Plant Molecular Biology (2005) 59:363–374
DOI 10.1007/s11103-005-8924-7

Endosperm-preferred expression of maize genes as revealed by transcriptome-wide analysis of expressed sequence tags

Natalia C. Verza[1,†], Thaís Rezende e Silva[1,†], Germano Cord Neto[1], Fábio T.S. Nogueira[1], Paulo H. Fisch[1], Vincente E. de Rosa Jr[1], Marcelo M. Rebello[1] André L. Vettore[3], Felipe Rodrigues da Silva[4] and Paulo Arruda[1,2,*]

*[1]Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas (UNICAMP), 13083-970, Campinas, SP, Brazil; [2]Departamento de Genética e Evolução, Instituto de BiologiaUniversidade Estadual de Campinas (UNICAMP), 13083-970, Campinas, SP, Brazil (*author for correspondence; e-mail parruda@unicamp.br); [3]Instituto Ludwig de Pesquisa sobre o Câncer, 01509-010, São Paulo, SP, Brazil; [4]Embrapa Recursos Genéticos e Biotecnologia–CENARGEN, Caixa Postal 2372, 70770-900, Brasília, DF, Brazil; [†]These authors contributed equally to this work*

Received 16 January 2005; accepted in revised form 19 June 2005

Key words: Endosperm, Endosperm-specific genes, ESTs, Maize transcriptome

Abstract

The transcriptome-wide endosperm-preferred expression of maize genes was addressed by analyzing a large database of expressed sequence tags (ESTs). We generated 30,531 high quality sequence-reads from the 5′-ends of cDNA libraries from maize endosperm harvested at 10, 15, and 20 days after pollination. A further 196,900 maize sequence-reads retrieved from public databases were added to this endosperm collection to generate MAIZEST, a database with tools for data storage and analysis. MAIZEST contains 227,431 ESTs, one third of which represents developing endosperm and the remaining two-thirds represent transcripts from 49 cDNA libraries constructed from different organs and tissues. Assembling the MAIZEST ESTs generated 29,206 putative transcripts, of which a set of 4032 assembled sequences was composed exclusively of sequences derived from endosperm cDNA libraries. After sequence analysis using overlapping parameters, a sub-set of 2403 assembled sequences was functionally annotated and revealed a wide variety of putative new genes involved in endosperm development and metabolism.

Introduction

Tissue and organ differentiation and development require a concerted network of signaling, regulatory and metabolic processes that is ultimately controlled by the qualitative and quantitative expression of a set of genes (Stolc *et al.*, 2004). In humans, the number of protein-coding genes has been estimated to be around 25,000 (International Human Genome Sequencing Consortium, 2004). Approximately 1000 of these genes are found to be expressed in all cell types and only a small fraction of transcripts are exclusively expressed in an individual tissue (Velculescu *et al.*, 1999). The quantity and nature of these tissue-specific genes are largely unknown. In plants, some mutations specifically affect differentiation, development and the metabolism of certain tissues or organs (Maizel and Weigel, 2004; Tuteja *et al.*, 2004), but the roles of most tissue-specific expressed genes remain unknown.

The availability of large databases of expressed genes offers a good opportunity to identify tissue-specific genes. Over 4 million expressed sequence tags (ESTs) from plant tissues are currently available at GenBank (http://www.ncbi.nlm.nih.

gov/dbEST/dbEST_summary.html; release 121004, December 10, 2004). When these data provide a detailed description of the tissue or organ from which the cDNA libraries were made, it is possible to annotate and compare different library sources and gain insight into tissue-specific expression.

The maize endosperm is a suitable model system for transcriptome analysis because it is formed by only three cell types: starchy endosperm cells, aleurone cells and transfer cells (Olsen 2004). The endosperm development is also well characterized at the cellular level. The first 7–12 days after pollination (DAP) characteristically involve cell division, after which the endosperm cells enlarge and undergo several metabolic processes that result in the deposition of starch and storage proteins (Lopes and Larkins 1993; Berger 1999; Olsen 2004). During endosperm development, a complex gene expression system integrate carbohydrate, amino acid and storage protein metabolism (Giroux *et al.*, 1994; Muller *et al.*, 1997; Arruda *et al.*, 2000; Hunter *et al.*, 2002).

