for large-scale sequencing of total cDNA libraries. The present study revealed the abundance of genes related to cell metabolism, mobile genetic elements, and transport in the leaf tissues with a significantly higher proportion of genes having roles in bioenergetics and photosynthetic pathways, indicating the high metabolic rate of the leaf tissues. In addition, some genes were identified as being stress- or defense- induced; whereas, a number of unigenes showed homology with genes with no significant role in plants. In some species, it could be possible that during evolution, genes associated with specialized mechanisms/critical functions may have been “borrowed” to form new genes with different functions, or which simply share some common functional domain, e.g., protein moonlighting (Jeffery 2003; Piatigorsky 2007). This study indicated the presence of several types of clones more than once in sequenced cDNAs, indicating the occurrence of multiple copies of specific genes and their relative frequency.

A considerable proportion (5%) of leaf unigene (non-redundant) EST sequences showed homology to transposable elements (TEs) (Table 3); this included gag/pol protein (Contig43), En/Spm sub-class (Contig16; SUM01-067-F06-A-047.g), transposon proteins (SUM01-067-E11-A-083.g, SUM01-075-E11-A-083.g, SUM01-070-C08-A-054.g), Ac-like transposase (SUM01-061-F09-A-075.g), putative polyproteins (SUM01-074-G01-A-004.g; SUM01-062-A01-A-001.g; SUM01-068-H04-A-032.g), and retro-transposon proteins (SUM01-069-B04-A-029.g; SUM01-070-A07-A-049.g). In most of the eukaryotes, TEs are present in high copy number and represent the largest component of the genetic material averaging 45% in the human genome and 50–80% in genomes of grasses (Feschotte et al. 2002). The TEs produce sequence alterations in the gene(s) or may cause large genome rearrangements, both of which can result to altered gene expression/function (Bennetzen 2000), indicating that such elements might have an adaptive function in eukaryote evolution (Feschotte et al. 2002; McClintock 1984). The occurrence of a large number of TEs in the current study is in agreement with the reports from the Brazilian Sugarcane EST Sequencing Project (SUCEST; Rossi et al. 2001; Vettore et al. 2003; Vincentz et al. 2004). In sugarcane, *in silico*

Table 2 Nucleotide sequence, expected product size and amplification details of 50 EST-SSRs primers derived from the unigenes of leaf library

Primer	Nucleotide sequence (5'–3')	Motif	T _m (°C)	Expected product size ^a (bp)	Amplification product
SUMSSR_01	F-gtatgacgacatgatgcaag R-ctacagcactcaatcgatca	(GAGT) ₁₉	54.9	287	P
SUMSSR_02	F-atgtgtgcatgtgtgctagt R-tcagctgctcactcactcta	(GTGA) ₁₃	54.9	353	P
SUMSSR_03	F-gtgcacgtatgatgaatgc R-tctcgtctctctcactctc	(GATA) ₁₂	54.9	336	P
SUMSSR_04	F-tagtgccatagctgataga R-atctcgctcactcactgtct	(AGTG) ₂₇	54.9	225	P
SUMSSR_05	F-gtatgtagacacgcacgac R-accatgatacacacagcgta	(GTAT) ₁₃	55.0	259	P
SUMSSR_06	F-ccatattgttagtgtgtgc R-ctgtcacgctctacacactg	(TGCG) ₁₃	54.8	371	P
SUMSSR_07	F-cgagagagagaggtatgcac R-actatctcagctctcgcaac	(GA) ₁₉	55.0	304	P
SUMSSR_08	F-cgagagagcgatagagacag R-tacgcacacacaatctcact	(ACGC) ₁₆	55.3	391	P
SUMSSR_09	F-ttagtatctgttcagtcgc R-gctacctcaccatctacagc	(AGT) ₁₃	54.6	223	P
SUMSSR_10	F-ggaggaggagtgtagtag R-tctcgttttcaactttgt	(GAA) ₁₅	55.2	231	F
SUMSSR_11	F-ctgatgctactagctgctga R-cagtccaatcgtctgcaat	(GATGT) ₁₅	55.4	356	F
SUMSSR_12	F-tgatgatcgactacgatacg R-gtatcgactcatctgcacac	(ATGAT) ₁₇	54.3	383	P
SUMSSR_13	F-tcgtatcgtagatgatgctg R-atagtcacgtcgtatccgt	(GAT) ₆₃	54.7	215	P
SUMSSR_14	F-agatgatgctgatgatgttg R-tcgtactcagaatcgtcaatc	(GAC) ₃₁	54.9	343	F
SUMSSR_15	F-gatggatgatgtggtatgtg R-ctccatctcaggtttctg	(ATGT) ₂₉	54.6	395	P
SUMSSR_16	F-aggagagagcaacaggag R-ctcgtccactctctcgt	(GGAG) ₂₀	57.2	387	P
SUMSSR_17	F-ggagagagagaaggcagg R-atactcactctcctcatcg	(GGAG) ₃₈	54.6	360	P
SUMSSR_18	F-gaggaagagaggagagaga R-gcctcaccatactcactc	(AGGG) ₁₄	55.3	209	F
SUMSSR_19	F-ccagaaccacaagaactagc R-gatttctcggctctttaattc	(CCTCT) ₁₈	55.0	123	M
SUMSSR_20	F-cccacagtggatagaggata R-atgatgcacatctcgccttat	(CTTTC) ₁₅	55.0	310	P
SUMSSR_21	F-cgtgagtgtagagtgtg R-actctctcttgaatcggtc	(TGCG) ₁₄	55.3	265	P
SUMSSR_22	F-gaagtatggtgtagtgctg R-tctctcctcaatcagtcctc	(GTGA) ₂₉	55.4	316	P
SUMSSR_23	F-gatccagtcgatcagttcac R-gtaattaaccacccggtaca	(TTG) ₁₅	55.2	245	F
SUMSSR_24	F-tgaaagcttcttctcaacc R-acaacaacgtcttccaattc	(TAAA) ₄₄	54.9	269	P

