Tropical Plant Biol.
DOI 10.1007/s12042-012-9103-z

Gene Content Analysis of Sugarcane Public ESTs Reveals Thousands of Missing Coding-Genes and an Unexpected Pool of Grasses Conserved ncRNAs

R. Vicentini · L. E. V. Del Bem · M. A. Van Sluys · F. T. S. Nogueira · M. Vincentz

Received: 12 December 2011 / Accepted: 12 April 2012

Abstract Sugarcane is the most important crop for sugar industry and raw material for bioethanol. Here we present a quantitative analysis of the gene content from publicly available sugarcane ESTs. The current sugarcane EST collection sampled orthologs for ~58 % of the closely-related sorghum proteome, suggesting that more than 10,000 sugarcane coding-genes remain undiscovered. Moreover the existence of more than 2,000 ncRNAs conserved between sugarcane and sorghum was revealed, among which over 500 are also detected in rice, supporting the existence of hundreds of conserved ncRNAs in grasses. New efforts towards sugarcane transcriptome sequencing were needed to sample the missing coding-genes as well as to expand the catalog of ncRNAs.

Keywords · ncRNAs · Orthology · Sorghum · Sugarcane · Transcriptome

Introduction

Sugarcane (*Saccharum* spp. L., Poaceae) is a C4 sucrose-accumulating grass, the most important crop for sugar industry (Lam et al. 2009) and probably the most successful bioenergy raw material nowadays been widely used in the production of bioethanol in Brazil (Moore 1995; Goldemberg 2006). Modern *Saccharum* hybrids are highly polyploid and aneuploid with chromosome numbers in somatic cells ranging from 100 to 130. This complex genome is derived from a few crosses between the sucrose-accumulating *Saccharum officinarum* L. (2n=8x=80) and the disease-resistant but low sucrose content *S. spontaneum* L. (2n=5x to 12x=40–128). Cultivars are vegetatively propagated and result from selection in populations derived from crosses between outcrossing heterozygous parents (Daniels and Roach 1987; Grivet et al. 2004; Garcia et al. 2006). Current sugarcane cultivars are estimated to possess 80–90 % of the genome from *S. officinarum* and 10–20 % from *S. spontaneum* (Grivet et al. 1996; Hoarau et al. 2002; D'Hont 2005; Piperidis et al. 2010). Sugarcane's basic monoploid genome ranges between 760 Mb and 930 Mb depending

R. Vicentini and L. E. V. Del Bem are first authors.

Communicated by: Robert Henry

Electronic supplementary material The online version of this article (doi:10.1007/s12042-012-9103-z) contains supplementary material, which is available to authorized users.

R. Vicentini
Systems Biology Laboratory, Center for Molecular Biology
and Genetic Engineering, State University of Campinas,
Campinas, SP, Brazil

L. E. V. Del Bem · M. Vincentz
Plant Genetics Laboratory, Center for Molecular Biology
and Genetic Engineering, State University of Campinas,
Campinas, SP, Brazil

F. T. S. Nogueira
Department of Genetics, Institute of Biosciences,
São Paulo State University,
Botucatu, SP, Brazil

M. A. Van Sluys
Genomes and Transposable Elements Laboratory, Department
of Botany, Institute of Biosciences, University of São Paulo,
Rua do Matão, 277,
05508-090, São Paulo, Brazil

R. Vicentini (✉) · L. E. V. Del Bem (✉)
Center for Molecular Biology and Genetic Engineering,
State University of Campinas,
Av. Cândido Rondon, 400,
Campinas, SP, Brazil CEP: 13083-875
e-mail: shinapes@unicamp.br

L. E. V. Del Bem
e-mail: lev.del.bem@gmail.com

Published online: 15 May 2012

on the cultivar breeding history, which represents more than twice the size of rice genome (389 Mb) and is close to sorghum (730 Mb) (D'Hont & Glaszmann 2001). Analyses of haplotype organization suggest that despite the elevated ploidy sugarcane's monoploid genome is highly conserved with sorghum in terms of gene retention and colinearity (Jannoo et al. 2007; Garsmeur et al. 2011). This result makes sorghum the most obvious model choice for sugarcane genomics.

