Plant Mol Biol
DOI 10.1007/s11103-007-9282-4

Expressed sequence tag analysis and development of gene associated markers in a near-isogenic plant system of *Eragrostis curvula*

Gerardo D. L. Cervigni · Norma Paniego · Marina Díaz · Juan P. Selva · Diego Zappacosta · Darío Zanazzi · Iñaki Landerreche · Luciano Martelotto · Silvina Felitti · Silvina Pessino · Germán Spangenberg · Viviana Echenique

Received: 15 February 2007 / Accepted: 22 December 2007

Abstract *Eragrostis curvula* (Schrad.) Nees is a forage grass native to the semiarid regions of Southern Africa, which reproduces mainly by pseudogamous diplosporous apomixis. A collection of ESTs was generated from four cDNA libraries, three of them obtained from panicles of near-isogenic lines with different ploidy levels and reproductive modes, and one obtained from 12 days-old plant leaves. A total of 12,295 high-quality ESTs were clustered and assembled, rendering 8,864 unigenes, including 1,490 contigs and 7,394 singletons, with a genome coverage of 22%. A total of 7,029 (79.11%) unigenes were functionally categorized by BLASTX analysis against sequences deposited in public databases, but only 37.80% could be classified according to Gene Ontology. Sequence comparison against the cereals genes indexes (GI) revealed 50% significant hits. A total of 254 EST-SSRs were detected from 219 singletons and 35 from contigs. Di- and tri- motifs were similarly represented with percentages of 38.95 and 40.16%, respectively. In addition, 190 SNPs and Indels were detected in 18 contigs generated from 3 to 4 libraries. The ESTs and the molecular markers obtained in this study will provide valuable resources for a wide range of applications including gene identification, genetic mapping, cultivar identification, analysis of genetic diversity, phenotype mapping and marker assisted selection.

Keywords cDNA libraries · *Eragrostis curvula* · ESTs · Gene index · Genic markers

Abbreviations

EST	Expressed sequence tags
2,4-D	2,4 Dichlorophenoxyacetic acid
BAP	6-Benzylamino purine
DMSO	Dimethyl sulfoxide
IPTG	Isopropyl-beta-D-thiogalactopyranoside

Gerardo D. L. Cervigni and Norma Paniego contributed equally to this manuscript.

G. D. L. Cervigni · M. Díaz · J. P. Selva · D. Zappacosta · V. Echenique
Centro de Recursos Naturales Renovables de la Zona Semiárida (CERZOS)–CONICET, Camino de La Carrindanga Km 7.0, Bahia Blanca 8000, Argentina

N. Paniego · V. Echenique
Instituto de Biotecnología, CICVyA, INTA-Castelar, Buenos Aires, Argentina

M. Díaz
Departamento de Biología, Bioquímica y Farmacia, Universidad Nacional del Sur, San Juan 670, Bahia Blanca 8000, Argentina

J. P. Selva · D. Zappacosta · V. Echenique (✉)
Departamento de Agronomía, Universidad Nacional del Sur, CERZOS–CONICET, San Andrés 800, Bahia Blanca 8000, Argentina
e-mail: echeniq@criba.edu.ar

D. Zanazzi · I. Landerreche
Bioaxioma, Venezuela 110-1P, Buenos Aires C1095AAD, Argentina

D. Zanazzi · I. Landerreche
Facultad de Ingeniería, Universidad de Buenos Aires, Av. Paseo Colón 850, Buenos Aires C1063ACV, Argentina

L. Martelotto · S. Felitti · S. Pessino
Facultad de Ciencias Agrarias, Universidad Nacional de Rosario, Zavalla S2125ZAA, Argentina

G. Spangenberg
Primary Industries Research Victoria, Plant Biotechnology Centre, La Trobe University, Bundoora, VIC 3086, Australia

Introduction

For the past two decades, random sequencing of cDNA has provided a simple and efficient method for the identification of many genes from different organisms (Adams et al. 1991). Large-scale EST/cDNA discovery and analysis are widely used for the study of gene expression and the identification of candidate genes for biological processes and phenotypic analysis (Hatey et al. 1998) and also provide the means to study processes of gene evolution (Van der Hoevent et al. 2002). ESTs are also a valuable source of highly polymorphic genetic markers, like simple sequence repeats (SSRs) (Lazo et al. 2004; Thiel et al. 2003; Eujayl et al. 2002; Gao et al. 2004) and single nucleotide polymorphisms (SNPs) (Garg et al. 1999; Yu et al. 2006). These markers are useful for genetic mapping, in which genes can be associated with variation for multiple pleiotropic traits (Ma et al. 2002; Rostoks et al. 2005; Xu et al. 2004; Ravel et al. 2005). SSRs, also known as microsatellites, have been shown to be one of the most powerful of genetic marker systems in biology. They are common, readily identified DNA features consisting of short (1–6 bp), tandemly repeated sequences, widely and ubiquitously distributed throughout eukaryotic genomes (Tóth et al. 2000) and have been found in all prokaryotic and eukaryotic genomes that have so far been analyzed (Katti et al. 2001). SSRs are highly polymorphic, due to the mutation affecting the number of repeat units. This hypervariability among related organisms makes them highly informative for a wide range of applications including high-density genetic mapping, molecular tagging of genes, genotype identification, analysis of genetic diversity, paternity exclusion, phenotype mapping and marker-assisted selection (MAS) of crop plants (Tautz 1989; Powell et al. 1996).

