Genetics and Molecular Biology, 24 (1-4), 77-84 (2001)

Dissecting the sugarcane expressed sequence tag (SUCEST) database: unraveling flower-specific genes

R.C. Figueiredo[1], M.S. Brito[1], L.H.M. Figueiredo[1], A.C. Quiapin[1], P.M. Vitorelli[1], L.R. Silva[1], R.V. Santos[2], J.B. Molfetta[1], G.H. Goldman[3] and M.H.S. Goldman[1,]*

Abstract

There are almost 260,000 independent clones sequenced from the 5' end in the Sugarcane Expressed Sequence Tag (SUCEST) database, which have been obtained from 37 cDNA libraries prepared from different tissues. This large number of expressed sequence tags (ESTs) provides an opportunity, unprecedented in plants, to perform 'digital differential screening' on selected cDNA libraries. In general, the frequency of a particular EST correlates with transcript accumulation in the tissues from which the cDNA libraries were constructed, so it is possible to compare the whole transcriptome from different tissues using computer-assisted analysis of an EST database. In our research we analyzed sugarcane ESTs according to tissue expression and identified more than 1,000 putative flower-specific genes. The fact that using this technique we were able to identify sugarcane homologues of several genes previously described as pollen-specific justifies this method of assessing tissue specificity. In addition, ESTs similar to genes specific to reproductive organs were detected *e.g.* a sugarcane gene encoding a meiotic protein essential for assembly of the synaptonemal complex and normal synapsis. This approach also allowed the identification of many flower-specific anonymous sequences that are good candidates for being novel genes involved in plant reproduction. This paper describes the analysis of the gene expression levels of 24 EST clusters during flower development using a 'digital northern blot' constructed from direct EST counts made on the non-normalized sugarcane cDNA libraries.

INTRODUCTION

At the time of writing this paper (February 2001), the number of plant expressed sequence tags (ESTs) deposited in public databases has been growing, but so far no other plant species has so many ESTs described as sugarcane (SUCEST Project - http://sucest.lad.dcc.unicamp.br/en/). Sugarcane (*Saccharum* spp.) is one of the most important crop plants and is cultivated in tropical and subtropical areas in more than 80 countries where it is mainly used for the production of sugar and ethanol.

A potential application of EST databases is the study of the expression of plant genes (Ewing *et al*., 1999). Digital analysis of gene expression can be achieved by generation of tags to expressed genes and transcript abundance inferred from the frequency of these tags (Ewing *et al*., 1999). Several studies have observed that the abundance of ESTs for many genes varies according to the tissue of origin of the cDNA library (Ewing *et al*., 1999). The availability of a significant EST database from a certain plant offers the possibility of studying gene expression for different tissues and organs. The extensive representation of the sugarcane genome in the SUCEST database can be used to detect genes exhibiting tissue-specific expression.

An important challenge in the study of the growth of higher organism, including plants, is to understand both the spatial and temporal regulation of gene expression. Indeed, the development of distinct tissues and cell-types is highly dependent on specific patterns of gene expression and transcript accumulation (Ewing *et al*., 1999; Pesole *et al*., 2000). The analysis of an EST database offers a complete overview of the genes expressed in a certain organ and their relative expression levels and may suggest possible interrelationships between them.

Flowers are responsible for reproduction in angiosperms, and understanding flowering and plant reproduction is not only a fundamental concern of plant biology but is also of practical interest in agriculture. The identification of genes involved in flowering and plant reproduction could greatly contribute to improvements in plant breeding and in establishing alternative techniques to obtain interesting agronomic traits.

MATERIAL AND METHODS

The construction of the sugarcane (SC) cDNA libraries, as well as the information about sequencing and clustering of the reads have already been described in previous papers in this volume. The clusters considered in our work were defined using the Phred/Phrap Program. The flower libraries (FL) which are the subject of this paper consist of cDNAs representing genes expressed in whole flowers at different developmental stages (FL1, FL3 and

[1]*Depto. Biologia FFCLRP/Universidade de São Paulo, Av. Bandeirantes 3900, 14040-901 Ribeirão Preto, SP, Brazil.*
[2]*Centro de Biologia Molecular e Engenharia Genética - UNICAMP, Campinas, SP, Brazil.*
[3]*Depto. C. Farmacêuticas FCFRP/Universidade de São Paulo, Av. do Café s/no., 14040-903 Ribeirão Preto, SP, Brazil.*
Send correspondence to M.H.S. Goldman. E-mail: mgoldman@ffclrp.usp.br.

FL2/FL5/FL6) and in flower stems in two subsequent developmental stages (FL8 and FL4).

Identification of flower-specific clusters

Our approach to identifying putative flower-specific genes was to use 'digital differential screening' (DDS) to search for clusters containing reads exclusively from flower cDNA libraries. To this end we wrote scripts in the Perl programming language to query the SUCEST MySQL relational database. The function of these scripts was to identify the tissue of origin (library) of the reads inside each cluster. If a cluster was entirely formed by reads from flower libraries the script would record its name and use the consensus sequence of the cluster to perform BLAST (Basic Local Assignment Search Tool) searches (Altschul *et al*., 1990) against the NR, NT and EST GenBank databases (Benson *et al*., 2000).

Developmental expression analysis of the putative flower-specific genes

The expression of the putative flower-specific genes was inferred from the number of independent reads from each library present in the cluster. When there was more than one read from the same clone (*i.e.* reads from the 5' and 3' ends, or control reads) only a single read was considered in the analysis. The number of reads was normalized (taking into account the total number of reads sequenced in each library), multiplied by 10,000, and the results used to prepare a 'digital northern blot'.