In the last ten years, several sugarcane ESTs collections have been developed (Casu et al. 2001; Carson and Botha 2002; Carson et al. 2002; Casu et al. 2003; Vettore et al. 2003; Ma et al. 2004; Bower et al. 2005; Gupta et al. 2010). The publicly available sugarcane ESTs were assembled into tentative consensus sequences referred to as the Sugarcane Gene Index, mainly composed by sequences from the Brazilian sugarcane EST project (SUCEST; Vettore et al. 2003). The SUCEST project generated 237,954 ESTs, which were organized into 43,141 putative unique sugarcane transcripts referred to as Sugarcane-Assembled Sequences (SASs). These ESTs were used to develop molecular markers such as microsatellite (SSR) and single nucleotide polymorphisms (SNPs), which were successfully used to produce linkage maps and identify QTL for important agronomical traits (Oliveira et al. 2007; Pastina et al. 2010; Somerville et al. 2010). Whether this ESTs collection represents the complete set of sugarcane genes is unclear since around 60 % of the SASs present an average two-fold redundancy with *Arabidopsis* proteome (Menossi et al. 2008). Whether this degree of redundancy could be attributed to the high degree of ploidy/aneuploidy found in sugarcane genome still needs to be further investigated. In order to improve the assessment of sugarcane genes in public ESTs we performed a comparative analysis of the sugarcane ESTs against sorghum and rice genomes.

Our approach uses orthology assignment based on high-throughput amino acids maximum-likelihood (ML) phylogenetic analysis, to identify sugarcane's sorghum and rice possible orthologs along with a complementary nucleotide mapping of sugarcane sequences against sorghum chromosomes. This strategy estimated the sugarcane sampled genes as corresponding to only ~58 % of the predicted sorghum proteome, and uncovers the possibility that more than two thousand putative non-coding RNAs (ncRNAs) are conserved between sugarcane and sorghum, been a quarter possibly shared by rice.

Results and Discussion

Sugarcane EST Collections

All sugarcane ESTs were compiled as the Sugarcane Gene Index database SoGI (http://compbio.dfci.harvard.edu/cgi-bin/tgi/gimain.pl?gudb=s_officinarum). The current version (3.0) of SoGI contains 121,342 unique sequences of which only 7,587 singletons and 1,192 tentative consensuses are composed exclusively of ESTs not generated by SUCEST project (~7 %). Although SoGI integrates all published sequences, its clustering strategy produces redundant clusters. This aspect makes the use of a less redundant assemblage strategy, like the one implemented by SUCEST, more appropriate for an orthology-based analysis. A detailed study on sugarcane's *Adh* genes using SNPs (Grivet et al. 2003) suggested that the SUCEST assembly does not merge the paralogous genes into chimeric clusters, which reinforces the reliability of the SASs in a gene-content study where the discrimination between paralogous genes is of critical importance.

In order to obtain a less redundant dataset that includes sequences that were not sampled by SUCEST, we performed a blastn alignment of the 43,141 SASs against the set of 8,779 sequences from SoGI composed exclusively by ESTs not generated by the SUCEST project. This step resulted in a set of 8,106 sequences lacking detectable similarity to SASs (e-value cutoff$<e^{-5}$). We generated a new dataset that integrates the SAS with the latter set leading to 51,247 consolidated clusters (Fig. 1) that will referred to as CSCs (Consolidated Sugarcane Clusters).

Estimation of the Non-Redundant Protein-Coding Gene set in Sugarcane EST Collections

To further contribute with the prediction of the non-redundant set of genes that were sampled by sugarcane ESTs we focused on assigning orthology of the CSC to proteins of sorghum which is the sugarcane closest related grass species whose genome has been sequenced. Since sugarcane and sorghum lineages have diverged recently (5–9 million years ago; Ming et al. 1998; Jannoo et al. 2007) it is reasonable to assume that the monoploid set of genes from sugarcane and sorghum is highly conserved. This assumption is further supported by a recent study that compared sugarcane and sorghum homologous genomic regions (Garsmeur et al. 2011).

To infer orthologous relationships we designed an algorithm that starts with a blastx search of CSC (51,247 sequences) against the sorghum and rice predicted proteomes (Fig. 1a). This initial step produced blastx alignments (e-value cutoff$<e^{-5}$) against sorghum proteome for 38,405 CSC (~75 %). An additional set of 900 CSC (~1.7 %) produced positive hits exclusively against rice proteins. This first step revealed that ~77 % (39,305) of the CSC collection is most likely derived from protein-coding genes (Fig. 1a). The remaining 11,942 CSC were compared by blastx against NCBI's NR database (e-value cutoff$<e^{-5}$) resulting in 2127 positive hits and 9815 no-hits. Among the positive hits 1,826 (~85 %) CSC had a non-Embryophyta (land plants) sequence as best hit. This later set was considered as corresponding to