Table 2 continued

Primer	Nucleotide sequence (5′–3′)	Motif	Tm (°C)	Expected product size ^a (bp)	Amplification product
SUMSSR_25	F-ggagagaagattgttggg R-cacaaacaacaccacacac	(GGGT) ₁₃	55.2	271	F
SUMSSR_26	F-gaaaggagggaaggtagta R-tcttcacactctacatccac	(GGA) ₁₅	55.0	328	F
SUMSSR_27	F-gcaaagagaagagcagaaga R-atcctcagtgccccatacag	(AGAAG) ₁₅	55.0	331	F
SUMSSR_28	F-gaggaggaggagatagagaga R-gtcgctcctccacgctac	(GAGG) ₁₅	57.3	372	M
SUMSSR_29	F-tcagttgagaatgatcacg3′A R-tgcttctacatcagcgtcta	(GAC) ₂₄	54.9	321	F
SUMSSR_30	F-gagtgctagagagtgcgt R-gcacagctcttcatcacag	(AGAC) ₁₄	54.7	356	F
SUMSSR_31	F-atgatgtcgtgatgaggat R-attcgtacgttcgtcgtagt	(TGA) ₁₅	55.0	293	P
SUMSSR_32	F-gcgtggagagagataatgc R-atcacgcttcatctcaatct	(TG) ₂₀	54.9	400	F
SUMSSR_33	F-gtgtgagtagtgtagaagtgtg R-acaatcactctcactcgc	(GTGA) ₂₅	52.8	336	P
SUMSSR_34	F-gagagagatagtcctgtc3′G R-gcactcagacgcataattgta	(AG) ₃₁	55.0	325	M
SUMSSR_35	F-gatgtgagtgatgggtgtg R-tctcactacattctatcggga	(GTAT) ₁₅	54.7	147	P
SUMSSR_36	F-cgtgactgactgtgtgagag R-acactcgcgcacatctatct	(AGTG) ₂₆	55.0	375	P
SUMSSR_37	F-tgatgatatgagctgctacg R-gcacaactatcatcacatcg	(GATGA) ₁₅	55.0	109	P
SUMSSR_38	F-gagtgaggagtgaggagaga R-cacagtgacacactcacaca	(AG) ₃₈	54.4	330	P
SUMSSR_39	F-tgagagagagagagagcg R-tcacactcagactcacacaag	(ATAG) ₁₆	54.9	356	F
SUMSSR_40	F-gaggtgatgtaggttgtgt R-tcatctcactcacatctctcc	(GTGTT) ₁₅	55.0	253	P
SUMSSR_41	F-gttggtgataggttagtgga R-ctcacacttttctcactctca	(GA) ₃₁	54.9	393	F
SUMSSR_42	F-gtgagcgagcgagttagt R-tgcagacgtacctactcaca	(ACGT) ₂₁	54.7	177	P
SUMSSR_43	F-atgcgtgtgtgatatgtgt R-tcacactcactctgtctctc	(AGTG) ₂₀	54.3	346	F
SUMSSR_44	F-tgagagtggtttccttgtc R-tcttcatctctcactcatca	(AG) ₂₂	55.2	184	P
SUMSSR_45	F-gggagagtgtagagtgaga R-acacaccaactctgtcacat	(GA) ₂₅	54.4	275	P
SUMSSR_46	F-agagagcgacagtgagagag R-ctcactatcacgcacatc	(TG) ₃₄	54.9	224	P
SUMSSR_47	F-gagagtggtgattcctcgt R-caagctcactatcacacaaca	(AG) ₄₀	55.0	374	M
SUMSSR_48	F-gtatgtatgtatcgaggag R-acatccactcacatactcacac	(GT) ₂₁	54.7	301	P

Table 2 continued

Primer	Nucleotide sequence (5′–3′)	Motif	T _m (°C)	Expected product size ^a (bp)	Amplification product
SUMSSR_49	F-gtgtgtgtgtgcttgtgagt R-ctctctctctctctcgctca	(AG) ₂₄	55.2	379	M
SUMSSR_50	F-ttgacgattctgagtacgac R-ctccatcgtcagtcagtc	(GAT) ₂₆	54.8	285	P

^aEstimated by the SSR identification program; *P* polymorphic, *M* monomorphic; *F* failed

Fig. 2 Validation of EST-SSR primer SUMSSR_38 in 24 sugarcane genotypes. Lanes: M: 50 bp DNA size marker; Lane 1–24: genotypes as mentioned at serial number 1–24 in supplementary Table 1

analysis identified 267 TE-like clones, of which 68 were assigned to 11 families based on their sequence alignment against a fully-characterized element. Recently, it was demonstrated that retrotransposons could be activated by interspecific hybridization that ultimately results in chromosomal rearrangements and the modification of adjacent gene expression, leading to polyploidization and bivalent prevalence at meiosis (Kashkush et al. 2003). Sugarcane is an example of such genomic rearrangements since it is believed to be the outcome of hybridization between two species, and such ‘hybridity’ is perpetuated through the routine vegetative propagation. The expression of retrotransposons were compared at different stages of polyploidization, and it was detected that out of 407,000 *Zea mays* (2n, 2500 Mbp) ESTs, 0.014% displayed similarity to retrotransposons (Meyers et al. 2001), while ~ 2.4% each of the transcriptomes were reported to be composed of TEs in the polyploid plant like *Triticum aestivum* (6n, 17,000 Mbp; Li et al. 2004) and sugarcane (8n, 3150 Mbp; Vettore et al. 2003). This suggests that copy number does change after polyploidization but progress after

transcriptional activation in the highly polyploid and complex genome of sugarcane.

The presence of one unigene EST (SUM01-068-G09-A-068.g) homologous to WD40 repeat proteins in leaf transcriptome is very interesting as these proteins play varied roles in developmental processes, e.g., RNA processing, transcriptional regulation (Williams et al. 1991; Hoey et al. 1993), mitotic spindle formation (Vaisman et al. 1995), regulation of vesicle formation and trafficking (Pryer et al. 1993) and cell division control (Feldman et al. 1997), etc., and also function in a variety of histone/chromatin-modifying complexes. The current study suggests that this protein may have some cross-talk with other proteins involved in development.

The unigene sequence (SUM01-075-F04-A-031.g) showed homology to homeodomain proteins expressed in leaf tissues. The homeobox is a semi-conserved sequence motif of about 180 bp that controls multiple biochemical pathways and cellular processes, it has been identified in different crops like rice (Jain et al. 2008) and sugarcane (Papini-Terzi et al. 2009). Knotted-like homeobox (*knox*) genes encode homeodomain-containing transcription factors and are required for meristem maintenance and organ initiation (Langdale 1998). In plants with simple leaves like *Arabidopsis*, expression of *knox* confined exclusively in the meristem and stem, while in those with dissected leaves, they are expressed in leaf primordia suggesting their possible role in leaf architecture diversity (Hake et al. 2004).

In the present study, the number of clusters found with repetitive sequences demonstrates its potential for SSR marker development in simple, quicker and economical manner. A total of 540 unique EST sequences were used for SSR search, of which 97 (17.9%) contained specified SSR motifs, generating 212 unique SSRs, which clearly demonstrate that sugarcane ESTs are a valuable resource for mining SSR markers. Notably, in this study, a relatively high frequency of SSRs in the ESTs was observed compared to the previous reports in other plant systems, e.g., maize (1.4%), *Medicago truncatula* (3%), wheat (3.2%),

Table 3 Putative functions associated with leaf ESTs under different functional categories