In recent decades, various studies have unraveled many aspects of the biochemistry and cellular and molecular biology of endosperm development (Guo *et al.*, 2003; Lai *et al.*, 2004). Neuffer and Sheridan (1980) estimated that at least 300 mutations specifically affect the endosperm phenotype, although only a small fraction of these have been characterized at the molecular level (Scanlon *et al.*, 1994; Scanlon and Meyers, 1998). The *Opaque-2* gene, one of the best characterized plant transcription factors, is a good example of the integration of carbohydrate, amino acid and storage protein metabolism. This gene regulates the expression of a set of enzymes involved in these metabolic pathways and therefore has a central role in endosperm development (Lohmer *et al.*, 1991; Schmidt *et al.*, 1992; Bass *et al.*, 1992; Habben *et al.*, 1993; Giroux *et al.*, 1994; Cord Neto *et al.*, 1995; Gallusci *et al.*, 1996; Arruda *et al.*, 2000; Hunter *et al.*, 2002). Recently, large databases of expressed genes have been made available for maize (http://www.maizegdb.org/est.php; http://genoplante-info.infobiogen.fr), and transcriptome analyses aimed at identifying the genes involved in endosperm development and metabolism have been published (Hunter *et al.*, 2002; Fernandes *et al.*, 2002; Guo *et al.*, 2003; Yu and Setter, 2003; Lai *et al.*, 2004).

We have created a large database of expressed maize genes, called MAIZEST (http://www.maizest.unicamp.br), that focuses on genes expressed in developing endosperm. The database was constructed by retrieving cDNA sequence-reads from MaizeGDB (http://www.maizegdb.org/est.php) and Génoplante (http://genoplante-info.infobiogen.fr) and subsequently enriching these with 30,531 new cDNA sequence-reads from 10, 15 and 20 DAP developing endosperm. Altogether, the MAIZEST database contains 227,431 maize ESTs, of which 64,537 came from developing endosperm. Bioinformatic tools were developed to help assembling and analyzing the endosperm transcriptome. In this report, we describe the analysis and annotation of endosperm-preferred expressed genes. Based on the large number of expressed sequences, we believe that the genes we have identified represent over 80% of the genes expressed in the endosperm, and that the endosperm-preferred set represents a significant contribution to understanding the molecular mechanisms underlying endosperm development and metabolism.

Materials and methods

Library construction

Field-grown maize (*Zea mays* L.) plants from the F352 inbred line (Kemper *et al.*, 1999) were self-pollinated and the ears were harvested at 10, 15 and 20 days after pollination (DAP). The upper third of the endosperms, containing only endosperm, aleurone and pericarp tissues, was removed, frozen in liquid nitrogen and stored at −80 °C. Total RNA was isolated from frozen developing endosperm as described by Manning (1991). Poly $(A)^+$RNA was purified from 500 μg of total RNA using Oligotex-dT (Qiagen) according to the manufacturer's instructions. The purity and integrity of the RNA were assessed by the absorbance at 260/280 nm and agarose gel electrophoresis. cDNA was synthesized using 1–5 μg of poly$(A)^+$RNA and directionally cloned into the pSPORT vector (Invitrogen) as described by Vettore *et al.* (2001). cDNAs ranging 500–800 bp (base pairs) in size were considered to be short libraries (S10, S15, S20), and those >800 bp were defined as long libraries (L10, L15, M15, N15, L20). Unamplified libraries were plated and individual colonies picked and transferred to 96 well plates containing liquid Circle Grow medium (Bio101), supplemented with 100 mg of ampicillin/l and 8% glycerol. The plates were stored at −80 °C.

cDNA Sequencing

DNA templates were prepared in 96-well plates in all stages, from bacterial growth through to purification after the sequencing reaction. DNA was prepared using a 96-well alkaline lysis method (http://sucest.lad.ic.unicamp.br/public). Sequencing reactions were done on plasmid templates using one-fourth of the standard volume of ABI Prism BigDye Terminator sequencing kits (Applied Biosystems) and the T7 promoter primer (5′-TAATACGACTCACTATAGGG-3′). The reaction products were precipitated with 95% ethanol using 3 M sodium acetate and glycogen (1 g/l) and the pellets were washed twice with 75% ethanol before drying under vacuum. The sequencing reaction products were analyzed using a 3700 ABI sequencer.