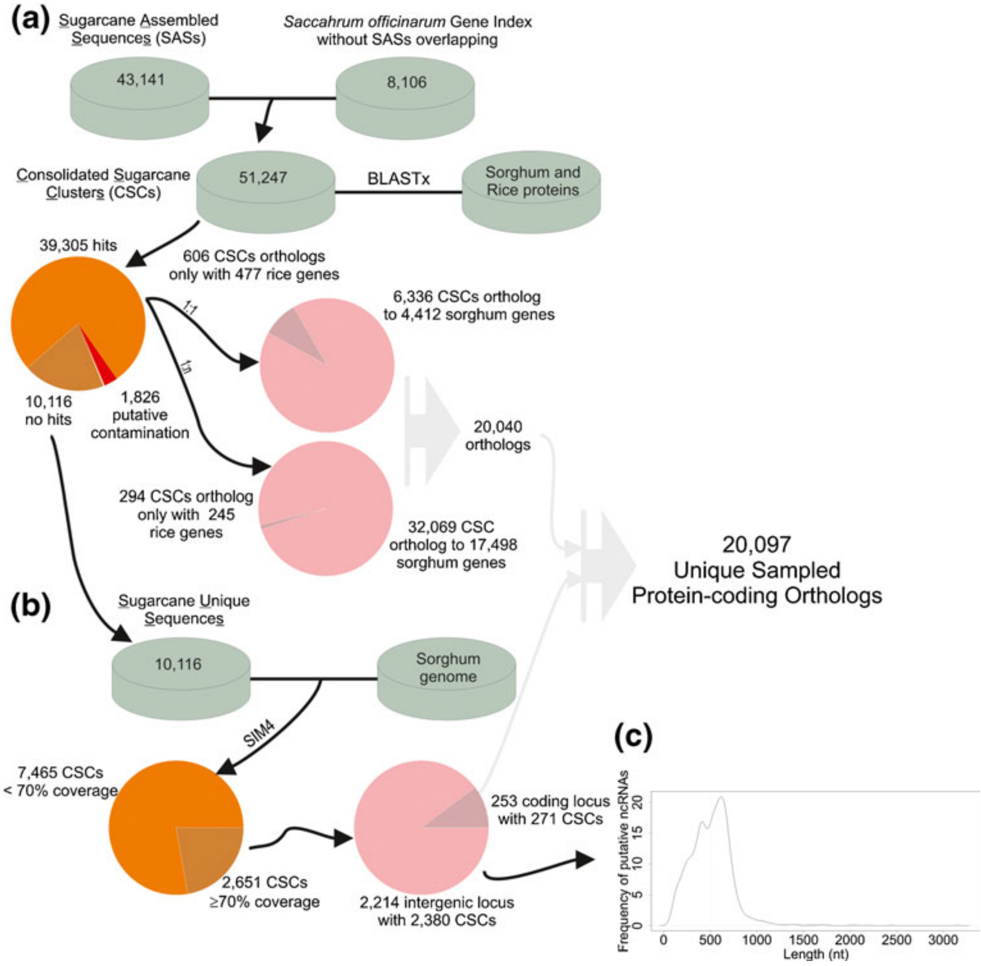
(a)
Sugarcane Assembled Sequences (SASs)
43,141
Saccahrum officinarum Gene Index without SASs overlapping
8,106
Consolidated Sugarcane Clusters (CSCs)
51,247
BLASTx
Sorghum and Rice proteins
39,305 hits
606 CSCs orthologs only with 477 rice genes
1:1
1:n
6,336 CSCs ortholog to 4,412 sorghum genes
10,116 no hits
1,826 putative contamination
20,040 orthologs
294 CSCs ortholog only with 245 rice genes
32,069 CSC ortholog to 17,498 sorghum genes
20,097 Unique Sampled Protein-coding Orthologs
(b)
Sugarcane Unique Sequences
10,116
Sorghum genome
SIM4
7,465 CSCs < 70% coverage
2,651 CSCs ≥70% coverage
253 coding locus with 271 CSCs
2,214 intergenic locus with 2,380 CSCs
(c)
Frequency of putative ncRNAs
0
5
10
15
20
0
500
1000
1500
2000
2500
3000
Length (nt)

Fig. 1 Estimation of the non-redundant gene content of sugarcane ESTs. (**a**) Schematic diagram describing the different filters applied to obtain the list of putative non-redundant sugarcane coding-genes and (**b**) the mapping of sugarcane sequences without detectable protein similarity to the sorghum genome. We identified 20,097 putative non-redundant coding-genes and 2,214 putative non-coding RNAs. (**c**) The sequence size distribution of the ncRNAs shown by frequency plot

non-plant contaminants and removed from subsequent analyses. All the remaining CSC lacking blastx positive hits (10,116; Fig. 1b) were analysed by mapping to sorghum genome and will be further discussed in the next section.