Functional class/putative function [#]	Unigene identifier [*]
<i>Bioenergetics and photosynthesis</i>	
Dehydroquinase, class II	SUM01-067-D01-A-010.g
Cytochrome c oxidase subunit 1	SUM01-064-A11-A-084.g
Quinone oxido-reductase	SUM01-064-B05-A-044.g
Photosystem II protein	SUM01-064-F09-A-078.g
Oxidoreductase	SUM01-066-H10-A-080.g
Alanine:glyoxylate aminotransferase 2 homolog	SUM01-069-F02-A-015.g
Photosystem I assembly protein	SUM01-072-E01-A-003.g, SUM01-073-H07-A-060.g
Photosystem I P700 apoprotein A2	SUM01-073-B03-A-025.g
NADH dehydrogenase	Contig23, Contig76, Contig77
Cytochrome c oxidase subunit 3/ subunit 4	Contig55, Contig62
Polyphosphate kinase 2	SUM01-062-C02-A-006.g
Pyruvate carboxylase	SUM01-063-E02-A-007.g
Aspartate1-decarboxylase	SUM01-067-G06-A-040.g
Cell metabolism	
Putative LytB protein	SUM01-061-H06-A-048.g
Cinnamyl alcohol dehydrogenase	SUM01-064-A05-A-036.g
Beta-expansin 1b	SUM01-074-D03-A-026.g
Putative cyclopropane synthase	SUM01-075-C08-A-054.g
BH3-only member B protein	Contig33
TOR2 kinase complex component	SUM01-061-C08-A-054.g
Chemotaxis protein PomB	SUM01-063-H11-A-092.g
Hyphal form cell wall	SUM01-061-B07-A-057.g
Cyto Kinesis defect family member (cyk-1)	SUM01-061-G09-A-068.g
Cell wall-anchored protein	SUM01-062-B09-A-073.g
Cell wall surface anchor family protein	SUM01-063-C08-A-054.g
Cinnamyl alcohol dehydrogenase	SUM01-064-A05-A-036.g
Protein-glutamatemethylsterase	SUM01-072-H05-A-044.g
Small GTP-binding protein domain	Contig37
Extensin precursor (Cell wall hydroxyproline-rich glycoprotein)/Putative extension/extensin precursor	SUM01-075-G12-A-088.g, SUM01-061-E04-A-023.g, SUM01-063-C01-A-002.g, SUM01-074-A06-A-037.g, SUM01-067-F02-A-015.g, SUM01-069-E11-A-083.g, SUM01-074-G09-A-068.g
<i>Development related protein</i>	
Formin/ Similar to formin 2	SUM01-071-F01-A-011.g, SUM01-074-F03-A-027.g, SUM01-071-F10-A-079.g
Maize egg apparatus-1 protein (ZmEA1)	SUM01-066-H02-A-016.g
WD-40 repeat protein	SUM01-068-G09-A-068.g
Putative Pollen specific protein C13 precursor	SUM01-071-E09-A-067.g
Similar to TEX14 protein	SUM01-064-A08-A-056.g
Early gene regulator	SUM01-064-E06-A-050.g
RPGR	SUM01-064-H11-A-095.g
DNA-directed RNA polymerase II largest subunit (RPB1) isoform 4	SUM01-065-B01-A-009.g
Similar to papilin	SUM01-066-A11-A-081.g
Disintegrin-like and metalloprotease	SUM01-066-D06-A-046.g
Regulatory protein, ArsR	SUM01-067-A03-A-017.g
DNA-directed RNA polymerase	SUM01-068-A02-A-005.g
Po protein	SUM01-074-A07-A-049.g
Putative homeodomain protein	SUM01-075-F04-A-031.g
<i>Membrane related</i>	

Table 3 continued

Functional class/putative function [#]	Unigene identifier [*]
SNAP91	SUM01-061-C09-A-066.g
Membrane glycoprotein	SUM01-061-F11-A-091.g
Putative membrane protein	SUM01-070-H07-A-060.g, SUM01-071-H06-A-048.g
Putative outer membrane protein	SUM01-072-F02-A-015.g
Uncharacterized conserved membrane protein	SUM01-067-B04-A-029.g
<i>Mobile genetic elements</i>	
Polyprotein	SUM01-074-G01-A-004.g, SUM01-062-A01-A-001.g, SUM01-068-H04-A-032.g, SUM01-074-G01-A-004.g
Gag-pol precursor	Contig43, SUM01-062-F06-A-047.g, SUM01-073-D01-A-010.g, SUM01-062-H02-A-016.g
Transposon protein, Putative, CACTA, En/Spm sub-class	Contig16, SUM01-067-F06-A-047.g
Putative TNP2 transposon	Contig64
Ac-like transposase	SUM01-061-F09-A-075.g
Putative copia-type pol polyprotein	SUM01-062-G04-A-024.g
RIRE2 orf3	SUM01-064-E07-A-054.g
Transposon protein/ Retrotransposon protein	SUM01-067-E11-A-083.g, SUM01-069-B04-A-029.g, SUM01-075-E11-A-083.g, SUM01-070-A07-A-049.g, SUM01-070-C08-A-054.g
Putative arm repeat-containing protein	SUM01-071-E03-A-019.g
hAT family dimerisation domain	SUM01-068-C02-A-006.g
Retrovirus-related Pol poly protein from transposon	Contig17
<i>Signal transduction</i>	
Protein kinase	Contig44, Contig50, SUM01-068-C05-A-034.g, SUM01-066-G10-A-072.g, SUM01-073-A12-A-085.g
Aluminium-induced protein	SUM01-075-H07-A-060.g
Similar to Ras and Rab interactor	SUM01-061-A02-A-005.g
aPP1c-relatedser/ Hr protein phosphataseZ isoform	SUM01-070-D07-A-058.g
Similar to synapse defective1, RhoGTPase, homolog 1	SUM01-072-D10-A-078.g
Metallo phosphoesterase	SUM01-074-A12-A-085.g
Probable transcriptional regulator, AraC family	SUM01-073-C08-A-054.g
<i>DNA metabolism</i>	
RNA-directed DNA polymerase	SUM01-064-E04-A-034.g
Similar to exonuclease1 isoform b	SUM01-065-C01-A-002.g
DNA packaging protein	SUM01-071-F03-A-027.g, SUM01-074-C06-A-038.g
Putative type I restriction-modification system methylase	SUM01-075-E06-A-039.g
UvrD/REP helicase	SUM01-075-B01-A-009.g
<i>RNA metabolism</i>	
RNA-binding protein	SUM01-061-D03-A-026.g
Zinc finger protein	SUM01-063-H03-A-028.g
Putative translation initiation factorIF-2	SUM01-069-F03-A-027.g
SFPQ protein	SUM01-074-E05-A-035.g
SRRM2 protein	SUM01-074-A03-A-017.g
<i>Protein metabolism</i>	
26S proteasome regulatory particle non-ATPase subunit 8	Contig12
Beta-1,3-glucan synthase	SUM01-065-D07-A-058.g
Myosin	Contig25
RibonucleaseA/angiogenin	SUM01-061-H05-A-044.g
Nicotinic acid mononucleotide adenylyl transferase	SUM01-065-B07-A-057.g
Pherophorin-C2proteinprecursor	SUM01-067-F05-A-043.g
Glycoprotein C-1	SUM01-075-F01-A-011.g