RESULTS AND DISCUSSION

Analysis of Flower-Specific Genes from the SUCEST database

To our knowledge, the only previous DDS study is that of Schmitt *et al*. (1999) and our study is the first one accomplished in plants. From 81,223 clusters defined using the Phred/Phrap program our DDS procedure resulted in 12,503 (15.4%) putative flower-specific clusters, of which we chose 24 clusters for more detailed analysis.

Table 1 shows the best results obtained in BLAST searches (BLAST EST, BLAST NR or BLAST NT) for each of the sequences. The highest similarity found for 20 out of the 24 (83.3%) clusters was a sequence of unknown

Table I - The best result obtained in Blast searches for each 24 putative flower-specific clusters of sugarcane.

Cluster id	Best hit	Accession number	BLAST	E-value
SCSBFL1039E08	Water-stressed 1 (WS1) *Sorghum bicolor* cDNA	gi7554919	BLAST EST	0.0
SCSBFL4011C12	Water-stressed 1 (WS1) *Sorghum bicolor* cDNA	gi7661256	BLAST EST	1e-99
SCQGFL3053E07	Water-stressed 1 (WS1) *Sorghum bicolor* cDNA	gi9304748	BLAST EST	e-100
SCSFFL4016G02	Ovary 1 (OV1) *Sorghum bicolor* cDNA	gi10420254	BLAST EST	0.0
SCEPFL3081D03	Ovary 1 (OV1) *Sorghum bicolor* cDNA	gi10420414	BLAST EST	3e-63
SCCCFL6002G10	Early embryo from Delaware *Zea mays* cDNA	gi6626589	BLAST EST	e-163
SCSBFL5017H12	Early embryo from Delaware *Zea mays* cDNA	gi6021224	BLAST EST	e-106
SCRLFL1006H06	Ear tissue cDNA library from Schmidt lab *Zea mays* cDNA	gi5268819	BLAST EST	0.0
SCSBFL1044A08	Ear tissue cDNA library from Schmidt lab *Zea mays* cDNA	gi5018253	BLAST EST	2e-29
SCCCFL4002D03	Light Grown 1 (LG1) *Sorghum bicolor* cDNA	gi7218411	BLAST EST	5e-90
SCCCFL5003C08	14 day immature embryo from Hake lab (HS) *Zea mays* cDNA	gi6022142	BLAST EST	e-156
SCCCFL5054H07	Inbred Tassel cDNA Library *Zea mays* cDNA	gi5688519	BLAST EST	e-128
SCJFFL3C02F01	putative serine/threonine protein kinase *Arabidopsis thaliana*	gi4678928	BLAST NR	1e-23
SCRLFL1008D07	*Zea mays* liguleless1 protein (liguleless1) mRNA, complete cds	gi1914844	BLAST NT	5e-37
SCRUFL4022D11	Mixed adult tissues cDNA library from Walbot lab *Zea mays* cDNA	gi6696021	BLAST EST	3e-81
SCRUFL4020G06	*Hordeum vulgare* 5-45 DAP spike EST library	gi9859663	BLAST EST	e-146
SCSBFL1042A08	Endosperm 10-14 DAP cDNA library from Schmidt lab *Zea mays* cDNA	gi5455809	BLAST EST	e-170
SCRLFL1014H04	Meiotic asynaptic mutant 1 – *Arabidopsis thaliana*	gi7939627	BLAST NR	1e-25
SCJLFL1052A03	Dark Grown 1 (DG1) *Sorghum bicolor* cDNA	gi9300843	BLAST EST	0.0
SCSGFL5C02C11	etiolated seedling *Zea mays* cDNA clone	gi453350	BLAST EST	6e-73
SCSGFL4C05D11	Pathogen induced 1 (PI1) *Sorghum bicolor* cDNA	gi9851789	BLAST EST	0.0
SCRUFL3069E08	tassel primordium prepared by Schmidt lab *Zea mays* cDNA	gi9961464	BLAST EST	e-125
SCCCFL6001E11	pectin methylesterase-like protein *Arabidopsis thaliana*	gi9759007	BLAST NR	6e-18
SCSBFL4012B01	Mixed stages of anther and pollen *Zea mays* cDNA	gi7137551	BLAST EST	3e-66

function from a monocotyledon EST project (*Zea mays*, *Sorghum bicolor* or *Hordeum vulgare*).

Table 2 presents the results of a search for the best match to a sequence with a known function, but when this criterion was not fulfilled the result of Table 1 was repeated on Table 2. It should be noted that the data in both these tables have expected E-values below e-05. Table 2 shows that 10 out of 24 clusters (41,7%) were anonymous sequences that may represent novel flower-specific genes. Although cluster SCSBFL5017H12 was found to encode a protein similar to an unknown protein of *Arabidopsis thaliana*, clusters SCSBFL1039E08, SCSBFL4011C12, SCRLFL1006H06, SCSBFL1044A08, SCCCFL5003C08, SCCCFL5054H07, SCRUFL4022D11, SCRUFL3069E08 and SCSBFL4012B01 showed no similarity to *Arabidopsis* sequences. Considering that the whole *Arabidopsis* genome is known it is tempting to suggest that this lack of similarity may indicate that these sequences are specific to monocotyledonous plants. The future analysis of genomes from other dicotyledonous and monocotyledonous plants may help to determine the existence, or not, of flower-specific genes that are specific to monocotyledonous plants.