The next step of the algorithm was to separate the CSC that produced a single blastx hit against sorghum or rice proteomes (6,942 or ~17.6 %) from those producing multiple hits (32,363 or 82.4 %). CSC from the first category were directly defined as orthologs to its unique sorghum or rice blastx hit which resulted in the assignment of 6,336 sugarcane CSC to 4,412 sorghum protein-coding orthologs (~43 % redundancy) and 606 sugarcane unigenes lacking sorghum blastx hits which were assigned to 477 rice orthologs (~27 % redundancy) (Fig. 1a).

Each one of the CSC producing multiple blastx hits went through a one-by-one ML phylogenetic analysis along with its 40 first sorghum and rice blastx hits. All phylogenetic analyses were done with amino acid sequences. The CSC were assigned to its closest sorghum or rice ortholog in the resulting phylogenetic trees. Whenever a CSC (4,744 sequences or ~14.6 %) was included in a clade containing multiple sorghum or rice putative paralogs, an additional step was performed in order to define a single ortholog such as to increase the resolution of the analysis (i.e. to limit overestimation). This additional analytic step essentially consisted in producing a distance matrix using WAG plus gamma substitution model among the sequences belonging to such clades and the closest sorghum or rice sequence within the clade was assigned as ortholog to the CSC under analysis. This process allowed us to assign 32,069 CSC to 17,498 sorghum orthologs (~83 % redundancy) and 294 CSC to 245 rice orthologs (~19 % redundancy) (Fig. 1a).

More than half (~53 %) of the CSC that produced a single blastx hit in the first step was assigned to the same sorghum (or rice) orthologs as CSC producing multiple blastx hits. This occurred mainly with CSC containing just a small part of the entire protein that due to our blastx e-value cutoff (e^{-5}) produced just one alignment below the threshold. The final estimation of the coding-gene content of sugarcane public ESTs was obtained by removing the redundancy between the set of orthologs assigned by blastx (single-hit CSC) and the set assigned by ML analysis. Based on the orthology assignment to sorghum and rice, we estimated that the CSC derived from coding-genes (39,305 sequences) sampled 20,040 unique protein-coding orthologs implying ~96 %

of internal redundancy. The total number of sorghum orthologs sampled represents ~58 % of its predicted proteome (34,496 unique coding-genes; http://genome.jgi.doe.gov/Sorbi1/Sorbi1.info.html). This estimation highlights a significant degree of redundancy among the public available sugarcane ESTs collection and points to the necessity of new sequencing efforts.

A Set of Conserved Potential ncRNAs was Revealed by Mapping Sugarcane Unigenes to Sorghum Genome

The 10,116 CSC lacking positive blastx hits against sorghum and rice predicted proteomes were further analyzed by mapping them to sorghum chromosomes using SIM4 (Florea et al. 1998) (Fig. 1b). To limit the number of false positives we applied a filter that recovered the CSC that had a minimum of 70 % of its sequence aligned to the same locus of the sorghum genome (Fig. 1b). Under this criterion 2,651 CSC were retained (7,465 were discarded) and further analysed to define their location relative to the sorghum annotated genes. Among this later set of sequences 271 CSC overlapped with 253 annotated sorghum coding-genes, of which just 56 represented previously unidentified sorghum orthologs (Fig. 1a) and it represents a marginal increment to the first assessment, leading to 20,097 unique sampled protein-coding orthologs (Fig. 1, Table S1). The remaining 2,380 CSC were mapped to 'intergenic' loci within the sorghum genome (Table S2). Any independent CSC overlapping by at least one nucleotide at the same sorghum locus were merged resulting in 2,214 possibly unique ncRNAs loci conserved with sorghum (redundancy of ~7,5 %; Fig. 1b). The size distribution of these non-coding CSCs shows that 54 % of them are longer than 500 pb (Fig. 1c). Furthermore, a blastn search against rice chromosomes (e-value cutoff<e^{-5}) was performed and revealed that 533 out of the 2,214 conserved sugarcane/sorghum ncRNAs (~24 %) were also detected suggesting that at least some of these putative ncRNAs are conserved among grasses and are therefore relevant in grass biology. None of the sugarcane miRNA precursors previously reported (Zanca et al. 2010) were recovered in our analysis due to low sequence alignment with sorghum counterparts.