Table 3 continued

Functional class/putative function [#]	Unigene identifier*
16SrRNAuridine-516pseudouridylate synthase and related pseudouridylate synthases	SUM01-070-B12-A-093.g
<i>Metabolic process</i>	
Alpha amylase, catalytic region	SUM01-072-C10-A-070.g
Lipoic acid synthetase	SUM01-073-C03-A-018.g
Putative glucose-6-phosphate dehydrogenase	Contig26
Putative 8-amino-7-oxononanoate synthase	Contig34
Expressed protein	SUM01-066-G01-A-004.g
Calcineurin-related phosphoesterase-like	SUM01-071-C04-A-022.g
Cytochrome P450	Contig13, Contig63
Glycosyl transferase	Contig70
Putative reductase component of monooxygenase	Contig72
SOCS box-containing WD protein SWiP-2	SUM01-063-E01-A-003.g
LacZ	SUM01-066-E09-A-067.g
<i>Transcription factors</i>	
Putative NEC 1	SUM01-074-D05-A-042.g
Zinc finger	SUM01-070-H12-A-096.g
Zinc knuckle domain containing protein-like	SUM01-074-C08-A-054.g
<i>Transport</i>	
POT family	Contig74
Putative calcium ATPase	SUM01-065-B12-A-093.g
Putative phosphate transporter	SUM01-071-D01-A-010.g
Drug resistance transporter EmrB/QacA subfamily	SUM01-068-C10-A-070.g
Monovalent cation-transporting P-type ATPase	SUM01-063-D03-A-026.g
POT1 andTIN2 interacting protein	SUM01-063-G09-A-068.g
Oligopeptide porter	SUM01-064-D08-A-073.g
Exocyst complex component Sec15B	SUM01-064-E10-A-082.g
Collagen alpha1(III) chain precursor	SUM01-065-A07-A-049.g
Transmembrane protein 45b	SUM01-065-A10-A-069.g
Putative portal protein	SUM01-065-G04-A-024.g
Outer membrane protein	SUM01-065-G05-A-036.g
Putativ emembrane protein	SUM01-067-F08-A-063.g
Phosphocarrier HPr protein	SUM01-067-H10-A-080.g
Sugar ABC transporter substrate-binding protein	SUM01-069-A06-A-037.g
Spermidine/putrescine ABC transporter	SUM01-069-B01-A-009.g
Sec Y-independent transporter protein	SUM01-072-D05-A-042.g
VDP/USO1/YBL047C family vesicular transport factor	SUM01-072-E09-A-067.g
ABC transporter putativ permease	SUM01-070-E09-A-067.g
<i>Stress/Defence response related</i>	
Similar to Killer cell immunoglobulin-like receptor 3DL1 precursor	SUM01-062-G08-A-056.g
Homospermidine synthase-like	SUM01-062-D01-A-010.g
Anti-freeze glycopeptide AFGP polypeptide precursor	SUM01-061-F12-A-095.g
Ligand-independent activating molecule for estrogen receptor	SUM01-061-D08-A-062.g
Cf2/Cf5 disease resistance protein homolog	SUM01-069-H10-A-080.g
Plant disease resistance polypeptide-like	SUM01-071-G12-A-088.g
Putative glutathione transporter	Contig39

Table 3 continued

Functional class/putative function [#]	Unigene identifier [*]
Proline-rich protein/proline-rich proteoglycan2	SUM01-067-H01-A-012.g, SUM01-075-F12-A-095.g, SUM01-069-D10-A-078.g, SUM01-070-C02-A-006.g, SUM01-073-D12-A-094.g, SUM01-074-A10-A-069.g, SUM01-074-C02-A-006.g, SUM01-075-B03-A-025.g, SUM01-063-B07-A-057.g, SUM01-068-F01-A-011.g
<i>Proteins with no 'obvious' plant functions</i>	
Myeloid/lymphoid or mixed-lineage leukemia protein 2	SUM01-063-G01-A-004.g
Bacterial surface proteins containing Ig-like domains	SUM01-067-D07-A-058.g
Histidine-rich protein	SUM01-061-C12-A-86.g, SUM01-061-E09-A-067.g
Proteo phosphoglycan	SUM01-061-E11-A-083.g
Similar to SRY (sex determining regionY)-box 3	SUM01-067-F04-A-031.g
Immediate early protein	SUM01-062-A12-A-085.g
Similar to tripartite motif protein	SUM01-062-C10-A-070.g
Phage-related protein	SUM01-066-A01-A-001.g
Cro protein	SUM01-070-B10-A-077.g
BHLF1 [Human herpes virus 4]	SUM01-070-C11-A-082.g
Immunoglobulin mu heavy chain	SUM01-072-B05-A-041.g
Special lobe-specific silk protein ssp 160	SUM01-070-E10-A-071.g
Mucin2/ MUC3B mucin/Mucin 2 precursor	Contig65, Contig49, SUM01-061-A11-A-081.g, SUM01-062-C06-A-038.g, SUM01-062-D03-A-026.g, SUM01-062-D04-A-030.g, SUM01-062-F08-A-063.g, SUM01-061-F08-A-063.g, SUM01-063-B11-A-089.g, SUM01-063-F11-A-091.g, SUM01-066-E04-A-023.g,
Viral protein	Contig38, Contig41, Contig46, Contig66, Contig67, Contig75, SUM01-061-E05-A-035.g, SUM01-062-B01-A-009.g
Tail component	SUM01-064-A06-A-040.g, SUM01-064-F11-A-094.g, SUM01-072-F07-A-059.g
<i>Unable to classify</i>	
Prpol	Contig19, Contig35
KED-like protein	SUM01-075-E03-A-019.g
Host specificity protein	Contig27
Proteo phospho glycan	SUM01-062-F09-A-075.g
TPA:Hornerin	SUM01-062-G07-A-052.g
Flocculin-likeprotein	SUM01-063-C10-A-070.g
E 2structural protein	SUM01-063-D06-A-046.g
Erythrocyte membrane protein	SUM01-067-D05-A-042.g
Putative hosts pecificity protein	SUM01-069-B06-A-045.g
Formin 2	SUM01-069-H01-A-012.g
Formin 3	SUM01-069-H05-A-044.g
Proteo phosphoglycan	SUM01-071-B06-A-045.g
Rrf2/aminotransferase, class V family protein	SUM01-074-D10-A-078.g
<i>Unknown protein</i>	
Germination specific N-acetylmuramoyl-L-alanine amidase	SUM01-062-C08-A-054.g
PREDICTED:similar to C61protein	SUM01-068-G07-A-052.g
gp2	SUM01-062-G10-A-072.g
PREDICTED: similar to plexinD1	SUM01-062-H11-A-092.g
Olfactory receptor	SUM01-063-G04-A-024.g
Granule cell marker protein	SUM01-067-G10-A-072.g
LigA	SUM01-068-A12-A-085.g
Pherophorin-C2 protein precursor	SUM01-068-B01-A-009.g
Salivary gland secretion ICG3047-PA	SUM01-068-H03-A-028.g