Clusters SCSBFL1039E08 and SCSBFL4011C12 had very strong similarity to ESTs from a water-stressed *Sorghum bicolor* cDNA library. Although the role of the proteins encoded by these clusters is still unknown, a pollen-specific and desiccation-associated transcript (LLA23) has been described by Huang *et al*. (2000) in *Lilium longiflorum*. The predicted LLA23 polypeptide has significant similarity to a group of water-deficit/ripening -induced proteins and seems to have a protective function during pollen maturation.

Clusters SCRLFL1006H06, SCSBFL1044A08, SCCCFL5054H07, SCRUFL3069E08 and SCSBFL4012B01 were significantly similar to ESTs from some *Z. mays* cDNA libraries (from ear, tassel and anther/pollen libraries) and may represent transcripts important to plant reproduction. Cluster SCCCFL5003C08 was similar to an EST sequence from a *Z. mays* 14 DAP immature embryo cDNA library and cluster SCRUFL4022D11 to an EST sequence from a *Z. mays* cDNA library of mixed adult tissues.

Table II - The best result obtained in Blast searches for each of the 24 putative flower-specific clusters of sugarcane, with a protein of known function.

Cluster id	Best hit with a known function	Accession number	BLAST	E-value
SCSBFL1039E08	Water-stressed 1 (WS1) *Sorghum bicolor* cDNA	gi7554919	BLAST EST	0.0
SCSBFL4011C12	Water-stressed 1 (WS1) *Sorghum bicolor* cDNA	gi7661256	BLAST EST	1e-99
SCQGFL3053E07	Type 2 Metallothionein-Like Protein expressed in the tapetum *Zea mays*	gi6689674	BLAST NT	1e-46
SCSFFL4016G02	*Oryza sativa* - 40S Ribosomal Protein S19	gi730456	BLAST NR	7e-75
SCEPFL3081D03	Nonspecific- Lipid-Transfer Protein 4 Precursor (LTP 4)- *Oryza sativa*	gi2497748	BLAST NR	8e-29
SCCCFL6002G10	myosin heavy chain-like protein *Arabidopsis thaliana*	gi8346549	BLAST NR	3e-05
SCSBFL5017H12	similar to unknown protein *Arabidopsis thaliana*	gi9758143	BLAST NR	4e-18
SCRLFL1006H06	Ear tissue cDNA library from Schmidt lab *Zea mays* cDNA	gi5268819	BLAST EST	0.0
SCSBFL1044A08	Ear tissue cDNA library from Schmidt lab *Zea mays* cDNA	gi5018253	BLAST EST	2e-29
SCCCFL4002D03	MFS18 protein precursor embX67324.1ZMFS18 *Zea mays* MFS18 mRNA	gi22646	BLAST NT	2e-46
SCCCFL5003C08	14 day immature embryo from Hake lab (HS) *Zea mays* cDNA	gi6022142	BLAST EST	e-156
SCCCFL5054H07	Inbred Tassel cDNA Library *Zea mays* cDNA	gi5688519	BLAST EST	e-128
SCJFFL3C02F01	Putative serine/threonine kinase - *Arabidopsis thaliana*	gi4678928	BLAST NR	1e-23
SCRLFL1008D07	*Zea mays* liguleless1 protein	gi1914844	BLAST NT	5e-37
SCRUFL4022D11	Mixed adult tissues from Walbot lab (SK) *Zea mays* cDNA	gi6696021	BLAST EST	3e-81
SCRUFL4020G06	TGF beta receptor associated protein *Homo sapiens*	gi4759260	BLAST NR	1e-19
SCSBFL1042A08	Similar to pollen specific like protein *Oryza sativa*	gi5922606	BLAST NR	6e-32
SCRLFL1014H04	Meiotic asynaptic mutant 1 - *Arabidopsis thaliana*	gi7939627	BLAST NR	1e-25
SCJLFL1052A03	*Oryza sativa* beta-expansin (EXPB4) mRNA, complete cds	gi8118424	BLAST NT	e-157
SCSGFL5C02C11	male sterility 2 protein – *Brassica napus*	gi7488458	BLAST NR	2e-35
SCSGFL4C05D11	*Oryza sativa* histone H3 mRNA, complete cds	gi3885889	BLAST NT	e-172
SCRUFL3069E08	tassel primordium prepared by Schmidt lab *Zea mays* cDNA	gi9961464	BLAST EST	e-125
SCCCFL6001E11	pectin methylesterase-like protein *Arabidopsis thaliana*	gi9759007	BLAST NR	6e-18
SCSBFL4012B01	Mixed stages of anther and pollen *Zea mays* cDNA	gi7137551	BLAST EST	3e-66

Some putative flower-specific genes and their possible functions

We found that cluster SCQGFL3053E07 was significantly similar to a type 2 metallothionein-like gene (MZm3-4) preferentially expressed in the tapetum of *Z. mays* (Charbonnel-Campaa *et al*., 2000). The MZm3-4 gene is highly expressed in the male reproductive organs engaged in microsporogenesis, the expression starting at the pollen mother-cell stage and reaching a maximum during meiosis. Charbonnel-Campaa *et al*. (2000) have conducted *in situ* hybridization experiments which have demonstrated that MZm3-4 mRNAs are only present in the highly metabolically active cells of the tapetum (an anther tissue with which microspores establish an essential collaboration) and have suggested a role for the MZm3-4 protein in the fine regulation of metal at the cellular compartment level. We suggest that cluster SCQGFL3053E07 encodes a sugarcane protein with equivalent localization and function.