We found out that ~18 % of the sugarcane/sorghum conserved ncRNA (440 sequences) presented a perfect match with at least one 23-25nt small RNA (sRNA) read from a sugarcane leaf sRNA library (42,218 mapped Illumina® reads out of 2,567,356). When using an arbitrary criterion of at least 15 perfect matched sRNA, only 117 putative sugarcane/sorghum ncRNAs were retained and 63 of them are also detected in rice (Table S2). Whether these putative ncRNAs are the precursors of the perfect matched sRNAs (*cis* action) or they are produced by other loci and act in *trans* remains an open question.

A more detailed analysis of the 13 ncRNA most enriched in perfectly matched sRNAs (i.e., >1,000 sRNAs) revealed a phased distribution of sRNAs (Figure S1). In rice, this kind of pattern was found to derive from miRNAs miR2118- and miR2275-mediated cleavage of a target RNA to define the starting point of the grass-specific Dicer-Like OsDCL3b–mediated production of phased 24-nt sRNAs (Johnson et al. 2009; Song et al. 2011) in a way resembling the biogenesis of the 21-nt *trans*-acting siRNAs (Yoshikawa et al. 2005). The mechanism of biogenesis of the phased 24-nt sRNAs also appears to be conserved in maize (Johnson et al. 2009) and our data suggests it is conserved in sugarcane. The function of these grass-specific 24-nt phased sRNAs is still to be explored.

We compared the whole set of sugarcane putative ncRNAs against the TIGR Plant Repeat Databases (Ouyang and Buell 2004) and only 93 positive hits (~4 %, blastn e-value cutoff < e^{-5}, Table S3) were found. This proportion is near 10 times higher in the sRNA-enriched set of ncRNAs (>15 perfect matched sRNA; 46 out of 117 or ~39 %). Performing the same search for repeats and low complexity sequences using the RepeatMasker software (http://www.repeatmasker.org) we obtained 369 positive hits for the whole set (~15.5 %, Table S4) and 85 among the sRNA-enriched sugarcane ncRNAs (~72 % or ~4.6 times higher). This latter result suggests that the pool of sRNA-enriched ncRNAs is enriched in repetitive and/or transposable element (TE)-derived sequences. Whether the remaining ncRNAs represent new TEs or even Pol IV-transcribed sequences remain to be defined.

Coverage of the Sorghum Exome by Sugarcane ESTs

We have shown in the previous sections that the publicly available sugarcane transcriptome could be linked to 20,097 out of 34,496 (~58 %) sorghum coding-genes. To access the completeness of the sugarcane sequences we mapped all the sugarcane CSC to the sorghum genome using SIM4 and recovered only the best alignment for each CSC. We limited the analysis to the CSC aligned to the sorghum exome that summed 17.654.812 aligned bases. This latter number corresponds to ~40 % of the sorghum exome (48.348.706 nucleotides within 34,496 coding-genes). Normalizing the coverage by the number of sampled sugarcane orthologs (20,097) we found an average coverage of ~63 % relative to the sorghum orthologs.

Conclusions

Our comparative approach leads to the conclusion that the publicly available EST collection for sugarcane accounts with orthologs sampled for at least ~58 % of the predicted sorghum proteome. Significantly, we also found more than two thousand conserved sugarcane/sorghum putative ncRNAs, of which 553 also have some degree of conservation in the rice genome. We were able to show that a subset of these putative ncRNAs has a considerable number of perfect matched 23-25nt sRNAs from a library of sugarcane leaf-expressed sRNAs. Some of these ncRNA may correspond to TEs while the function of most of them remains to be investigated with special attention to their involvement in epigenetic-related processes (Mattick 2001; Mattick 2005; Mattick and Makunin 2006; Mercer et al. 2009; Ben Amor et al. 2009; Matzke et al. 2009; De Lucia and Dean 2011; Zhu and Wang 2012). We also show that the total coverage of the sorghum exome by the sugarcane coding-sequences available up to now is ~40 % and the average coverage of the sampled orthologs is ~63 %. The fact that possibly more than ten thousand sugarcane coding-genes are undiscovered shows the need of new sequencing efforts of sugarcane transcriptome to increase the panel of possible molecular markers and sequence information for sugarcane breeding programs and biotechnological improvement.