Table 3 continued

Functional class/putative function [#]	Unigene identifier*
Unknown protein	
Contig15, Contig36, Contig54, Contig71,	
SUM01-067-H05-A-044.g, SUM01-068-C01-A-002.g, SUM01-069-C06-A-038.g, SUM01-070-H06-A-048.g, SUM01-073-C12-A-086.g, SUM01-071-H08-A-064.g, SUM01-061-G02-A-008.g, SUM01-061-H11-A-092.g, SUM01-062-A11-A-081.g, SUM01-062-C05-A-034.g, SUM01-067-C10-A-070.g, SUM01-068-B02-A-013.g, SUM01-069-H06-A-048.g, SUM01-070-E08-A-055.g, SUM01-070-H05-A-044.g, SUM01-071-D05-A-042.g, SUM01-071-D09-A-074.g, SUM01-071-E05-A-035.g, SUM01-072-A05-A-033.g, SUM01-075-D06-A-046.g, SUM01-072-E05-A-035.g, SUM01-075-D01-A-010.g, SUM01-074-B01-A-009.g, SUM01-074-G11-A-084.g	
Unnamed protein	
Contig3, Contig11, Contig20, Contig58,	
SUM01-062-B06-A-045.g, SUM01-062-F03-A-027.g, SUM01-066-G07-A-052.g, SUM01-067-G11-A-084.g, SUM01-075-D10-A-078.g, SUM01-069-A05-A-033.g, SUM01-071-A06-A-037.g, SUM01-074-A11-A-081.g, SUM01-074-E06-A-039.g, SUM01-074-G04-A-024.g, SUM01-075-B06-A-045.g, SUM01-061-C02-A-006.g, SUM01-061-C11-A-082.g, SUM01-061-D04-A-030.g, SUM01-061-E03-A-019.g, SUM01-061-G07-A-052.g, SUM01-061-H12-A-096.g, SUM01-062-C09-A-066.g, SUM01-062-D10-A-078.g, SUM01-062-E07-A-051.g, SUM01-062-G09-A-068.g, SUM01-068-H05-A-044.g, SUM01-063-A10-A-069.g, SUM01-063-B06-A-045.g, SUM01-063-C11-A-082.g, SUM01-063-C12-A-086.g, SUM01-063-D10-A-078.g, SUM01-063-D11-A-090.g, SUM01-063-E06-A-039.g, SUM01-063-E11-A-083.g, SUM01-063-F01-A-011.g, SUM01-063-F12-A-095.g, SUM01-063-G06-A-040.g, SUM01-064-A01-A-002.g, SUM01-069-E04-A-023.g, SUM01-070-G03-A-020.g, SUM01-066-F01-A-011.g, SUM01-075-G05-A-036.g, SUM01-075-G03-A-020.g, SUM01-068-G06-A-040.g, SUM01-069-F06-A-047.g, SUM01-071-C08-A-054.g, SUM01-071-F08-A-063.g, SUM01-071-G11-A-084.g, SUM01-073-D07-A-058.g, SUM01-073-H09-A-076.g, SUM01-074-D06-A-046.g, SUM01-074-E07-A-051.g, SUM01-074-E11-A-083.g, SUM01-075-A10-A-069.g	
Hypothetical protein	
Contig5, Contig6, Contig7, Contig10, Contig14, Contig18, Contig21, Contig22, Contig24, Contig28, Contig29, Contig30, Contig31, Contig40, Contig42, Contig47, Contig51, Contig56, Contig57, Contig61	
SUM01-061-A06-A-037.g, SUM01-061-B12-A-093.g, SUM01-061-D06-A-046.g, SUM01-062-D02-A-014.g, SUM01-062-F11-A-091.g, SUM01-063-B03-A-025.g, SUM01-063-H08-A-064.g, SUM01-064-F07-A-062.g, SUM01-067-D02-A-014.g, SUM01-068-F05-A-043.g, SUM01-075-D05-A-042.g, SUM01-075-D03-A-026.g, SUM01-075-C05-A-034.g, SUM01-061-A08-A-053.g, SUM01-061-A09-A-065.g, SUM01-068-C06-A-038.g, SUM01-061-B02-A-013.g, SUM01-061-B08-A-061.g, SUM01-061-B09-A-073.g, SUM01-061-B10-A-077.g, SUM01-061-B11-A-089.g, SUM01-061-C10-A-070.g, SUM01-061-D05-A-042.g, SUM01-061-D07-A-058.g, SUM01-061-D10-A-078.g, SUM01-061-D11-A-090.g, SUM01-061-E07-A-051.g, SUM01-061-E10-A-071.g, SUM01-061-E12-A-087.g, SUM01-061-F01-A-011.g, SUM01-061-F05-A-043.g, SUM01-061-F10-A-079.g, SUM01-061-G01-A-004.g, SUM01-061-G05-A-036.g, SUM01-061-G10-A-072.g, SUM01-061-G12-A-088.g, SUM01-062-B08-A-061.g, SUM01-062-C04-A-022.g, SUM01-062-C07-A-050.g, SUM01-062-D09-A-074.g, SUM01-062-D12-A-094.g, SUM01-062-E05-A-035.g, SUM01-062-E09-A-067.g, SUM01-062-E10-A-071.g, SUM01-062-G02-A-008.g, SUM01-062-H10-A-080.g, SUM01-063-A03-A-017.g, SUM01-063-A08-A-053.g, SUM01-063-A09-A-065.g, SUM01-063-A12-A-085.g, SUM01-063-B08-A-061.g, SUM01-063-B09-A-073.g, SUM01-063-B10-A-077.g, SUM01-063-B12-A-093.g, SUM01-063-D08-A-062.g, SUM01-063-D12-A-094.g, SUM01-063-E08-A-055.g, SUM01-063-E10-A-071.g, SUM01-063-F08-A-063.g, SUM01-063-F09-A-075.g, SUM01-063-F10-A-079.g, SUM01-063-G03-A-020.g, SUM01-064-B04-A-032.g, SUM01-064-C07-A-053.g, SUM01-075-H04-A-032.g, SUM01-075-H03-A-028.g, SUM01-064-G07-A-055.g, SUM01-064-G08-A-067.g, SUM01-075-H02-A-016.g, SUM01-070-A03-A-017.g, SUM01-065-C07-A-050.g, SUM01-065-E02-A-007.g, SUM01-065-F03-A-027.g, SUM01-065-F04-A-031.g, SUM01-065-G06-A-040.g, SUM01-065-H06-A-048.g, SUM01-066-D09-A-074.g, SUM01-066-E02-A-007.g, SUM01-066-E08-A-055.g, SUM01-066-E12-A-087.g, SUM01-066-H04-A-032.g, SUM01-066-H09-A-076.g, SUM01-067-E06-A-039.g, SUM01-067-G08-A-056.g, SUM01-067-H12-A-096.g, SUM01-068-B09-A-073.g, SUM01-075-F09-A-075.g, SUM01-072-A12-A-085.g, SUM01-069-D11-A-090.g, SUM01-070-B02-A-013.g, SUM01-070-B08-A-061.g, SUM01-070-E07-A-051.g, SUM01-070-E12-A-087.g, SUM01-070-F11-A-091.g, SUM01-070-F12-A-095.g, SUM01-070-G09-A-068.g, SUM01-070-H04-A-032.g, SUM01-070-H09-A-076.g, SUM01-071-B02-A-013.g, SUM01-071-B04-A-029.g, SUM01-071-H12-A-096.g, SUM01-073-G10-A-072.g, SUM01-072-F06-A-047.g, SUM01-072-G03-A-020.g, SUM01-073-B12-A-093.g, SUM01-073-D06-A-046.g, SUM01-074-C10-A-070.g, SUM01-074-E02-A-007.g, SUM01-074-E03-A-019.g, SUM01-073-G11-A-084.g, SUM01-074-E08-A-055.g, SUM01-074-A04-A-021.g, SUM01-074-B08-A-061.g, SUM01-074-F04-A-031.g, SUM01-074-F09-A-075.g, SUM01-074-B11-A-089.g, SUM01-074-G03-A-020.g, SUM01-074-H02-A-016.g, SUM01-074-F11-A-091.g, SUM01-074-G08-A-056.g, SUM01-074-H09-A-076.g, SUM01-074-H10-A-080.g, SUM01-075-A03-A-017.g	