Cluster SCEPFL3081D03 encoded a protein which resembles a nonspecific lipid-transfer protein 4 precursor from *O. sativa* (gi2497748). Plant nonspecific lipid-transfer proteins transfer phospholipids as well as galactolipids across membranes, and may play a role in wax or cutin deposition in the cell wall of expanding epidermal cells and certain secretory tissues. Anther-specific lipid-transfer genes have been described in *A. thaliana* (Rubinelli *et al*., 1998) and *Z. mays* (Lauga *et al*., 2000). In *Z. mays*, this gene (MZm3-3) is highly and preferentially expressed in the tapetum, from the start of the pollen mother-cell stage to the uni-nucleated microspore stage, suggesting that MZm3-3 contributes to the formation of the pollen coat (Lauga *et al*., 2000). In addition, Mollet *et al*. (2000) have described a lipid transfer protein in lily, now called stigma/stylar cysteine-rich adhesin which is required for lily pollen tube adhesion to the stylar transmitting tissue. At the moment, the information we have on cluster SCEPFL 3081D03 is not enough to suggest that this cluster is anther or stigma/stylar specific.

Cluster SCCCFL4002D03 was similar to a *Z. mays* cDNA designated MFS18 that encodes a polypeptide (rich in glycine, proline and serine) which has similarities with plant structural proteins and which has enhanced expression in male flowers (Wright *et al*., 1993). According to Wright *et al*. (1993), MFS18 mRNA accumulates in the glumes, anther walls, paleas and lemmas of mature florets but is particularly associated with the vascular bundle in the glumes. We suggest that cluster SCCCFL4002D03 encodes a protein with a comparable localization and function.

Members of the protein kinase superfamily catalyze the reversible transfer of γ-phosphate from ATP to serine, threonine, or tyrosine amino acids on target proteins. One protein kinase can phosphorylate many hundreds of target proteins, greatly amplifying weak signals (Buchanan *et al*., 2000). Protein kinases have been implicated in a great number of plant processes, including the development of plant reproductive organs in *Petunia hybrida* and *A. thaliana* (Decroocq-Ferrant *et al*., 1995; Dornelas *et al*., 2000) and it is reasonable to assume that flower-specific protein kinases also exist in sugarcane. We found that cluster SCJFFL 3C02F01 encoded a protein similar to a putative serine/threonine protein kinase from *A. thaliana* (Table 1).

Cluster SCRLFL1008D07 encoded a protein which resembles the *Z. mays* LIGULELESS 1 protein, a nuclear-localized protein required for the induction of ligules and auricles during leaf organogenesis (Moreno *et al*., 1997). This protein contains an internal domain similar to a conserved DNA-binding domain called the SQUAMOSA promoter-binding protein (SBP) domain (Moreno *et al*., 1997). A protein containing the SBP domain is involved in floral transition in *Arabidopsis* by binding to the APETALA 1 promoter (Cardon *et al*., 1997). The sugarcane homologues for the SQUAMOSA and APETALA 1 genes have been identified in the SUCEST project (unpublished data) and we suggest that cluster SCRLFL1008D07 encodes a protein that may bind to the promoter of at least one of these genes in sugarcane.

Cluster SCRUFL4020G06 encoded a protein similar to the transforming growth factor beta-receptor associated protein from *Homo sapiens*. The possible role of such a protein in plants is completely speculative and the fact that cluster SCRUFL4020G06 matched best to a protein from humans, instead of to other proteins from plants, is definitely intriguing.

The protein encoded by cluster SCSBFL1042A08 is similar to a putative pollen-specific protein identified in *O. sativa*, although no further information is available about the possible function of this protein.

Synapses of homologous chromosomes is a key event in meiosis as it is essential for normal chromosome segregation and is implicated in the regulation of the frequency of crossovers (Caryl *et al*., 2000). In our study, the protein encoded by cluster SCRLFL1014H04 (Table 1) showed significant similarity to an *A. thaliana* protein essential for synaptonemal complex assembly and normal synapses which was recognized by Caryl *et al*., (2000) during the study of the meiotic asynaptic mutant 1 of *A. thaliana*.

Cluster SCJLFL1052A03 had a very strong similarity to the nucleotide sequence of a cDNA encoding the *O. sativa* beta-expansin, extracellular proteins that loosen plant cell walls and are thought to function (in grasses) in pollen tube invasion of the stigma (Cosgrove, 2000). Multigene families encode expansins and each gene is often expressed in highly specific locations and cell types (Cosgrove, 2000), and it is probable that cluster SCJLFL1052A03 belongs to the sugarcane expansin multigene family and is specifically expressed in some flower tissue.

The *A. thaliana* MALE STERILITY 2 (MS2) gene product is involved in male gametogenesis and expression of this gene has been observed in tapetum of flowers at (and shortly after) the release of microspores from the tetrads (Aarts *et al*., 1997). The MS2 gene product shows sequence similarity to a jojoba protein that converts waxy fatty acids to fatty alcohols and, according to Aarts *et al*. (1997), may function as a fatty acyl reductase in the formation of pollen wall components. We found that cluster SCSGFL5C02C11 encoded a protein similar to the MALE STERILITY 2 protein from *Brassica napus*.

Cluster SCCCFL6001E11 encoded a protein similar to a pectin methylesterase-like protein from *A. thaliana*. A flower-specific gene (exclusive to male flowers) for a pectin methylesterase has been described in *Salix gilgiana* by Futamura *et al*. (2000), who found that the expression of this gene was developmentally regulated in male reproductive organs and that maximal expression occurred when male flowers were fully open and mature. Futamura *et al*. (2000) have demonstrated by *in situ* hybridization that expression was restricted to mature pollen grains after microspore mitosis. In addition, a pollen-specific gene for pectin methylesterase has been described in *Lotus japonicus* by Endo *et al*. (2000). It is reasonable to suppose that sugarcane may use comparable mechanisms during its reproductive process, and express similar genes only in male flowers or pollen.