Materials and Methods

Public Sequence Datasets

The SASs (Vettore et al. 2003) were obtained from Sugarcane Functional Genomics Database (http://www.sucest-fun.org/) and are available for download, including the option of batch downloading (https://sucest-fun.org/cgi-bin/cane_regnet/sucamet/search_transcript.cgi). The Sugarcane Gene Index sequences were obtained from The Gene Index Project (http://compbio.dfci.harvard.edu/cgi-bin/tgi/gimain.pl?gudb=s_officinarum) and the *Sorghum bicolor* (Paterson et al. 2009) and *Oryza sativa* (Yu et al. 2002) complete genomic sequences were downloaded from Phytozome (http://www.phytozome.net/). *Oryza sativa* and *Sorghum bicolor* protein data sets were obtained from the Rice Genome Annotation Project (version 5.0, http://rice.plantbiology.msu.edu) and DOE JGI's *Sorghum bicolor* (version 1.4, http://genome.jgi-psf.org/Sorbi1/Sorbi1.home.html), respectively.

Consolidated Sugarcane Clusters Phylogenetic Analyses

Blastx searches using the Consolidated Sugarcane Clusters (CSC; see description in Results and Discussion) as queries were performed against the predicted rice and sorghum proteomes, and the CSC coding-sequences were deduced from the amino acids alignment with the blastx best-match (sorghum or rice). Whenever a CSC was aligned in more than one frame with its blastx best hit (i.e. frameshifts; ~20 % of the protein-coding CSC) we recovered the aligned blocks with e-value below e^{-5} to produce a concatenated sequence keeping the same order of the aligned blocks in relation to its best hit. The average coverage over the sorghum counterparts of these concatenated proteins inferred by this method was 76 %.

The translated CSC were then aligned with the 40 first blastx hits from sorghum and/or rice by MAFFT (Katoh et al. 2005) using default parameters. The phylogenetic relationship of the aligned protein sequences was then inferred by ML using PhyML (Guindon et al. 2010) with WAG plus gamma substitution model and aLTR test.

Estimation of the Non-Redundant Coding-Gene set in Consolidated Sugarcane Clusters

The set of phylogenetic trees generated was analysed by a script that searches for the closest sorghum protein sequence to each of the inputted CSC in a given phylogenetic tree to assign ortologous relationships. Similarly, the CSC that only had blastx hits with rice proteins were assigned to the phylogenetically closest rice sequences. We estimated the redundancy among the CSC by merging different CSC that were assigned as orthologs to the same sorghum or rice protein. A marginal proportion of the whole set of sugarcane coding-genes comes from blastx no-hit sequences that were mapped into a sorghum coding-gene as described below.

Consolidated Sugarcane Clusters Sequences Mapping to Sorghum Genome

We used the SIM4 software (Florea et al. 1998) to align the CSC against sorghum genome. Only CSCs without protein similarity and with >70 % of its total length aligned to a single locus were retained. We removed potential redundancies merging different CSC that overlapped (at least one nucleotide) over their best alignment on the sorghum genome.

Small RNA Library Construction and Bioinformatic Analysis

To evaluate the small RNA landscape of putative sugarcane ncRNAs, we analyzed Illumina® sequences from a small RNA library generated from leaves of 1-month old SP80-3280 sugarcane cultivar plants, grown under greenhouse conditions. Ten micrograms of total RNA, prepared using TRizol reagent (Invitrogen®) according to the manufacturer's instructions, were used to generate a sRNA library following Illumina's modified protocol. The sRNA fraction of 19–28 nt was purified by size fractionation on a 15 % TBE–Urea polyacrylamide gel. A 5`-adenylated single-stranded adapter was first ligated to the 3'-end of the sRNAs using T4 RNA ligase without ATP, followed by a second single-stranded adapter ligation at the 5'-end of the RNA using T4 RNA ligase in the presence of ATP. The resulting products were fractioned on a 10 % TBE–Urea polyacrylamide gel and then used for cDNA synthesis and PCR amplification. The resulting library was sequenced on an Illumina® Genome Analyzer (GA-IIx) following the manufacturer's protocol available at http://www.fasteris.com. Raw sequences were retrieved in a FASTQ formatted file and the adapter sequences were removed using Perl® scripts. After trimming of the adapter sequences, we used the software MAQ (http://maq.sourceforge.net) to map 23–25 nt sRNA reads against the CSC representing the set of putative sugarcane ncRNAs. A total of 42,218 high quality raw sequences of 23 to 25 nucleotides shows perfect match against the sugarcane putative ncRNAs, representing ~1.6 % of the whole sRNA library (Supplemental File 1).