[#]Putative annotation based on GenBank nr (BLASTx); *Unigene (non redundant) EST sequences identity number

barley (3.4%), sorghum (3.6%) and rice (4.7%) (Kantety et al. 2002; Eujayl et al. 2004). The frequency of SSRs has been reported to be dependent on a variety of factors, e.g., SSR search criteria, size of the dataset, database-mining tools, etc. (Varshney et al. 2005). Sugarcane is highly heterozygous and harbor higher levels of genomic diversity (Ming et al. 1998; Soltis and Soltis 2000); both the

conditions are favored by the occurrence of SSRs within coding sequences. In this study, redundancy was searched and eliminated manually prior to analysis so as to obtain an effective dataset. Since random sequencing within cDNA libraries generally leads to a high proportion of redundant ESTs, NR ESTs was used to avoid such overestimations. In the case of *Arabidopsis thaliana*, using a similar approach,

Kumapatla and Mukhopadhyay (2005) reported 37.3% fewer SSR-ESTs.

In the present study, tetra-nucleotide repeat was the most abundant, followed by tri-nucleotide repeats; interestingly, in sugarcane, it was found that AC/TG repeat was abundant. A similar high frequency of such a repeat has also been reported in soybean, maize, rice and wheat (Gao et al. 2003). In terms of the reading frames, it has been opined that dinucleotide motif could represent multiple codons leading to the translation event (Kantety et al. 2002). For example, the AG repeat motif may also be read as GA repeat which could be translated in to amino acids like Ser, Arg, Asp or Glu. In this study, among the dinucleotide motifs, AG and AC were the most abundant types, whereas, (CG)_n motif was rare. The AG and AC motifs have been previously reported to be the most frequently observed SSRs in other plant systems (Gao et al. 2003; Saha et al. 2004; Scott et al. 2000). In this scenario, the putative functions of these abundant species-specific motifs need detailed investigations. In plant systems, TC and CTT repeats were found to be typical of transcribed regions that occur in high frequency in the 5' untranslated regions (Kantety et al. 2002; Moccia et al. 2009). Such high frequency of (TC)_n repeats might be due to its translation into Ala and Leu (Kantety et al. 2002), which are present in proteins at high proportions of 8 and 10%, respectively. The (AT)_n repeats are reported to be abundant in many plant systems (Morgante and Olivieri 1993), but these were found to be rare in the current ESTs of sugarcane. Such rare occurrence of (AT)_n SSRs is in accordance with previous reports in maize (Chin et al. 1996), rice (Temnykh et al. 2000), and *Arabidopsis* (Cardle et al. 2000).

The trimeric motifs, (ACG)_n, (AGA)_n and (ATC)_n were the most common (70%) in this study, which is in agreement with the reports in rice where 60% of EST-derived microsatellite sequences were trinucleotide repeats [e.g., (CCG)_n, (ACG)_n, (AGG)_n, (ACC)_n] (Temnykh et al. 2000). A similar high frequency of (ATC)_n repeats has also been reported in *Arabidopsis* (Cardle et al. 2000). In a previous report in sugarcane (Cordeiro et al. 2001), the (CCG)_n motif was reported to be the most common motif, but interestingly, this motif was found to be rare in this study. Such discrepancy may be attributed to the differences in data size and SSR search criteria since these factors may significantly alter the estimates of frequency/distribution of EST-SSRs (Varshney et al. 2005). This suggests a need to formulate universally acceptable criteria to identify SSR. Notably, the motif (AAT)_n, reported to be rare in different members of Poaceae, e.g., barley, rice, maize, and sugarcane, as well as in *Arabidopsis*, was not detected in this study. This could be attributed to the fact

that TAA-based variants code for stop codons (Chin et al. 1996).

The putative functions of tri-nucleotide repeats in ESTs are under debate. Morgante et al. (2002) opined that two-fold frequency of trinucleotide repeats in the coding region may lead to mutations including bias selection for specific single amino acid stretches. Gao et al. (2003) reported that trinucleotides were associated with genes for important functions like stress tolerance, transcription regulation, enzyme biosynthesis, signal transduction, etc. Notably, AAC repeats prevailing in wheat ESTs are associated with storage proteins like glutenin and gliadin, which impart quality. The establishment of a relationship between a trinucleotide repeat and its specific functional categories of genes could be the first step in characterizing novel trinucleotide repeat-containing genes.

The EST-derived SSR markers possess the inherent merits of the conventional genomic SSRs, in addition, they improve the ability to detect marker-trait associations since they are part of the transcribed domain(s). Due to this, in recent years, the emphasis has been shifted towards development of functional molecular markers to assay diversity in a natural population (Andersen and Lubberstedt 2003). In addition, a practical advantage of EST-SSR markers is that these are highly transferable across the species mainly due to their higher sequence conservation (Varshney et al. 2005). For EST-SSRs, an amplification success rate of 60–80% has been reported in different studies, e.g., 60–70% (sugarcane; Cordeiro et al. 2001, Oliveira et al. 2009), 64% (barley; Thiel et al. 2003), 80.9% (coffee; Aggarwal et al. 2007). In the present study, a high proportion (72%) of EST-SSRs turn out to be functional, which could be due to the fact that primer pair encompasses large introns in the genomic regions. Oliveira et al. (2009) also reported such differences in the expected and observed size of the amplicons, which could be due to amplification of small introns. Nicot et al. (2004) opined that EST-SSRs that had smaller amplicon than expected size may result from a single deletion within the sequence framed by the two primers, non-specific annealing of the primers, duplication of the EST sequences in the genome, EST from multigenic family, or a primer designed from the conserved domain.

The identified EST-SSRs were validated in six sugarcane genotypes, and one primer pair was also tested with 24 genotypes. Results from these experiments revealed a high level of polymorphism. Such observations are in conformity with the previous reports of Oliveira et al. (2009) and Singh et al. (2013). This suggests that the markers revealing high polymorphism could be useful in the assessment of parentage as well as linkage mapping studies. The EST-SSR primers on account of greater DNA sequence conservation have been reported to yield less

polymorphism compared to the genomic SSRs (Pinto et al. 2006; Oliveira et al. 2009). In this study, using ESTs-derived SSRs, the putative function could be deduced for 28% of ESTs, and which may increase with further enrichment of protein databases. The remaining (72%) ESTs-derived SSRs revealed ‘no-hit’ (18%), ‘other proteins’ (10%) or ‘hypothetical protein’ (44%); these may illustrate the specific transcriptome of sugarcane genotype CoS 767, which is yet to be well characterized for its putative functions.