Very little basic information is available about weeping lovegrass [*Eragrostis curvula* (Schrad.) Nees]. It is a member of the family *Poaceae*, subfamily *Chloridoideae* (Wartson and Dallwitz 1992), native to Southern Africa and an important cultivated forage resource for semiarid regions (Covas and Cairnie 1991). Understanding the origin of crop plants is important because it can provide valuable information for plant breeders. *Eragrostis pilosa* is considered to be the immediate progenitor to *Eragrostis tef* (Ingram and Doyle 2003) but wild ancestors of *E. curvula* were not identified. Size genome of *E. curvula* was not reported but the content of DNA per haploid nucleus was 0.35 pg for the genus *Eragrostis* (Bennett and Smith 1976). Weeping lovegrass has small chromosomes, ranging from 2 to 3 μm, with a variable ploidy level (Poverene and Curvetto 1991). Related to the auto- or allopolyploid origin, Poverene et al. (1986), based on the meiotic behavior of 12 *Eragrostis curvula* cultivars stated that cv. Tanganyika shows a high frequency of multivalents. Based on this observations these authors concluded that it is autopolyploid.

The most useful cultivars as forage are tetraploid ($2n = 4x = 40$) and reproduce by pseudogamous diplosporous apomixis (Voigt et al. 2004). This mode of asexual reproduction by seeds is typical of some grass species, where a non-reduced megaspore originates directly from the reproductive cell itself following a lack or a failure of meiosis. All progeny of a given plant genotype, consequently constitutes a maternally derived clone.

In previous work, a series of related *Eragrostis curvula* lines with different ploidy levels and reproductive modes was obtained through in vitro culture of inflorescences followed by colchicine treatment (Cardone et al. 2006; Mecchia et al. 2007). The series consists of a natural apomictic tetraploid cultivar Tanganyika (T) ($2n = 4x = 40$), a diploid sexual line (cv. Victoria = D) ($2n = 2x = 20$) obtained from T after a tissue culture procedure, and a tetraploid sexual strain (cv. Bahiense = M) originated from D by colchicine treatment. These plants provide a suitable system for the identification of gene(s) involved in diplosporous apomixis, ploidy level-regulated expression and forage quality candidate variation using a transcriptomic approach, as well as the development of potential markers for genetic mapping. The aim of this work was to construct, sequence and analyse four cDNA libraries of *Eragrostis curvula* constructed from inflorescences and leaves of the near-isogenic plant series. Categorization of *E. curvula* genes, the first gene index for the species and the development of a group of SSRs and SNPs/indels molecular markers useful to construct genetic maps is described here. In a parallel (see this issue) study a comparison between the ESTs profiles in order to identify genes associated to apomixis and ploidy level gene expression control was conducted.

Materials and methods

Plant material

Plants were obtained as previously reported (Echenique et al. 1996; Cardone et al. 2006). Briefly, inflorescences from plants of the apomictic cultivar Tanganyika (T) ($2n = 4x = 40$) just emerging from the flag leaf were cultured on Murashige and Skoog's (1962) medium supplemented with 2,4-dichlorofenoxiacetic acid (2,4-D) and 6-bencilaminopurine (BAP). One of 23 R_0 plants—first generation of plants obtained from in vitro culture—had half of the normal chromosome number ($2n = 2x = 20$) (cv. Victoria = D). In order to duplicate the chromosome number, 500 seeds of one diploid R_1 plant were treated with a solution of 0.05% colchicine and 2% dimethyl sulfoxide (DMSO). A plant with 40 chromosomes (2n = 4*x* = 40) and sexual reproduction was obtained (UNST1131- cv. Bahiense = M) (Cardone et al. 2006).

cDNA library construction and sequencing

Four cDNA libraries were constructed: three were derived from panicles at the same developmental stage; one from T (2n = 4*x* = 40, apomictic, EC02); one from plant M (2n = 4*x* = 40, sexual, EC04); and one from plant D (2n = 4*x* = 20, sexual, EC01). The fourth library was constructed using leaves of 12 days old plants (2n = 4*x* = 40, apomictic, EC03). The appropriate precautions were taken to minimize the risk of false positive detection: all plants were located together, treated under identical conditions and samples were all collected at the same time of the day (4 pm). Total RNA was extracted from inflorescences or leaves using the RNeasy total RNA isolation kit (Promega). cDNAs were obtained using the SMART PCR synthesis kit (Clontech) which allows maintenance of original transcript levels, cloned into the pGEM-T easy vector (Promega) and used for transformation of XL10-Gold Ultracompetent *E. coli* cells (Stratagene). The cDNA libraries were amplified by adding 4 ml of NZY media and incubated for 3 h at 37°C, with a rotation of 250 rpm. Subsequently, 5 ml of 50% glycerol were added, mixed, aliquoted (200 µl/tube) and stored at −80°C. For titering, 50, 75 and 100 µl of each amplified cDNA library were plated onto LB media containing ampicillin (10 mg/µl). IPTG (isopropyl-beta-D-thiogalactopyranoside) (0.5 mmol/l) and X-Gal (5-bromo-4-chloro-3-indolyl-beta-D-galactopyranoside) (50 mg/l). Plates were incubated overnight at 37°C and the titer was calculated by measurement of number of white colonies (which contain the processed insert).

To isolate the plasmids for sequencing, bacterial cultures were performed in 96-well blocks using LB media supplemented with ampicillin (10 mg/µl). DNA isolation was performed by using a QIAprep 96 turbo Miniprep kit Qiagen Biorobot 9600. Twelve plasmids from each library were chosen at random in order to estimate the average size of the inserts: 2 µl of plasmid DNA were digested with 0.5 µl of *Eco*-RI (10 U/µl) during 30 min at 37°C. Fragments were run in 1% (w/v) agarose gels and visualized with ethidium bromide (0.5 µg ml^{-1}). A total of 12,600 randomly selected clones from the libraries were sequenced from the 3′ end using a MegaBACE DNA sequence analyzer performing an average read of 300–500 bp from each clone.

For EST sequencing 300–500 bp of the inserts were amplified using the SMART primer (5′-TAA CAA CGC AGA GTA CGC GG-3′) (BD Bioscience) using a Big Dye Terminator kit (Applied Biosystems) with Ampli-Taq FS DNA polymerase (Applied Biosystems). Each insert was characterized by a single read using a MegaBACE 4000 sequencer. Sequences were deposited in Gen Bank (Nos: EH183417 to EH195711).