Acknowledgments This work was funded by grants 08/58031-0 (RV) and 08/52071-0 (MV) from FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) and LEVDB received a PhD scholarship from FAPESP (2008/09105-1).

References

Ben Amor B, Wirth S, Merchan F, Laporte P, d'Aubenton-Carafa Y, Hirsch J, Maizel A, Mallory A, Lucas A, Deragon JM, Vaucheret H, Thermes C, Crespi M (2009) Novel long non-protein coding RNAs involved in Arabidopsis differentiation and stress responses. Genome Res 19:57–69

Bower NI, Casu RE, Maclean DJ, Reverter A, Chapman SC, Manners JM (2005) Transcriptional response of sugarcane roots tomethyl jasmonate. Plant Sci 168:761–772

Carson D, Botha F (2002) Genes expressed in sugarcane maturing internodal tissue. Plant Cell Rep 20:1075–1081

Carson DL, Huckett BI, Botha FC (2002) Sugarcane ESTs differentially expressed in immature and maturing intermodal tissue. Plant Sci 162:289–300

Casu RE, Dimmock CM, Thomas M, Bower N, Knight D (2001) Genetic and expression profiling in sugarcane. Proc Int Soc Sugar Cane Technol 24:542–546

Casu RE, Grof CPL, Rae AL, McIntyre CL, Dimmock CM, Manners JM (2003) Identification of a novel sugar transporter homologue strongly expressed in maturing stem vascular tissues of sugarcane by expressed sequence tag and microarray analysis. Plant Mol Biol 52:371–386

De Lucia F, Dean C (2011) Long non-coding RNAs and chromatin regulation. Curr Opin Plant Biol 14(2):168–173

D'Hont A (2005) Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. Cytogenet Genome Res 109:27–33

D'Hont A, Glaszmann JC (2001) Sugarcane genome analysis with molecular markers, a first decade of research. Proc Int Soc Sugar Cane Technol 24:556–559

Daniels J, Roach BT (1987) Taxonomy and evolution in sugarcane. In: Heinz D (ed) Sugarcane improvement through breeding. Elsevier Press, Amsterdam, pp 7–84

Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res 8:967–974

Garcia AA, Kido EA, Meza AN, Souza HM, Pinto LR, Pastina MM, Leite CS, Silva JA, Ulian EC, Figueira A et al (2006) Development of an integrated genetic map of a sugarcane (*Saccharum* spp.) commercial cross, based on a maximum-likelihood approach for estimation of linkage and linkage phases. Theor Appl Genet 112:298–314

Garsmeur O, Charron C, Bocs S, Jouffe V, Samain S, Couloux A, Droc G, Zini C, Glaszmann JC, Van Sluys MA et al (2011) High homologous gene conservation despite extreme autopolyploid redundancy in sugarcane. New Phytol 189:629–642

Goldemberg J (2006) The ethanol program in Brazil. Environ Res Lett 1:014008

Grivet L, D'Hont A, Roques D, Feldmann P, Lanaud C, Glaszmann JC (1996) RFLP mapping in cultivated sugarcane (*Saccharum* spp.): genome organization in a highly polyploid and aneuploid interspecific hybrid. Genetics 142:987–1000

Grivet L, Glaszmann JC, Vincentz M, da Silva F, Arruda P (2003) ESTs as a source for sequence polymorphism discovery in sugarcane: example of the *Adh* genes. Theor Appl Genet 106(2):190–197

Grivet L, Daniels C, Glaszmann JC, D'Hont A (2004) A review of recent molecular genetics evidence for sugarcane evolution and domestication. Ethnobot Res Appl 2:9–17

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321

Gupta V, Raghuvanshi S, Gupta A, Saini N, Gaur A, Khan MS, Gupta RS, Singh J, Duttamajumder SK, Srtivastava S et al (2010) The water-deficit stress- and red-rot-related genes in sugarcane. Funct Integr Genomics 10:207–214

Hoarau JY, Grivet L, Offmann B, Raboin LM, Diorflar JP, Payet J, Hellmann M, D'Hont A, Glaszmann JC (2002) Genetic dissection of a modern sugarcane cultivar (*Saccharum* spp.). II. Detection of QTLs for yield components. Theor Appl Genet 105:1027–1037

Jannoo N, Grivet L, Chantret N, Garsmeur O, Glaszmann J-C, Arruda P, D'Hont A (2007) Orthologous comparison in a gene-rich region among grasses reveals stability in the sugarcane polyploid genome. Plant J 50:574–585