Can proteins involved in basic cellular functions be flower-specific?

Cluster SCSGFL4C05D11 encoded a protein very similar to *O. sativa* histone H3, and this result raises the question whether or not it is possible for a flower-specific gene to encode a histone. Ueda *et al.* (2000) have described three novel histone genes which are specifically expressed in male gametic cells of *L. longiflorum*. They have proposed that these male gamete-specific core histones contribute to chromatin condensation in male gametes or to chromatin remodeling which results in the repression of gene expression in male gametes. It is therefore reasonable to assume that cluster SCSGFL4C05D11 encodes a truly flower-specific histone.

Other examples of proteins, involved in basic cellular functions and already described as specific to certain organs/tissues are an actin preferentially expressed in pollen (Huang *et al*., 1996), a pollen-specific actin-depolymerizing factor (Smertenko *et al*., 2001) and a pollen-specific calmodulin-binding protein (Safadi *et al*., 2000).

In plant cells, myosin is believed to be the molecular motor responsible for actin-based motility processes such as cytoplasmic streaming and directed vesicle transport (Kinkema and Schiefelbein, 1994), two processes which are important during pollen tube growth. Since an actin preferentially expressed in pollen has already been reported (Huang *et al*., 1996), it is reasonable to speculate about the existence of a pollen-specific myosin, and we found (Table 2) a flower-specific sugarcane cluster, SCCCFL6002G10, encoding a protein similar to the myosin heavy chain-like protein of *A. thaliana*.

The protein encoded by cluster SCSFFL4016G02 was similar to a 40S ribosomal protein S19 coded by a *O. sativa* gene, but to our knowledge no ribosomal protein has been described as specific to a certain organ or tissue. However, it is interesting to note that this *O. sativa* sequence (gi730456) belongs to an anther cDNA library.

Over the last few years there has been an increase in the number of papers reporting detailed analysis of gene expression and the discovery of new organ/tissue specific genes. Future experiments using northern blotting, arrays and/or *in situ* hybridization could confirm the results obtained in our digital differential screening.

Most of the flower-specific genes reported in this paper were related to male flower organs such as pollen or anther-specific genes. We believe this reflects the fact that these organs have been studied for much longer (and many more of their genes described) than female flower organs rather than the existence of fewer female-specific genes in plants.

Digital expression profile

Mekhedov *et al*. (2000) have stated that one of the major contributions that can emerge from large-scale EST sequencing projects is information on gene expression levels. Generally, the frequency of a particular EST correlates with transcript accumulation in the tissues from which the cDNA library originated. In our study it was possible to analyze the EST frequency of 24 clusters in 7 flower cDNA libraries because the sugarcane flower cDNA libraries we used were non-normalized.

Figure 1 shows the frequency of reads per library found for each cluster. As expected by the very low number of reads sequenced from library FL2, reads from this library

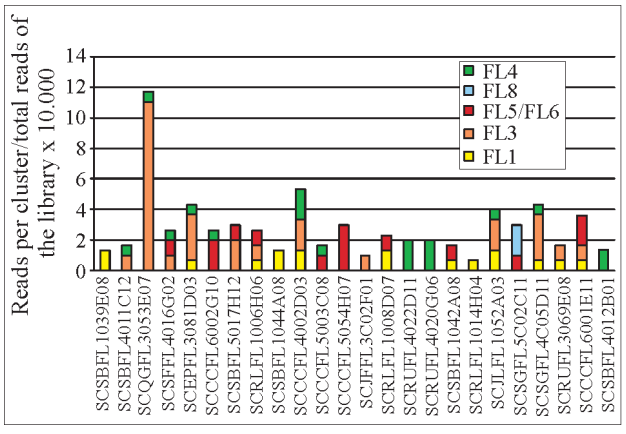

Figure 1 - EST frequency for 24 of the flower-specific clusters identified in the SUCEST database. The colors indicate the contribution of each cDNA library.

were not represented on the clusters analyzed. Genes corresponding to sugarcane clusters SCSBFL1039E08, SCSBFL1044A08 and SCRLFL1014H04 were only expressed in the FL1 cDNA library. Cluster SCJFFL 3C02F01, encoding a putative serine/threonine kinase, was the only cluster expressed solely in library FL3. All the reads of cluster SCCCFL5054H07 were from the FL5/FL6 cDNA libraries. Clusters SCRUFL4022D11, SCRUFL 4020G06 and SCSBFL4012B01 were detected exclusively in FL4 library. From all the 24 flower-specific clusters, the single one containing reads from the FL8 cDNA library is cluster SCSGFL5C02C11.

A 'digital northern blot' of the 24 flower-specific clusters (Figure 2) allows the visualization of the relative gene expression levels of these clusters in each of the libraries. Based on this, it seems that sugarcane clusters SCRUFL4022D11, SCRUFL4020G06 and SCSBFL4012 B01 were specific to flower stems and were not expressed in tissues involved in plant reproduction. This is in agreement with the fact that each of these clusters encoded a protein to which no plant reproduction function could be associated. On the other hand, clusters SCSBFL1039E08, SCSBFL5017H12, SCRLFL1006H06, SCSBFL1044A08, SCCCFL5054H07, SCJFFL3C02F01, SCRLFL1008D07, SCSBFL1042A08, SCRLFL1014H04, SCRUFL3069E08 and SCCCFL6001E11 were specific to flowers where the tissues specialized in plant reproduction are located. As mentioned above, all these clusters, except cluster SCSBFL 5017H12, encode proteins that may have a role in plant reproduction. Cluster SCSBFL5017H12 is similar to an *A. thaliana* protein with no recognized function, and is, there-