ESTs analysis and annotation

EST sequences were cleaned from vector and polyA tail using Seqclean software (TGIR, the Institute for Genomic Research, Rockville, MD, USA). Sequences containing more than 100 bases with Phred quality 20 (Ewing et al. 1998) were stored in fasta format, together with the corresponding quality file Phred. Sequences were assembled using CAP3 software (Huang and Madan 1999) with a high stringency criteria (sequence identity P = 90) and overlapping lengths (O = 40). The actual number of unigenes and the redundancy level for each cDNA library, both independently and considering the ESTs of the four libraries combined were estimated. This process was carried out as follow: (i) the gene cluster profile for each library and the combined one were obtained using the ESTstat program (Wang et al. 2004) adjusted for partial 5′cDNA libraries and Type III error correction by clustering and (ii) the actual number of unigenes and the redundancy level were estimated with the ESTstatCF program (Wang et al. 2005). The number of new genes to be captured and the redundancy level simulating the addition of a proportional sample of ESTs (S: 0.5, 1.0, 1.5 and 2.0–referred to the initial number of ESTs) to each one and all the four combined libraries was also calculated.

All unique sequences were queried against the SwissProt/Trembl databases using the BLASTX algorithm (Altschul et al. 1994) with an *E*-value $\leq 1 \times 10^{-5}$. Annotations were based on the Gene Ontology (GO) (http://www.geneontology.org/) terms using the program Blast2GO (Conesa et al. 2005). Sequences comparisons against Plant Gene Index (GI) such as: wheat (*Triticum aestivum*–TaGI), maize (*Zea mays*—ZmGI), oat (*Avena sativa*—OGI), barley (*Hordem vulgare*—HvGI) and *Arabidopsis thaliana* (AtGI) from TIGR (http://compbio.dfci.harvard.edu/tgi/tgipage.html) were performed locally using the BLASTN algorithm (Altschul et al. 1994).

Development and detection of genic molecular markers

EST-derived simple sequence repeats (EST-SSRs)

Consensus sequences taken from the total number of high quality ESTs (12,295) from all the four libraries) were used to develop EST-SSR loci. The SSR Discovery program (Robinson et al. 2004) was used for the identification of SSR for design of the primers. For EST-SSR selection a minimum of six repeats for di-, 5 for tri- and 4 for tetra- and pentanucleotide arrays were considered. Sixty primers were synthesized (AlphaDNA, Canada) in order to determine amplification efficiency in *E. curvula* cv. Tanganyika. Primer design criteria were selected as follows to obtain amplification products of 200–500 bp, annealing temperature 56–62°C, CG content of 40–60%, 18–25 bp. Genomic DNA was extracted from leaves of different cultivars of *E. curvula* (DW: Don Walter, DJ: Don Juan, DP: Don Pablo, DE: Don Eduardo, K: Kromdraii, DA: Don Arturo, M: Morpa, E: Ermelo, T: Tanganyika, V: Victoria, and B:Bahiense) and from line UNST1112 using the protocol described at the CIMMYT Molecular Genetics Applied Laboratory protocols, CIMMYT, Mexico (http://www.cimmyt.org). PCR conditions were as follows: a final volume of 20 µl containing 50 ng of genomic DNA, 1 µmol/l of each primer, 2.5 mmol/l $MgCl_2$, 0.125 mmol/l of each dNTP, 1X reaction buffer and 2.5 U *Taq* polymerase (Invitrogen). A touchdown program consisting in a denaturation step of 4 min at 94°C, 15 cycles of 30 s at 94°C, 30 s at 65°C (−1°C/cycle) and 1 min at 72°C and 30 cycles of 30 s at 94°C, 30 s at 50°C and 1 min at 72°C was used. The final extension step was of 5 min at 72°C. PCRs were performed in a MJ Research thermocycler. Samples were mixed with denaturing loading buffer, treated for 5 min at 95°C and separated in a 6% polyacrylamide gel. Amplification products were silver-stained following the DNA silver staining system procedure (Promega).

EST-derived single nucleotide polymorphisms (EST-SNPs)

SNPs were identified in the ESTs using the SNP Discovery program (Barker et al. 2003), where for each candidate SNP two measures of confidence are calculated, the redundancy of the polymorphism at a SNP locus and the co segregation of the candidate SNP with other SNPs in the alignment. ESTs from each of the four libraries were aligned and SNPs with a minimum redundancy of 10 and a co-segregation of five were considered.

Results and discussion

Characterization of cDNA libraries

Table 1 provides a description of the main characteristics of each library, including the RNA source, titer, the average size of the inserts and the average EST read length. The estimated insert size for all four libraries was similar and ranged from 500 to 3,500 bp and the insert average read length per library fall into 354 to 544 bp. The partial insert read accomplished allowed a fast and comprehensive sampling of the transcripts present in the different libraries at a reasonable cost, nevertheless, the SMART technology used here yield larger inserts. This kit provides cDNA libraries enriched of full-length transcripts when high-quality starting mRNA is used. The representation of full coding clones in the libraries cannot be informed because it is out of the scope of this study. A similar strategy for library construction and EST analysis was used by other authors like Sawbridge et al. (2003a, b) and Ji et al. (2006). The accessibility of large insert cDNA collection will allow in the near future the complete characterization of those transcripts differentially expressed in relation to ploidy level or reproductive mode in *E. curvula*. In order to evaluate the differential expression among developmental stages, the representative cDNA libraries were built avoiding any step of normalization and/or enrichment.