The genomic complexity and high polyploidy of the present-day sugarcane cultivars necessitate identifying a large number of markers for the comprehensive genome analysis. The abundance of polymorphic microsatellites in the transcribed regions makes EST libraries a valuable resource for an enrichment of genomic information in genetic studies. It is expected that the EST-derived markers developed in this study would make a substantial contribution in the delineation of structure/function of sugarcane genome, and ultimately benefit ongoing sugarcane improvement programs.

Acknowledgements The authors gratefully acknowledge the help at the Interdisciplinary Centre for Plant Genomics (ICPG), Department of Plant Molecular Biology, UDSC, New Delhi in cDNA library sequencing and ESTs generation. The financial support provided by Department of Biotechnology (DBT) Ministry of Science and Technology, Government of India, New Delhi is gratefully acknowledged.

Authors’ contributions MSK and RK conceived, designed and performed the experiments. MSK and SK prepared the manuscript. MSK, RK, and RKS analyzed the data. JS and SKD helped in the field experiments. RK and SKD provided important inputs to improve the manuscript. All the authors read and approved the final manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Aggarwal RK, Hendre PS, Varshney RK, Bhat PR, Krishnakumar V, Singh L (2007) Identification, characterization and utilization of EST-derived genic microsatellite markers for genome analyses of coffee and related species. *Theor Appl Genet* 114:359–372. <https://doi.org/10.1007/s00122-006-0440-x>

Aitken KS, Jackson PA, McIntyre CL (2006) Quantitative trait loci identified for sugar related traits in a sugarcane (*Saccharum* spp.) cultivar x *Saccharum officinarum* population. *Theor Appl Genet* 112:1306–1317. <https://doi.org/10.1007/s00122-006-0233-2>

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)

Andersen JR, Lubberstedt T (2003) Functional markers in plants. *Trends Plant Sci* 8:554–560. <https://doi.org/10.1016/j.tplants.2003.09.010>

Andrietta MH, Eloy NB, Hemerly AS, Ferreira PCG (2001) Identification of sugarcane cDNAs encoding components of the cell cycle machinery. *Genet Mol Bio* 24:61–68

Bennetzen JL (2000) Transposable element contributions to plant gene and genome evolution. *Plant Mol Biol* 42:251–269

Borges JCPM, Ramos CHI (2001) Molecular chaperone genes in the sugarcane expressed sequence database (SUCEST). *Genet Mol Biol* 24:85–92

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

Carson DL, Botha FC (2000) Preliminary analysis of expressed sequence tags for sugarcane. *Crop Sci* 40:1769–1779

Chin EC, Senior ML, Shu H, Smith JS (1996) Maize simple repetitive DNA sequences: abundance and allele variation. *Genome* 39:866–873

Chomczynski P, Sacchi N (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal Biochem* 162:156–159. <https://doi.org/10.1006/abio.1987.9999>

Cordeiro GM, Casu R, McIntyre CL, Manners JM, Henry RJ (2001) Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to erianthus and sorghum. *Plant Sci* 160:1115–1123

Costa RMA et al (2001) DNA repair-related genes in sugarcane expressed sequence tags (ESTs). *Genet Mol Bio* 24:131–140

Costet L et al (2012) Haplotype structure around Bru1 reveals a narrow genetic basis for brown rust resistance in modern sugarcane cultivars. *Theor Appl Genet* 125:825–836. <https://doi.org/10.1007/s00122-012-1875-x>

Da Silva JA, Bressiani JA (2005) Sucrose synthase molecular marker associated with sugar content in elite sugarcane progeny. *Genet Mol Bio* 28:294–298

D’Hont A, Grivet L, Feldmann P, Rao S, Berding N, Glaszmann JC (1996) Characterisation of the double genome structure of modern sugarcane cultivars (*Saccharum* spp.) by molecular cytogenetics. *Mol Gen Genet* 250:405–413

Dornelas MC, Rodrigue APM (2001) A genomic approach to elucidating grass flower development. *Genet Mol Bio* 24:69–76

Doyle JJ, Doyle JL (1990) Isolation of plant DNA from fresh tissue. *Focus* 12:13–14

Eujayl I, Sledge MK, Wang L, May GD, Chekhovskiy K, Zwonitzer JC, Mian MA (2004) *Medicago truncatula* EST-SSRs reveal cross-species genetic markers for *Medicago* spp. *Theor Appl Genet* 108:414–422. <https://doi.org/10.1007/s00122-003-1450-6>

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8:186–194

FAOSTAT (2016) <http://faostat.fao.org>. Accessed October 20, 2016

Feldman RM, Correll CC, Kaplan KB, Deshaies RJ (1997) A complex of Cdc4p, Skp1p, and Cdc53p/cullin catalyzes ubiquitination of the phosphorylated CDK inhibitor Sic1p. *Cell* 91:221–230

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: where genetics meets genomics. *Nat Rev Genet* 3:329–341. <https://doi.org/10.1038/nrg793>

Gao LF, Tang JF, Li HW, Jia JZ (2003) Analysis of microsatellites in major crops assessed by computational and experimental approaches. *Mol Breed* 12:245–261

Gupta V, Raghuvanshi S, Gupta A, Saini N, Gaur A, Khan MS et al (2010) The water-deficit stress- and red-rot-related genes in sugarcane. *Funct Integr Genomics* 10:207–214. <https://doi.org/10.1007/s10142-009-0144-9>

Hake S, Smith HM, Holtan H, Magnani E, Mele G, Ramirez J (2004) The role of knox genes in plant development. *Annu Rev Cell*