Johnson C, Kasprzewska A, Tennessen K, Fernandes J, Nan GL, Walbot V, Sundaresan V, Vance V, Bowman LH (2009) Clusters and superclusters of phased small RNAs in the developing inflorescence of rice. Genome Res 19:1429–1440

Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33:511–518

Lam E, Shine J, da Silva J, Lawton M, Bonos S, Calvino M, Carrer H, Silva-Filho MC, Glynn N, Helsel Z et al (2009) Improving sugarcane for biofuel: engineering for an even better feedstock. Glob Chang Biol Bioenergy 1:251–255

Ma HH, Schulze S, Lee S, Yang M, Mirkov E, Irvine J, Moore P, Paterson A (2004) An EST survey of the sugarcane transcriptome. Theor Appl Genet 108:851–863

Mattick JS (2001) Non-coding RNAs: the architects of eukaryotic complexity. EMBO Rep 2:986–991

Mattick JS (2005) The functional genomics of noncoding RNA. Science 309:1527–1528

Mattick JS, Makunin IV (2006) Non-coding RNA. Hum Mol Genet 15:R17–R29

Matzke M, Kanno T, Daxinger L, Huettel B, Matzke AJM (2009) RNA-mediated chromatin-based silencing in plants. Curr Opin Cell Biol 21(3):367–376

Ming R, Liu SC, Lin YR, da Silva J, Wilson W, Braga D, van Deynze A, Wenslaff TF, Wu KK, Moore PH, Burnquist W, Sorrells ME, Irvine JE, Paterson AH (1998) Detailed alignment of saccharum and sorghum chromosomes: comparative organization of closely related diploid and polyploid genomes. Genetics 150:1663–1682

Menossi M, Silva-Filho MC, Vincentz M, Van-Sluys M, Souza GM (2008) Sugarcane functional genomics: gene discovery for agronomic trait development. Int J Plant Genomics 2008:458732

Mercer TR, Dinger ME, Mattick JS (2009) Long noncoding RNAs: insights into function. Nat Rev Genet 10:155–159

Moore PH (1995) Temporal and spatial regulation of sucrose accumulation in the sugarcane stem. Aust J Plant Physiol 22:661–679

Oliveira KM, Pinto LR, Marconi TG, Margarido GRA, Pastina MM, Teixeira LHM, Figueira AV, Ulian EC, Garcia AAF, Souza AP (2007) Functional integrated genetic linkage map based on EST-markers for a sugarcane (*Saccharum* spp.) commercial cross. Mol Breed 20:189–208

Ouyang S, Buell CR (2004) The TIGR plant repeat databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res 32:D360–D363

Pastina MM, Pinto LR, Oliveira KM, Souza KM, Garcia AAF (2010) Molecular mapping of complex traits. In: Henry (ed) Genetics, genomics and breeding of sugarcane. CRC Press, Science Publishers

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A et al (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556

Piperidis G, Piperidis N, D'Hont A (2010) Molecular cytogenetic investigation of chromosome composition and transmission in sugarcane. Mol Genet Genomics 284:65–73

Somerville C, Youngs H, Taylor C, Davis SC, Long SP (2010) Feedstocks for lignocellulosic biofuels. Science 329:790–792

Song X, Li P, Zhai J, Zhou M, Ma L, Liu B, Jeong DH, Nakano M, Cao S, Liu C, Chu C, Wang XJ, Green PJ, Meyers BC, Cao X (2011) Roles of DCL4 and DCL3b in rice phased small RNA biogenesis. Plant J 69:462–474

Vettore AL, da Silva FR, Kemper EL, Souza GM, da Silva AM, Ferro MIs, Henrique-Silva F, Giglioti EA, Lemos MVF, Coutinho LL et al (2003) Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. Genome Res 13:2725–2735

Yoshikawa M, Peragine A, Park MY, Poethig RS (2005) A pathway for the biogenesis of trans-acting siRNAs in Arabidopsis. Genes Development. 15; 19(18):2164–2175

Yu J et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296:79–92

Zanca AS, Vicentini R, Ortiz-Morea FA, Del Bem LEV, da Silva MJ, Vincentz M, Nogueira FTS (2010) Identification and expression analysis of microRNAs and targets in the biofuel crop sugarcane. BMC Plant Biol 10:260

Zhu QH, Wang MB (2012) Molecular functions of long non-coding RNAs in plants. Genes 3(1):176–190