EST assembly and establishment of unigene sets

After cleaning of low-quality and vector sequences, 12,295 ESTs were retained from the original 12,425 that had been generated from the four cDNA libraries. Those sequences included 3,650 from library EC01; 3,644 from library EC02; 1,609 from library EC03 and 3,392 from library EC04. After clustering and assembly, 8,884 unigenes were obtained, including 1,490 contigs and 7,394 singletons (Table 2). The proportion of singletons, calculated in relation to the total number of ESTs in each library, was similar for the four libraries: 66.04% for library EC01 (2,427 ESTs), 74.09% for EC02 (2,700 ESTs), 65.69% for EC03 (1,057 ESTs), 63.92% for EC04 (2,157) and 60.01% (7,394) when the libraries were combined. A gene cluster

Table 1 Source materials and characteristics of the *Eragrostis curvula* cDNA libraries

Library designation	Tissue	Titer (colonies with insert/ml)	Average insert size (bp)	Average EST length (bp)
EC01	Inflorescence	7,635	1,568	480
EC02	Inflorescence	8,720	1,922	544
EC03	Leaf	13,680	1,562	456
EC04	Inflorescence	7,152	1,427	354

Table 2 Features of the *Eragrostis curvula* cDNA libraries: number of ESTs, singletons, contigs, unigenes, and redundancy level

	cDNA library				
	EC01	EC02	EC03	EC04	Combined libraries
ESTs	3,650	3,644	1,609	3,392	12,295
Unigenes	2,954	3,053	1,188	2,590	8,884
Singletons	2,427	2,700	1,057	2,157	7,394
Contigs	527	353	131	433	1,490
ESTs in Contigs	1,248	944	552	1,235	4,926
Redundancy	1.25	1.19	1.34	1.34	1.38

Redundancy level: estimated according Wang et al. (2005)

profile of each library and a general one, taking into account all the sequences from *Eragrostis curvula* was obtained using the ESTstat program (Wang et al. 2004). The number of ESTs/contig ranged from 2–19 for library EC01, 2–17 for EC02 and EC04 and 2–43 for EC03. The main features of the libraries, such as number of ESTs, unigenes, singletons, contigs, ESTs in contigs and redundancy level, are shown in Table 2.

The eight most prevalent unigenes were constituted from 14 to 46 ESTs. Among them, the gene for the RuBisCO enzyme (46 ESTs and 1,678 bp length), involved in CO_2 fixation; enzymes from the energetic metabolism: cytosolic glyceraldehyde-3-phosphate dehydrogenase (23 ESTs and 1,298 bp length), β-galactosidase (15 ESTs and 564 bp length), ATP synthase (14 ESTs and 1,417 bp length); genes associated to protein synthesis: elongation factor 1-alpha (18 ESTs and 1,582 bp length) and 60S ribosomal protein (22 ESTs and 1,353 bp length) and two genes involved in cell division, like tubulin β-2 chain (25 ESTs and 1,354 bp length) and the checkpoint protein 1 (CHK1) (30 ESTs and 1,066 bp length). Several of the above-mentioned genes are typically expressed in tissues and organs undergoing active division and growth, such as young leaves and inflorescences.

In order to avoid extra sequencing efforts, an important factor to estimate is the number of new genes that can be discovered considering the level of redundancy that would be reached after adding new sets of ESTs. This is an important consideration to guide the selection of the number of clones to be sampled from various cDNA libraries in order to maximize the rate of gene discovery. The redundancy level, estimated according to Wang et al. (2005), was low for each library, indicating that new sequencing runs can be conducted and new genes will be captured (Table 2). As was expected from the singleton ratios, the lowest value was obtained for library EC02. If a typical plant genome contains 30–40,000 genes or less (Bennetzen et al. 2004), each library and the pooled 8,884 unigenes would represent 5–10 and 22% of the *Eragrostis curvula* genome, respectively. Table 3 shows the putative number of new genes that can be discovered taking into account the addition of a proportional set of ESTs. Doubling the EST number (2S) would allow the identification of 3,735 new genes from library EC01, 4,151 from library EC02, 1,705 from library EC03 and up to 3,657 from library EC04. In all these cases the level of redundancy will be similar to the actual one. This means that, independently of the actual condition of each library, it will be possible to identify up to 12,243 new genes from the total number of ESTs, reaching a total of 21,127 genes captured and an *Eragrostis curvula* genome coverage of approximately 53%, while still maintaining a low redundancy level.

Functional annotation of *Eragrostis* unigenes

Eragrostis unigenes were searched by BLASTX for homology analysis against the protein sequences deposited in public databases like SW/Trembl. These unigenes were further annotated on the basis of existing annotation for the proteome of other species, in which functions were categorized according to the Gene Ontology Consortium. During the annotation, when multiple hits were found, the one with the lowest *E*-value was selected. From 8,884 unigenes whose products showed homology (E-value $\leq 1 \times 10^{-5}$) to sequences present in the public databases mentioned 3,358 (37.80%) significantly matched with categorized proteins and putative functions were assigned (Fig. 1). Twenty-two percent of the classified genes were categorized in relation to the molecular function, 23% related to the biological process and 21% to the cellular component. The largest proportion of the functionally categorized unigenes felt into six categories: catalytic activity, nucleotide binding, protein metabolism, binding, transport and energy pathways, with a high nuclear activity, typical from young tissues (Fig. 1). The biological function of other 3,671 (41.32%) unigenes was assigned taking into account the SW/Trembl databases, but it was not possible to categorized them through GO (Table 4). Combining both annotations a total 7,029 (79.11%) unigenes were functionally categorized. However, 1,855 (20.88%) unigenes gave no significant matches in the databases mentioned above.