- Dev Biol 20:125–151. <https://doi.org/10.1146/annurev.cellbio.20.031803.093824>
- Hoey T, Weinzierl RO, Gill G, Chen JL, Dynlacht BD, Tjian R (1993) Molecular cloning and functional analysis of Drosophila TAF110 reveal properties expected of coactivators. Cell 72:247–260
- Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877
- Jain M, Tyagi AK, Khurana JP (2008) Genome-wide identification, classification, evolutionary expansion and expression analyses of homeobox genes in rice. FEBS J 275:2845–2861. <https://doi.org/10.1111/j.1742-4658.2008.06424.x>
- Jeffery CJ (2003) Moonlighting proteins: old proteins learning new tricks. Trends Genet 19:415–417. [https://doi.org/10.1016/S0168-9525\(03\)00167-7](https://doi.org/10.1016/S0168-9525(03)00167-7)
- Jordan DR, Casu RE, Besse P, Carroll BC, Berding N, McIntyre CL (2004) Markers associated with stalk number and suckering in sugarcane colocate with tillering and rhizomatousness QTLs in sorghum. Genome 47:988–993. <https://doi.org/10.1139/g04-040>
- Kantety RV, La Rota M, Matthews DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol Biol 48:501–510
- Kashkush K, Feldman M, Levy AA (2003) Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat Genet 33:102–106. <https://doi.org/10.1038/ng1063>
- Keith CS, Hoang DO, Barrett BM, Feigelman B, Nelson MC, Thai H, Baysdorfer C (1993) Partial sequence analysis of 130 randomly selected maize cDNA clones. Plant Physiol 101:329–332
- Khan MS, Khraiweh B, Pugalenth G, Gupta RS, Singh J, Duttamajumder SK, Kapur R (2014) Subtractive hybridization-mediated analysis of genes and in silico prediction of associated microRNAs under waterlogged conditions in sugarcane (*Saccharum* spp.). FEBS Open Bio 4:533–541. <https://doi.org/10.1016/j.fob.2014.05.007>
- Kretz PL, Danylchuk T, Hareld W, Wells S, Provost GS (1994) Gigapack III high-efficiency lambda packaging extract with single-tube convenience. Strategies 7:44–45
- Kumpatla SP, Mukhopadhyay S (2005) Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. Genome 48:985–998. <https://doi.org/10.1139/g05-060>
- Langdale JA (1998) Cellular differentiation in the leaf. Curr Opin Cell Biol 10:734–738
- Li W, Zhang P, Fellers JP, Friebe B, Gill BS (2004) Sequence composition, organization, and evolution of the core Triticeae genome. Plant J 40:500–511. <https://doi.org/10.1111/j.1365-313X.2004.02228.x>
- McClintock B (1984) The significance of responses of the genome to challenge. Science 226:792–801
- Meyers BC, Tingey SV, Morgante M (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res 11:1660–1676. <https://doi.org/10.1101/gr.188201>
- Ming R et al (1998) Detailed alignment of *saccharum* and sorghum chromosomes: comparative organization of closely related diploid and polyploid genomes. Genetics 150:1663–1682
- Moccia MD, De-Desfeux CO, Marais GAB, Widmer A (2009) A White Campion (*Silene latifolia*) floral expressed sequence tag (EST) library: annotation, EST-SSR characterization, transferability, and utility for comparative mapping. BMC Genom 10:243
- Morgante M, Olivieri AM (1993) PCR-amplified microsatellites as markers in plant genetics. Plant J 3:175–182
- Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30:194–200. <https://doi.org/10.1038/ng822>
- Newman T et al (1994) Genes galore: a summary of methods for accessing results from large-scale partial sequencing of anonymous Arabidopsis cDNA clones. Plant Physiol 106:1241–1255
- Nicot N et al (2004) Study of simple sequence repeat (SSR) markers from wheat expressed sequence tags (ESTs). Theor Appl Genet 109:800–805. <https://doi.org/10.1007/s00122-004-1685-x>
- Oliveira KM et al (2009) Characterization of new polymorphic functional markers for sugarcane. Genome 52:191–209. <https://doi.org/10.1139/g08-105>
- Papini-Terzi FS et al (2009) Sugarcane genes associated with sucrose content. BMC Genom 10:120. <https://doi.org/10.1186/1471-2164-10-120>
- Piatigorsky J (2007) Gene sharing and evolution: the diversity of protein functions. Harvard University Press, Cambridge
- Pinto LR, Oliveira KM, Marconi TG, Garcia AAF, Ulian EC, Souza AP (2006) Characterization of novel sugarcane expressed sequence tags microsatellites and their comparison with genomic SSRs. Plant Breed 125:378–384
- Pryer NK, Salama NR, Schekman R, Kaiser CA (1993) Cytosolic Sec13p complex is required for vesicle formation from the endoplasmic reticulum in vitro. J Cell Biol 120:865–875
- Rossi M, Araujo PG, Van Sluys MA (2001) Survey of transposable elements in sugarcane expressed sequence tags (ESTs). Genet Mol Biol 24:147–154
- Saha MC, Mian MA, Eujayl I, Zwonitzer JC, Wang L, May GD (2004) Tall fescue EST-SSR markers with transferability across several grass species. Theor Appl Genet 109:783–791. <https://doi.org/10.1007/s00122-004-1681-1>
- Santelli RV, Siviero F (2001) A search for homologues of plant photoreceptor genes and their signaling partners in the sugarcane expressed sequence tag (SUCEST) database. Genet Mol Bio 24:49–53
- Scott KD, Eggler P, Seaton G, Rossetto M, Ablett EM, Lee LS, Henry RJ (2000) Analysis of SSRs derived from grape ESTs. Theor Appl Genet 100(5):723–726
- Singh RK et al (2013) Development, cross-species/genera transferability of novel EST-SSR markers and their utility in revealing population structure and genetic diversity in sugarcane. Gene 524:309–329. <https://doi.org/10.1016/j.gene.2013.03.125>
- Singh RK et al (2016) Identification of putative candidate genes for red rot resistance in sugarcane (*Saccharum* species hybrid) using LD-based association mapping. Mol Genet Genomics 291:1363–1377. <https://doi.org/10.1007/s00438-016-1190-3>
- Soares Netto LE (2001) Oxidative stress response in sugarcane. Genet Mol Biol 24:93–102
- Soltis PS, Soltis DE (2000) The role of genetic and genomic attributes in the success of polyploids. Proc Natl Acad Sci USA 97:7051–7057
- Srivastava HM, Pandey S, Tripathi BK (1989) Recent advances in genetics and plant breeding. Paper presented at the national symposium B.H.U., Varanasi, Nov 15–16
- Stevenson GC (1965) Genetics and breeding of sugarcane. Longmans, Green Co. Ltd., London
- Sugarcane.org (2016) <http://sugarcane.org/sugarcane-products/ethanol>. Accessed October 20, 2016
- Temnykh S et al (2000) Mapping and genome organization of microsatellite sequences in rice (*Oryza sativa* L.). Theor Appl Genet 100(5):697–712
- Thiel T, Michalek W, Varshney RK, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet 106:411–422. <https://doi.org/10.1007/s00122-002-1031-0>
- Vaisman N, Tsouladze A, Robzyk K, Ben-Yehuda S, Kupiec M, Kassir Y (1995) The role of *Saccharomyces cerevisiae* Cdc40p

- in DNA replication and mitotic spindle formation and/or maintenance. *Mol Gen Genet* 247:123–136
- Van De Loo FJ, Turner S, Somerville C (1995) Expressed sequence tags from developing castor seeds. *Plant Physiol* 108:1141–1150
- Varshney RK et al (2005) Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. *Plant Sci* 168(1):195–202
- Vettore AL et al (2003) Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. *Genome Res* 13:2725–2735. <https://doi.org/10.1101/gr.1532103>
- Vincentz M et al (2004) Evaluation of monocot and eudicot divergence using the sugarcane transcriptome. *Plant Physiol* 134:951–959. <https://doi.org/10.1104/pp.103.033878>
- Williams FE, Varanasi U, Trumbly RJ (1991) The CYC8 and TUP1 proteins involved in glucose repression in *Saccharomyces cerevisiae* are associated in a protein complex. *Mol Cell Biol* 11:3307–3316
- Yamamoto K, Sasaki T (1997) Large-scale EST sequencing in rice. *Plant Mol Biol* 35:135–144