The new sequence data described in this paper have been submitted to Genbank under accession numbers CO439027–CO469579.

Database implementation and sequence analysis

All scripts used in trimming, assembly, sequence analysis and web interface were developed using Perl version 5.6.1 (http://www.cpan.org). The data were stored in an Oracle version 8.1.6 relational database (http://otn.oracle.com) and made available on the Web through the Apache 1.3.14 server (http://www.apache.org).

For ESTs generated in our laboratory, base calling and quality assessment were done using the Phred program (Ewing *et al.*, 1998). The trimming process, which involved the removal of ribosomal RNA, poly-A tails, low quality sequences, bacterial sequences and vector/adapter sequences, was done essentially as described by Telles and Da Silva (2001), with minor modifications. After trimming, the resulting ESTs had an average length of 776 bp and a minimum sequence length of 100 bp with a Phred quality ≤ 20. For ESTs available from the public databases MaizeGDB (http://www.maizegdb.org/est.php) and Génoplante (http://genoplante-info.infobiogen.fr), FASTA sequences were retrieved and base quality values were arbitrarily assigned: the first 30 bases received a Phred value of 15, the last 20 bases received a Phred value of 12 and the remaining bases received a Phred value of 20. Although they were below the average value obtained for ESTs generated in our laboratory, these quality values improved the accuracy of the EST assembling (data not shown). The CAP3 assembler (Huang and Madan, 1999) set to default parameters was used to assemble the ESTs. The assembled ESTs were referred to as Maize Assembled Sequences (MASs hereafter) and each consisted of a consensus sequence of a group of clustered ESTs. MASs can be either contigs, containing at least two ESTs, or can be singletons, formed by only one EST. Each MAS is likely to represent a transcript rather than a gene, allele or other biological entity, as discussed elsewhere (Telles *et al.*, 2001).

Annotation of all MASs was initially automated (GO evidence code IEA; http://www.geneontology.org/GO.evidence.html) by searching Swiss-Prot, and its computer-annotated supplement, the TrEMBL database (Boeckmann *et al.*, 2003; http://us.expasy.org/sprot/). The highest significant similarity score was used for provisional IEA annotation of the corresponding MAS following analysis of the BLASTX results, using a cutoff value of $E \leq 10^{-15}$. The protein name, BLASTX reports, descriptions, keywords and associated Gene Ontology terms (http://www.godatabase.org), if any, were compiled for each MAS entry. For the subset of MASs containing ESTs exclusively from endosperm cDNA libraries, curator-revised annotation (GO evidence code ISS) was done when the BLASTX hit against the NCBI nr database ($E \leq 10^{-5}$) resulted in an alignment length ≥50% of the maximum overlapping length between the query MAS and the NCBI entry (scheme in Figure 1 of supplemental material).

Expression profiling analysis

Three 96-well plates containing EST clones were randomly sampled from the 10, 15 and 20 DAP endosperm cDNA libraries. Additionally, a 96-well plate containing DNA of the empty plasmid vector pSPORT1 (Life Technologies, USA) was used as a negative hybridization control. The plasmid DNA was spotted onto nylon membranes and three replicate filters were produced containing 384 clones each. Total RNA from 10, 15 and 20 DAP endosperm, leaf and root of 7-days-old maize seedlings were isolated and used for probe synthesis. cDNA array hybridization and washing steps were performed essentially as described by Nogueira *et al.* (2003). The average and CV among the signal intensities of four replicated spots representing each EST spotted onto filters was estimated. The CV values were used to access the signal variation among replicate spots. The ESTs displaying CV values lower than 30% in all replicate filters were considered for analysis.