Table 3 Putative number of new *Eragrostis curvula* genes that can be discovered taking into account the addition of a proportional number of ESTs

	Libraries									
S	EC01 (3,675)[a]		EC02 (3,644)		EC03 (1,609)		EC04 (3,392)		Combined libraries (12,320)	
	NGC	Rd.	NGC	Rd	NGC	Rd	NGC	Rd	NGC	Rd
0.5	1,124[b]	1.36[c]	1,258	1.26	497	1.41	1,016	1.41	3,475	1.49
	(1,086–1,156)	(1.35–1.37)	(1,251–1,341)	(1.25–1.27)	(495–516)	(1.40–1.43)	(977–1,048)	(1.39–1.42)	(3,390–3,524)	(1.49–1.51)
1.0	2,099	1.46	2,452	1.32	953	1.48	1,944	1.49	6,628	1.59
	(2,016–2,169)	(1.44–1.49)	(2,381–2,525)	(1.31–1.34)	(901–993)	(1.46–1.52)	(1,860–2,013)	(1.47–1.52)	(6,421–6,728)	(1.58–1.61)
1.5	2,951	1.56	3,527	1.38	1,370	1.55	2,804	1.57	9,536	1.67
	(2,817–3,079)	(1.53–1.60)	(3,400–3,648)	(1.36–1.41)	(1,287–1,436)	(1.51–1.60)	(2,656–2,910)	(1.54–1.61)	(9,205–9,693)	(1.66–1.70)
2.0	3,699	1.66	4,519	1.44	1,752	1.62	3,607	1.64	12,243	1.75
	(3,510–3,895)	(1.62–1.71)	(4,315–4,700)	(1.41–1.48)	(1,638–1,851)	(1.57–1.69)	(3,381–3,751)	(1.60–1.70)	(11,771–12,446)	(1.73–1.79)

S, Proportion of ESTs to be added related to the total number of EST/library; NGC, new captured genes (genes to be discovered); Rd, redundancy level; [a]total number of EST; [b,c]95% bootstrap confidence interval

In order to investigate gene conservation the *Eragrostis* unigenes were compared with the available genes indexes (GI) from maize (*Zea mays*), wheat (*Triticum aestivum*), oat (*Avena sativa*), barley (*Hordeum vulgare)* and *Arabidposis thaliana* using the BLASTN algorithm (Table 5). Forty-three percent of the *Eragrostis* genes matched sequences from wheat (391,939 annotated genes) and barley (50,423 annotated genes), 49.25% from oat (89,147 annotated genes) and 46.48% from maize (58,582 annotated genes). Despite the very different number of annotated genes available for these grasses, the percentage of *E. curvula* genes present in other plant species was very close to 50%, reaching the highest value when the search was made against the GI of maize, which is taxonomically more related to *Eragrostis* than the other grass species (Yeoh and Watson 1986). However, only 22.32% of the *E. curvula* genes revealed some level of sequence similarity with those from *Arabidopsis*, an anticipated result considering the close taxonomic distance between these species. As approximately 50% of the *E. curvula* genes were represented in the gene index from cereals, the other 28% may represent unique genes from *E. curvula* or closely related species, making them candidates for traits not present in the mentioned species. It is interesting to note that gene representation is independent of the size of the cereal GI used. The lack of correlation between the number of genes in the GI and the matches with *Eragrostis* sequences could be due to different factors such as: (1) the number of genes in the GI could be overestimated. According to Bennetzen et al. (2004), the total number of genes in monocots genomes could be overestimated because of the presence of retrotransposons, (2) unknown function from *Eragrostis* genes and (3) some sequences could correspond to non-functional transcripts.

Molecular markers development based on *Eragrostis* ESTs

Molecular markers provide powerful methods for differentiation of genotypes. A priority goal was to develop molecular markers based on functional coding sequences to be used for future mapping and taxonomic analysis. Combining the ESTs from our four cDNA libraries from *E. curvula* we identified two types of genic molecular markers: EST-SSRs and SNP/indels.

ESTs-SSR were identified over the unigenes taking into account minimal repeats of 6 for di-, 5 for tri- and 4 for tetra- and pentanucleotides. 254 EST-SSRs were detected from 219 singletons and 35 contigs. Unlike other grass species including *Eragrostis tef* (Yu et al. 2006), in *E. curvula* the di- and tri- motifs were equally represented in the 254 EST-SSR, in a percentage of 38.95 and 40.16%

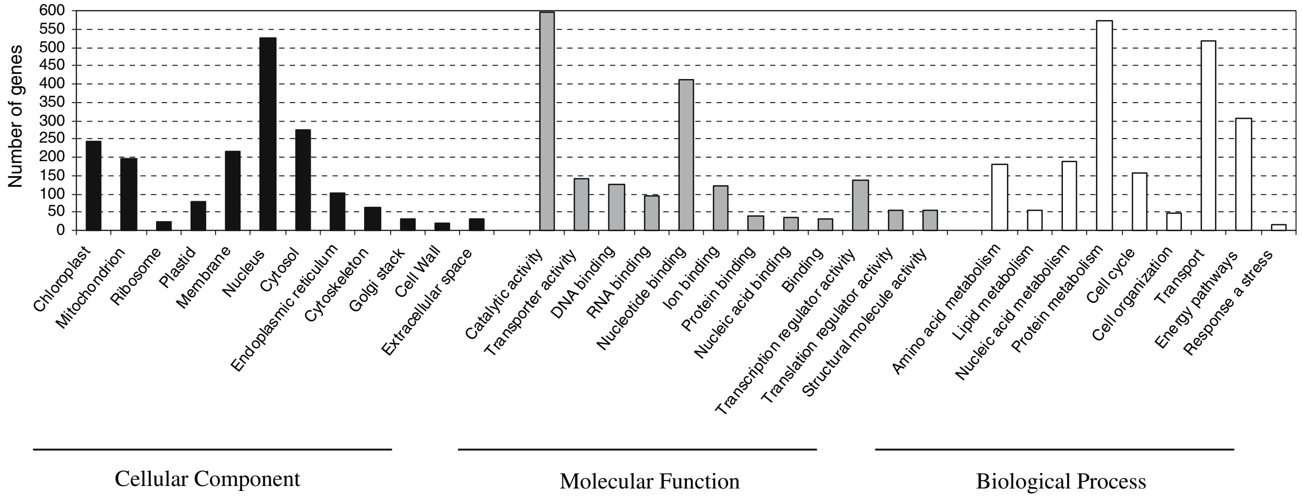

Fig. 1 Functional classification of weeping lovegrass unigenes according to Gene Ontology, *E*-value $\leq 1 \times 10^{-5}$ as the cut-off threshold