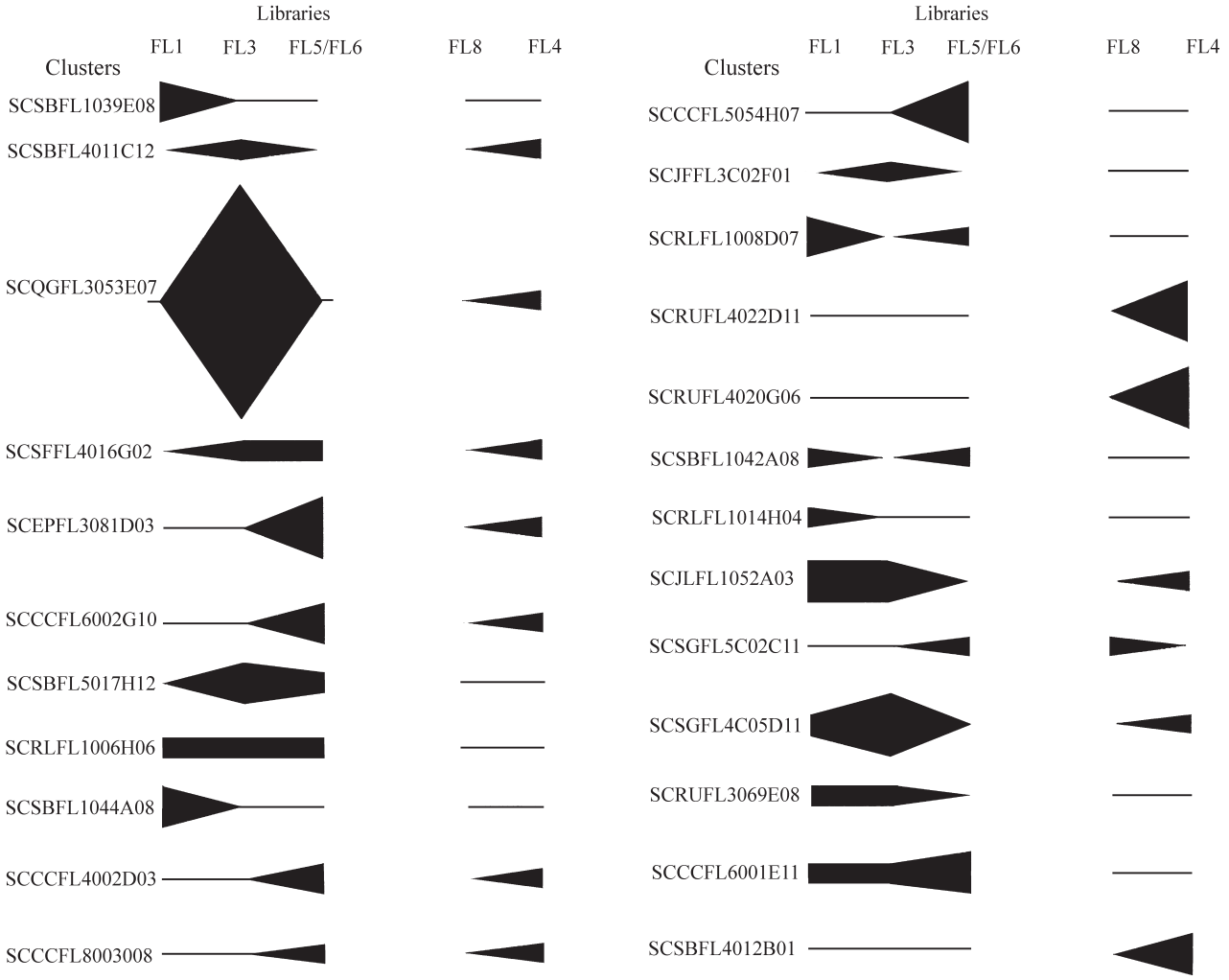

Figure 2 - 'Digital northern blot' of the 24 flower-specific clusters. The figure represents gene expression during sugarcane flower development, based on expressed sequence tag (EST) counts at each developmental stage as inferred from the description of the material used for the construction of the cDNA libraries. The flower cDNA libraries (FL) were prepared from 1cm sugarcane flowers (FL1), 5 cm flowers (FL3) and 20 cm flowers (FL5/FL6), while FL8 was constructed from 10 cm long flower stems and the FL4 library from 20 cm long flower stems. We considered these cDNA libraries to represent two distinct groups of flower material, which is why a virtual empty lane separates these two groups in the figure. For each cluster there is a virtual axis, the zero read position being centralized on the cluster name and represented by a straight line. The wider the area under a flower cDNA library, the higher is the expression level of the corresponding cluster in that stage of development.

fore, a good candidate for a novel gene involved in sexual plant reproduction.

The gene expression level of cluster SCRLFL1006 H06 is low and with little variation throughout flower development. Meanwhile, the gene corresponding to cluster SCQGFL3053E07 is not expressed at the first flower developmental stage (FL1) but increases its expression rapidly reaching a high level of expression at the FL3 stage and decreases to undetectable levels at the FL5/FL6 stage. Although not so pronounced, all the other clusters have changes in their expression levels, which increase or decrease during flower development (Figure 2).