Results

Generation and assembly of maize ESTs

The MAIZEST database was constructed by integrating cDNA sequence-reads from three distinct sources: sequences generated in our laboratory, sequences retrieved from MaizeGDB (Gai *et al.*, 2000; Dong *et al.*, 2003; http://www.maizegdb.org/est.php) and sequences retrieved from Génoplante (Job, 2002, Samsom *et al.*, 2003; http://genoplante-info.infobiogen.fr). The information about tissue or organ used for cDNA library construction, as well as the number of sequences from each library from the three EST sources are shown in Table 1.

Sixty-seven cDNA libraries from different maize tissues, developmental stages or culture conditions were used. The data generated in our laboratory were derived from 41,450 cDNA 5′-end sequence-reads from standard, non-normalized, unidirectional cDNA libraries prepared from maize endosperm sampled at 10, 15 and 20 DAP. After trimming low quality and vector sequences and removing contaminant bacterial and ribosomal RNA sequences, the resulting data set contained 30,531 high-quality sequence-reads (>100 bp, Q20) with an average length of 776 bp. The tissue source information from MaizeGDB and Génoplante libraries, was retrieved from these two databases. Because we were interested in finding genes that were preferably expressed in developing endosperm, data from non-endosperm libraries that contained some endosperm-specific sequences were not included in the database. To exclude these libraries, we used the BLASTN tool (Altschul *et al.*, 1997) to screen the data set of each non-endosperm library for the presence of well-described, highly expressed endosperm-specific genes (Supplemental Table 1). In total, we retrieved 160,019 cDNA sequence-reads from MaizeGDB and 41,998 cDNA sequence-reads from Génoplante. The retrieved cDNA sequence-reads were trimmed for vector sequences, bacterial sequences and ribosomal RNA sequences, resulting in 196,900 validated cDNA sequence-reads. The MaizeGDB and Génoplante sequences were added to 30,531 cDNA sequence-reads from our laboratory, resulting in 227,431 ESTs. Of these, 64,537 originated from developing endosperm libraries.

CAP3 program (Huang and Madan, 1999) was used to assembly the 227,431 sequence-reads. A total of 217,665 sequence-reads were assembled into 19,440 contigs, while 9766 remained as singletons (Table 2). The combined set of contigs and singletons resulted in 29,206 sequences (hereafter referred to as MAS for Maize Assembled Sequence) representing putatively different transcripts. A search of the GenBank (Benson *et al.*, 2000) non-redundant protein database (cutoff BLASTX E value $\leq 10^{-5}$) indicated that approximately 68% of the MASs were similar to known protein sequences.

To estimate the level of redundancy among assembled sequences, the 29,206 MASs were compared to a set of 745 complete maize coding sequences (CDSs) retrieved from GenBank (Supplemental Table 2). Using a highly stringent selection parameter (BLASTN $E = 0.0$) and the requirement that a complete CDS had to cover at least 90% of the MAS extension, a total of 382 CDSs matched to 465 MAS sequences. This result suggested that there was approximately 17.8% redundancy among the MAS sequences. This level was in good agreement with the redundancy calculated for other large EST assemblages, e.g. 19.6% for *Apis mellifera* (Whitfield *et al.* 2003) and 22% for *Saccharum* spp. (Vettore *et al.* 2003), and indicates that MAIZEST may have identified around 24,000 expressed maize genes.

Identification of genes preferentially expressed in endosperm

The MAIZEST database was designed to provide tools for data storage and analysis. The assembling

Table 1. Description of the maize cDNA libraries and number of ESTs in the database.