Table 4 *Eragrostis curvula* unigenes annotated against TIGR Gene Index of wheat, oat, maize, barley and *Arabidopsis*

Gene Index (GI)	GI size	*E. curvula* annotated genes
Wheat	391,939	3,823 (43.03%)
Oat	89,147	4,508 (49.25%)
Maize	58,582	4,130 (46.48%)
Barley	50,453	3,846 (43.21%)
Arabidopsis	81,826	1,983 (22.32%)

Table 5 Singletons and contigs with positive hits against SW or Trembl, not included in GO classification

Databases			
Unigenes	SW	Trembl	Total[a]
Singlets	171	2,504	2,675
contigs	643	353	996
Total[b]	814	2,837	3,671

[a] Singletons and contigs annotated against SW/Trembl

[b] Total number of unigenes

respectively. The tetra- and penta-nucleotides showed significant lower values (12.04 and 8.33%, respectively). However, the proportions of SSR-containing EST as well as the frequency of in silico-detected SSRs even for the same species has been reported as variable in the literature (Peng and Lapitan 2005).

As was found for other species (Nicot et al. 2004), the most abundant dinucleotide motifs were $(GA)_n$ and $(AC)_n$, representing 63.37 and 27.72% of the total, respectively. As in the case of wheat (Nicot et al. 2004), the GC motif was poorly represented (1.98%). Although $(AT)_n$ microsatellites are thought to be very abundant in the genomic sequences of plants (Lagercrantz et al. 1993; Morgante and Olivieri 1993), this was not the case for *Eragrostis* ESTs (4.95%) . Similar results were found by Nicot et al. (2004) in wheat. These authors suggested that the self-complementation existing for $(AT)_n$ microsatellites makes them difficult to isolate. Kantety et al. (2002) suggested that the high occurrence of $(GA)_n$ motifs is related to the potential for translation into polyAla or polyLeu, depending on the reading frame, which are very frequent aminoacid motifs in proteins (Lewin 1994).

Among the trinucleotide combinations, the predominant types were $(CCG)_n$ and $(CCT)_n$, representing the 25,71 and 23,81% of the total microsatellites, respectively. In wheat and rice these motifs represent 83 and 79% of the total (Kantety et al. 2002). La Rota et al. (2005) showed that in wheat, barley, and rice the di- tri-, tetra- and penta-motifs vary in relation to the length of the repeat. In this study, only in rice the di- and tri-nucleotides ratio were similar, but in those cases in which the length of the sequence was higher than 30 bp. In our study only 5 di- and 1 tri- motif EST-SSR were longer than 30 bp. This difference with our results may depend on the number of unigenes detected in both cases, 8,884 in *E curvula* and the complete gene index of rice.

Primers were designed for the 254 EST-SSRs found in the 8,884 unigenes from *E. curvula.* Sixty were synthesized and amplified for validation using template DNA from cultivar Tanganyika. From these, 58% (35) were effectively amplified under our experimental conditions. As was mentioned by Nicot et al. (2004) the lack of amplification could be due to several reasons: (1) one of the primers designed from EST sequences could overlap two exons; (2) the amplicon may contain a long intron producing a PCR product that could not be visualized on the electrophoretic profile; (3) sequence errors or problems during primer synthesis could occur; (4) as consensus sequences obtained

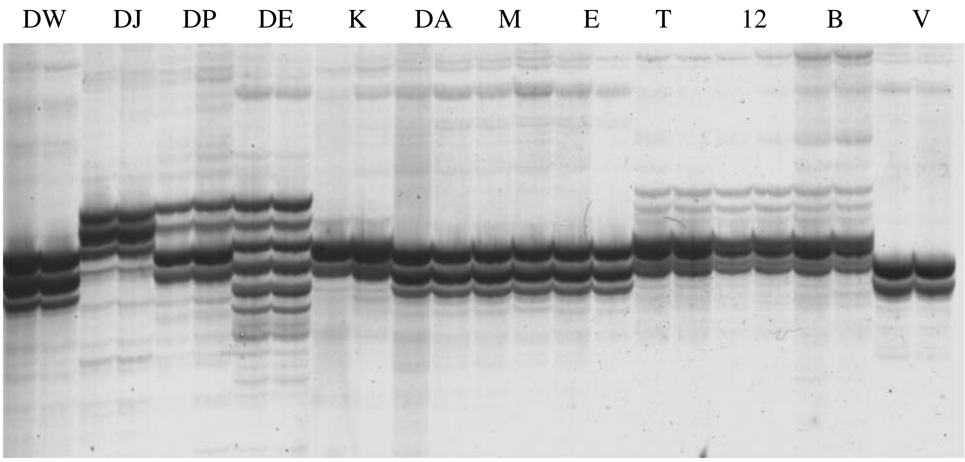

Fig. 2 EST-SSR amplification (by duplicated) in different weeping lovegrass cultivars using one of the designed primer: DW, Don Walter; DJ, Don Juan; DP, Don Pablo; DE, Don Eduardo; K, Kromdraii; DA, Don Arturo; M, Morpa; E, Ermelo; T, Tanganyika; V, Victoria; 12, UNST1112 and B, Bahiense

from the compilation of several ESTs were used, some of them may result from the addition of different copies of the same gene (homoeologs or paralogs). Fifty two *per cent* (13) of the selected primer pairs were polymorphic for at least one of the 12 cultivars tested. Figure 2 shows one of the polymorphic primers used in this work.