The sugarcane clusters encoding proteins involved in basic cellular functions were not expressed throughout development (*e.g.* SCSFFL4016G02 encoding the 40S ribosomal protein S19, SCCCFL6002G10 encoding the myosin heavy chain-like protein and SCSGFL4C05D11 encoding histone H3), which suggests that their floral-specific functions were not necessary in all stages of development. In the developmental stages when these clusters were not expressed there were probably other genes/clusters which provided the gene products necessary for the corresponding cellular function.

The polyploidy nature of the sugarcane genome and the existence of different flower-specific alleles

A general analysis of the 12,503 flower-specific clusters revealed the presence of independent clusters with similar consensus sequences and these clusters also showed similarity in BLAST searches with the same sequence from public databases. For example, sugarcane cluster SCQGFL 3053E07 (Tables 1 and 2) encoded a protein similar to a metallothionein-like protein (BLAST NR E-value = 2.e-16) from rice, but there were also five additional flower-specific clusters with similarity to exactly the same sequence from rice (sugarcane clusters SCJLFL3014G10 (1.e-13), SCSFFL4083D03 (3.e-18), SCEQFL5053H11 (2.e-14), SCRUFL3063A10 (3.e-18) and SCCCFL4090 B09 (0.0005)). Another example is sugarcane cluster SCCCFL4002D03 which showed similarity to the MFS 18 protein precursor from *Z. mays* (BLAST NR E-value = 0.005) and there were six other flower-specific clusters with similarity to the same protein (SCSBFL1043H12 (5.e-09), SCQGFL8018B05 (5.e-06), SCSFFL4017H11 (2.e-06), SCSBFL5015H05 (6.e-06), SCRLFL8049E11 (0.005) and SCSBFL1043H05 (4.e-09)). Among the 24 sugarcane clusters discussed in this paper, the same type of observation applies for clusters SCSFFL4016G02 (with three other flower-specific clusters showing the best BLAST result to the same database sequence), cluster SCEPFL3081D03 (with four other clusters), cluster SCJLFL1052A03 (with two other clusters) and cluster SCSGFL4C05D11 (with eight other clusters).

One possible explanation for this phenomenon is that these clusters represent members of a gene family, although different members of a gene family usually have differential expression (spatial, temporal, or inducible) but all the clusters mentioned above are flower-specific. Since sugarcane is an octaploid it is to be expected that different alleles may exist for several loci, and it is also possible that these clusters are alleles of one single gene or one flower-specific member of a gene family. If this is true, the number of sugarcane flower-specific genes may be much lower than the 12,503 clusters identified.

It is interesting to note that in the two examples mentioned above, there is at least one cluster with a high E-value, very distinct from the E-values of the other clusters. The significance of this observation and its possible correlation with the evolution of divergent alleles remains to be elucidated.

RESUMO

Existem quase 260.000 clones independentes, seqüenciados a partir da extremidade 5', no banco de dados do SUCEST (Sugarcane Expressed Sequence Tag), os quais foram obtidos a partir de 37 bibliotecas de cDNA preparadas de diferentes tecidos. Este grande número de etiquetas de sequências expressas (ESTs) fornece uma oportunidade, sem precedentes em plantas, de realizar um 'digital differential screening' em bibliotecas de cDNA selecionadas. Geralmente, a frequência de um determinado EST está correlacionada ao acúmulo de transcritos nos tecidos dos quais as bibliotecas de cDNA foram construídas, e desta forma, é possível comparar o transcriptoma completo de diferentes tecidos, usando uma análise computacional de um banco de dados de ESTs. Em nossa pesquisa, analisamos os ESTs de cana-de-açúcar de acordo com sua expressão tecidual e identificamos mais de 1.000 putativos genes específicos de flor. O fato de que usando esta técnica fomos capazes de identificar homológos em cana-de-açúcar, de vários genes previamente descritos como específicos de pólen, sustenta este método de estimar especificidade tecidual. Além disto, ESTs com similaridade a genes específicos de órgãos reprodutivos foram revelados, como por exemplo, o gene que codifica uma proteína meiótica essencial para a montagem do complexo sinaptonêmico e sinapse normal. Esta abordagem também permitiu a identificação de muitas sequências anônimas, específicas de flor, que são boas candidatas para novos genes envolvidos com a reprodução de plantas. Este trabalho descreve a análise dos níveis de expressão gênica de 24 clusters de ESTs, durante o desenvolvimento floral, usando um 'northern blot digital' construído a partir da contagem direta dos ESTs das bibliotecas não-normalizadas de cDNAs de cana-de-açúcar.

ACKNOWLEDGMENTS

We thank all the members of the SUCEST project involved in construction of the cDNA libraries, sequencing and bioinformatics and especially Dr. André Luiz Vettore

de Oliveira. We also thank Dr. Junior Barrera for suggestions concerning bioinformatics. We are indebted to the Brazilian agency FAPESP for financial support for this project and for fellowships to R.C.F., M.S.B., L.H.M.F., A.C.Q., P.M.V., L.R.S., R.V.S. and J.B.M., and to the Brazilian agency CNPq for fellowships to G.H.G. and M.H.S.G.

REFERENCES

Aarts, M.G., Hodge,R., Kalantidis, K., Florack, D., Wilson, Z.A., Mulligan, B.J. Stiekema, W.J., Scott, R. and **Pereira, A.** (1997). The *Arabidopsis* MALE STERILITY 2 protein shares similarity with reductases in elongation/condensation complexes. *Plant J. 12*: 615-623.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and **Lipman, D.J.** (1990). Basic local alignment search tool. *J. Mol. Biol. 215*: 403-410.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. and **Wheeler, D.L.** (2000). GenBank. *Nucleic Acid Res. 28*: 15-18.