The EST-SNP/Indel identification was performed using the sequences from all the four libraries. Alignment of 10 or more ESTs and a minimum co-segregation of five were analyzed with this objective. A total of 190 SNP/Indel were detected in 18 contigs from 3 to 4 libraries.

Conclusions and perspectives

Four cDNA libraries containing full length cDNAs were constructed and 12,295 high quality ESTs sequences were obtained and analyzed. These ESTs sequences were assembled into 8,884 unigenes (1,490 contigs and 7,394 singletons). From these unigenes, 7,029 (79.11%) were functionally categorized and deposited in public databases, but only 37.80% could be classified according to Gene Ontology. Twenty-two percent of the classified genes were categorized in relation to the molecular function, 23% related to the biological process and 21% to the cellular component. The largest proportion of the functionally categorized unigenes felt into six categories: catalytic activity, nucleotide binding, protein metabolism, binding, transport and energy pathways, with a high nuclear activity. Forty-three percent of the *Eragrostis* unigenes matched sequences from wheat and barley, 49.25% from oat and 46.48% from maize.

These cDNA libraries will provide important tools for gene identification and molecular breeding in *Eragrostis curvula*. EST characterization and comparison of cDNA libraries from different genotypes with variable ploidy levels and reproductive modes will allow the identification of genes related to apomixis and ploidy-regulated gene expression. These candidate genes will be completely sequenced and characterized. In addition, the 254 SSR and 190 SNP markers developed from the libraries will facilitate the mapping on populations segregating for apomixis or forage quality traits. Markers co-segregating with the mentioned traits could be used as probes for positional cloning and genomic landmarks to identify these genomic regions. Comparison between mapping results and expression studies can also help to identify candidate genes. Alternatively sequences identified here can be used to improve forage quality. In this regard, one of the most important parameters to be considered is the lignin profile, which strongly influences dry matter digestibility (Spangenberg et al. 1998). The availability of sequences involved in this pathway will allow molecular breeding by down-regulating lignin biosynthetic enzymes through antisense and sense suppression.

Acknowledgments The project was supported by the Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT), Grants PAV 137/4 and PICT 14625 and Secretaría de Ciencia y Tecnología (SECyT UNS) PGI 24/A133 (Argentina).

References

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651–1656

Altschul SF, Boguski MS, Gish W, Wootton JC (1994) Issues in searching molecular sequence databases. Nat Genet 6:119–129

Barker G, Batley J, O'Sullivan H, Edwards KJ, Edwards D (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. Bioinformatics 19:421–422

Bennett MD, Smith JB (1976) Nuclear DNA amounts in angiosperms. Phil Trans, Royal Soc L, Series B 274:227–274

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

Cardone S, Polci P, Selva JP, Mecchia M, Pessino S, Hermann P, Cambi V, Voigt P, Spangenberg G, Echenique V (2006) Novel genotypes of the subtropical grass *Eragrostis curvula* for the study of apomixis (diplospory). Euphytica 151:263–272

Conesa A, Götz S, García-Gómez JM, Perol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatic 2:3674–3676

Covas G, Cairnie A (1991) Introducción del pasto llorón en la Argentina. In: Fernandez O, Brevedad R, Gargajo (eds) El Pasto llorón su biología y manejo. CERZOS, Bahía Blanca, Argentina, pp 1–6

Echenique CV, Polci P, Mroginski L (1996) Plant regeneration in weeping lovegrass, (*Eragrostis curvula*) through inflorescence culture. Plant Cell Tissue Organ Cult 46:123–130

Eujayl I, Sorrells ME, Baum M, Wolters P, Powell W (2002) Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. Theor Appl Genet 104:399–407

Ewing B, Hillier LD, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using Phred. I. Accuracy Assessment. Genome Res 8:175–185

Gao LF, Jing RL, Huo NX, Li Y, Li XP, Zhou RH, Chang XP, Tang JF,·Ma ZY, Jia UZ (2004) One hundred and one new microsatellite loci derived from ESTs (EST-SSRs) in bread wheat. Theor Appl Genet 108:1392–1400

Garg K, Green P, Nickerson DA (1999) Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. Genome Res 9:1087–1092

Hatey F, Tosser-Klopp G, Clouscard-Martinato C, Mulsant P, Gasser F (1998) Expressed sequence tags for genes: a review. Genet Sel Evol 30:521–541

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877

Ingram AL, Doyle JJ (2003) The origin and evolution of *Eragrostis tef* (Poaceae) and related polyploids: evidence from unclear waxy and plastid rps 16. Am J Bot 90:116–122

Ji W, Li Y, Li J, Dai C-h, Wang X, Bai X, Cai H, Liang Yang L, Zhu Y-m (2006) Generation and analysis of expressed sequence tags from NaCl-treated *Glycine soja*. BMC Plant Biol 6:4

Kantety RV, La Rota M, Matthews DE, Sorrells ME (2002) Data mining for simple sequences repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol Biol 48:501–510

Katti MV, Ranjekar PK, Gupta VS (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol 18:1161–1167

La Rota M, Kantety RV, Yu J-K, Sorrells ME (2005) Nonrandom distribution and frequencies of genomic and ESTderived microsatellite markers in rice, wheat, and barley. BMC Genomics 6:23

Lagercrantz U, Ellegren H, Anderson L (1993) The abundance of various polymorphic microsatellite motifs differs between plants and vertebrates. Nucleic Acids Res 21:1111–1115

Lazo GR, Chao S, Hummel DD, Edwards H, Crossman CC, Lui N, Matthews DE, Carollo VL, Hane DL, You FM, Butler GE, Miller RE, Close TJ, Peng JH, Lapitan NLV, Gustafson JP, Qi LL, Echalier B, Gill BS, Dilbirligi M, Randhawa HS, Gill KS, Greene RA, Sorrells ME, Akhunov ED, Dvořák J, Linkiewicz AM, Dubcovsky J, Hossain KG, Kalavacharla V, Kianian, Mahmoud AA, Miftahudin, Ma XF, Conley EJ, Anderson JA, Pathan MS, Nguyen HT, McGuire PE, Qualset CO, Anderson OD (2004) Development of an Expressed Sequence Tag (EST) Resource for Wheat (*Triticum aestivum* L.): EST generation, unigene analysis, probe selection and bioinformatics for a 16,000-Locus Bin-Delineated Map. Genetics 168:585–593

Lewin B (1994) Genes V. Oxford University Press, New York

Ma C-X, Casella G, Wu R (2002) Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. Genetics 16:1751–1762

Mecchia MA, Ochogavía A, Selva JP, Laspina N, Felitti S, Martelotto LG, Spangenberg G, Echenique V, Pessino SC (2007) Genome polymorphisms and gene differential expression in a 'back-and-forth' ploidy-altered series of weeping lovegrass (*Eragrostis curvula*). J Plant Physiol 164(8):1051–1061

Morgante M, Olivieri AM (1993) PCR-amplified microsatellites as markers in plant genetics. Plant J 3:175–182

Murashige T, Skoog F (1962) A revised medium for rapid growth and bioassay with tobacco tissue cultures. Physiol Plant 15:473–497

Nicot N, Chiquet V, Gandon B, Amilhat L, Legeai F, Leroy P, Bernard M, Sourdille P (2004) Study of simple sequence repeat (SSR) markers from wheat expressed sequence tags (ESTs). Theor Appl Genet 109(4):800–805

Peng JH, Lapitan NLV (2005) Characterization of EST-derived microsatellites in the wheat genome and development of eSSR markers. Funct Integr Genomics 5:80–96

Poverene MM, Curvetto NR (1991) Citogenética. In: Fernandez O, Brevedad R, Gargajo (eds) El Pasto llorón su biología y manejo. CERZOS, Bahía Blanca, Argentina, pp 19–38

Poverene MM, Gardey C, Curvetto NR (1986) Estudios citogenéticos en pasto llorón, *Eragrostis curvula* (Schrad.) Nees *s. lat.* II. Comportamiento meiótico. Rev Univ Nac Río cuarto 6:67–78

Powell W, Machray GC, Provan J (1996) Polymorphism revealed by simple sequence repeats. Trends Plant Sci 1:215–222

Ravel C, Praud S, Linossier AML, Balfourier MDF, Brunel PDD, Charmet G (2005) Identification of *Glu-B1-1* as a candidate gene for the quantity of high-molecular-weight glutenin in bread wheat (*Triticum aestivum* L.) by means of an association study. Theor Appl Genet 112:738–743

Robinson AJ, Christopher G, Love CG, Batley J, Barker G, Edwards D (2004) Simple sequence repeat marker loci discovery using SSR primer. Bioinformatics 20:1475–1476

Rostoks N, Mudie S, Cardle L, Russell J, Ramsay L, Svensson ABJT, Wanamaker AI, Walia H, Hedley EM, Liu H, Close JM, Marshall DF, Waugh R (2005) Genome-wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress. Mol Gen Genomics 274:515–527

Sawbridge T, Ong E-K, Binnion C, Emmerling M, Meath K, Nunan K, O'Neill M, O'Toole F, Simmonds J, Wearne K, Winkworth A, Spangenberg G (2003a) Generation and analysis of expressed sequence tags in white clover (*Trifolium repens L.*) Plant Sci 165:1077–1087

Sawbridge T, Ong E-K, Binnion C, Emmerling M, McInnes R, Meath K, Nguyen N, Nunan K, O'Neill M, O'Toole F, Rhodes C, Simmonds J, Tian P, Wearne K, Webster T, Winkwort A, Spangenberg G (2003b) Generation and analysis of expressed sequence tags in perennial ryegrass (*Lolium perenne L.*) Plant Sci 165:1089–1100

Tautz D (1989) Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucl Acids Res 17: 6463–6471

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

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

Van der Hoevent R, Ronning C, Giovannoni J, Martin G, Tanksley S (2002) Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. Plant Cell 14:1441–1456

Voigt P, Rethman N, Poverene M (2004) Lovegrasses. In: American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Warm-Season (C4) Grasses, Agronomy Monograph No.45. Chapter 32:1027–1056

Wang JP, Lindsay BG, Leebens-Mack J, Cui L, Wall K, Miller WC, dePamphilis CW (2004) EST clustering error evaluation and correction. Bioinformatics 20:2973–2984

Wang J-PZ, Lindsay BG, Liying Cui L, Wall PK, Marion J, Zhang J, de Pamphilis CW (2005) Gene capture prediction and overlap estimation in EST sequencing from one or multiple libraries. BMC Bioinformatics 6:300

Wartson L, Dallwitz WJ (1992) The grass genera of the world. CAB International, Wallingford

Xu SS, Khan K, Klindworth DL, Faris JD, Nygar G (2004) Chromosomal location of genes for novel glutenin subunits and gliadins in wild emmer wheat *Triticum turgidum* L. var. dicoccoides). Theor Appl Genet 108:1221–1228

Yeho H-H, Watson L (1986) Taxonomic patterns in protein amino acid profiles of grass leaves and caryopses. In: Soderstrom TR, Hilu KW, Campbell CS, Barkworth ME (eds) Grass system and evolution. Smithsonian Insitution Press, Washington, DC, pp 88–96

Yu J-K, Sun Q, La Rota M, Edwards H, Tefera H, Sorrells ME (2006) Expressed sequence tag analysis in tef [*Eragrostis tef* (Zucc.) Trotter]. Genome 49:365–372