Buchanan, B.B., Gruissem, W. and **Jones, R.L.** (2000). *Biochemistry and Molecular Biology of Plants.* 1st Edition. American Society of Plant Physiology, Maryland.

Cardon, G.H., Hohmann, S., Nettesheim, K., Saedler, H. and **Huijser, P.** (1997). Functional analysis of the *Arabidopsis thaliana* SBP-box gene SPL3: a novel gene involved in the floral transition. *Plant J. 12*: 367-377.

Caryl, A.P., Armstrong, S.J., Jones, G.H. and **Franklin, F.C.** (2000). A homologue of the yeast HOP1 gene is inactivated in the *Arabidopsis* meiotic mutant asy1. *Chromosoma 109*: 62-71.

Charbonnel-Campaa, L., Lauga, B. and **Combes, D.** (2000). Isolation of a type 2 metallothionein-like gene preferentially expressed in the tapetum in *Zea mays*. *Gene 254*: 199-208.

Cosgrove, D.J. (2000). New genes and new biological roles for expansins. *Curr. Opin. Plant Biol. 3*: 73-78.

Decroocq-Ferrant, V., Van Went, J., Bianchi, M.W., de Vries, S.C. and **Kreis, M.** (1995). *Petunia hybrida* homologues of shaggy/zeste-white 3 expressed in female and male reproductive organs. *Plant J. 7*: 897-911.

Dornelas, M.C., Van Lammeren, A.A. and **Kreis, M.** (2000). *Arabidopsis thaliana* SHAGGY-related protein kinase (AtSK11 and 12) function in gynoecium development. *Plant J. 21*: 419-429.

Endo, M., Kokubun, T., Takahata, Y., Higashitani, A., Tabata, S. and **Watanabe, M.** (2000). Analysis of expressed sequence tags of flower buds in *Lotus japonicus*. *DNA Res. 7*: 213-216.

Ewing, R.M., Kahla, A.B., Poirot, O., Lopez, F., Audic, S. and **Claverie,J.M**. (1999). Large scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res. 9*: 950-959.

Futamura, N., Mori, H., Kouchi, H. and **Shinohara, K.** (2000). Male flower-specific expression of genes for polygalacturonase, pectin methylesterase and beta-1,3-glucanase in a dioecious willow (*Salix gilgiana* Seemen). *Plant Cell Physiol. 41*: 16-26.

Huang, J.C., Lin, S.M. and **Wang, C.S.** (2000). A pollen-specific and dessication-associated transcript in *Lilium longiflorum* during development and stress. *Plant Cell Physiol. 41*: 477-485.

Huang, S., Na, Y.Q., McDowell, J.M., McKinney, E.C. and **Meagher, R.B.** (1996). The *Arabidopsis thaliana* ACT4/ACT12 actin gene subclass is strongly expressed throughout pollen development. *Plant J. 10*: 189-202.

Kinkema, M. and **Schiefelbein, J.** (1994). A myosin from a higher plant has structural similarities to class V myosins. *J. Mol. Biol. 239*: 591-597.

Lauga, B., Charbonnel-Campaa, L. and **Combes, D.** (2000). Characterization of MZm3-3, a *Zea mays* tapetum-specific transcript. *Plant Sci. 157*: 65-75.

Mekhedov, S., Ilárduya, O.M. and **Ohlrogge, J.** (2000). Toward a functional catalog of the plant genome. A survey of genes for lipid biosynthesis. *Plant Physiol. 122*: 389-401.

Mollet, J.C., Park, S.Y., Nothnagel, E.A. and **Lord, E.M**. (2000). A lily stylar pectin is necessary for pollen tube adhesion to an in vitro stylar matrix. *Plant Cell 12*: 1737-1750.

Moreno, M.A., Harper, L.C., Kruegger, R.W., Dellaporta, S.L. and **Freeling, M**. (1997). LIGULELESS 1 encodes a nuclear-localized protein required for induction of ligules and auricles during maize leaf organogenesis. *Genes Dev. 11*: 616-628.

Pesole, G., Liuni, S. and **D'Souza, M.** (2000). PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. *Bioinformatics 16*: 439-450.

Rubinelli, P., Hu, Y. and **Ma, H.** (1998). Identification, sequence analysis and expression studies of novel anther-specific genes of *Arabidopsis thaliana*. *Plant Mol. Biol. 37*: 607-619.

Safadi, F., Reddy, V.S. and **Reddy, A.S.** (2000). A pollen-specific novel calmodulin-binding protein with tetratricopeptide repeats. *J. Biol. Chem. 275*: 35457-35470.

Schmitt, A.O., Specht,T., Beckmann, G., Dahl, E., Pilarsky, C.P., Hinzmann, B. and **Rosenthal, A.** (1999). Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. *Nucleic Acid Res. 27*: 4251-4260.

Smertenko, A.P. Allwood, E.G., Khan, S., Jiang, C.J., MacIver, S.K., Weeds, A.G. and **Hussey, P.J.** (2001). Interaction of pollen-specific actin-depolymerizing factor with actin. *Plant J. 25*: 203-212.

Ueda, K., Kinoshita Y., Xu, Z.J., Ide, N., Ono, M., Akahori, Y., Tanaka, I. and **Inoue, M.** (2000). Unusual core histones specifically expressed in male gametic cells of *Lilium longiflorum*. *Chromosoma 108*: 491-500.

Wright, S.Y., Suner, M.M., Bell, P.J., Vaudin, M. and **Greenland, A.J.** (1993). Isolation and characterization of male flower cDNAs from maize. *Plant J. 3*: 41-49.