Source	Library code	Description	No. of ESTs
PGL[a]	L10, S10	Endosperm harvested at 10 DAP	16,100
PGL	L15, M15, N15, S15	Endosperm harvested at 15 DAP	6387
PGL	L20, S20	Endosperm harvested at 20 DAP	8044
MaizeGDB[b]	CC1	Mixed logarithmic and stationary growth phases of suspension culture in BMS	581
MaizeGDB	CC2	Mixed logarithmic and stationary growth phases of suspension culture in BMS	13,264
MaizeGDB	EA4	Field-grown unpollinated ears silk channel-inoculated with *F. graminearum*	628
MaizeGDB	EA5	2 mm ear	19,082
MaizeGDB	EM2	Embryos harvested at 14 DAP	1088
MaizeGDB	EN1	Kernel endosperm	6506
MaizeGDB	EN2	Membrane-free polysomes from endosperm	609
MaizeGDB	EN3	Endosperm harvested at 7–23 DAP	1075
MaizeGDB	EN4	Endosperm harvested at 4–6 DAP	96
MaizeGDB	EN5	Endosperm harvested at 7–23 DAP	6389
MaizeGDB	EN6	Endosperm harvested at 7–23 DAP	10,092
MaizeGDB	EN7	Endosperm harvested at 7–23 DAP	4309
MaizeGDB	EN8	Endosperm harvested at 7–23 DAP	909
MaizeGDB	ES1	Embryonic sacs isolated with enzymatic maceration and manual micro dissection	368
MaizeGDB	GL1	Glume (2 weeks post-pollination)	2125
MaizeGDB	IN1	Developing female inflorescence	468
MaizeGDB	LF1	Immature leaf primordium and vegetative meristem	10,340
MaizeGDB	LF2	Shoot leaf primordia	5615
MaizeGDB	LF3	Illuminated leaves and sheaths of 5-week-old plant	829
MaizeGDB	MR1	Apical meristem from immature shoot	676
MaizeGDB	PA1	Whole premeiotic anthers to pollen shed	6366
MaizeGDB	PO1	Mature pollen	3916
MaizeGDB	PO2	Mature pollen	413
MaizeGDB	RT1	3–4-day-old root tissue	10,487
MaizeGDB	RT2	2-week-old roots stressed for 24 hours at 150 mM NaCl	483
MaizeGDB	RT3	Stressed seedling root	1981
MaizeGDB	SC1	Sperm cells sorted by fluorescent-activation	2048
MaizeGDB	SH1	Leaf and stem, including leaf base from 2-week-old seedling	8628
MaizeGDB	SH2	Stressed seedling shoot	1250
MaizeGDB	SK1	Silk channel of field-grown corn inoculated with *F. graminearum*	706
MaizeGDB	SL1	Seedling and silk	606
MaizeGDB	SL2	Cold stressed leaf and crown	589
MaizeGDB	SL3	Seedling and silk	8958
MaizeGDB	SL4	Cold stressed leaf and crown (seedlings at 4-leaf stage)	900
MaizeGDB	TA1	Immature tassels after transition from vegetative to inflorescence development	20,674
MaizeGDB	TA2	Tassels (length from 0.1 to 2.5 cm)	3348
Genoplante[c]	AL1	3rd adult leaf	1663
Genoplante	AL2	3rd adult leaf	1753
Genoplante	AL3	3rd adult leaf	1007
Genoplante	AL4	3rd adult leaf	2293
Genoplante	CD1	Cell division (part of the 6th leaf)	2200
Genoplante	CL1	Cell lignification (part of the 6th leaf)	1994
Genoplante	EM3	Embryo	2066
Genoplante	NE1	Endosperm	2377
Genoplante	NE2	Endosperm	1644
Genoplante	OV1	Ovary	683
Genoplante	PD1	Pedicel, whole kernel	691
Genoplante	PR1	Pericarp	4216

Table 1. (Continued).

Source	Library code	Description	No. of ESTs
Genoplante	PR2	Pericarp	3200
Genoplante	PR3	Pericarp	1871
Genoplante	PR4	Pericarp	835
Genoplante	RE1	Root extremities	581
Genoplante	RE2	Root extremities	550
Genoplante	RE3	Root extremities	1198
Genoplante	SM1	Seedling minus kernel	1049
Genoplante	SM2	Seedling minus kernel	1850
Genoplante	SM3	Seedling minus kernel	1991
Genoplante	SM4	Seedling minus kernel	1781
Genoplante	ST1	Sheath	3005
Total			227,431

[a] Plant Genome Laboratory, Brazil (http://est.cbmeg.unicamp.br/pgl).
[b] MaizeGDB project, USA (http://www.maizegdb.org/est.php).
[c] Génoplante, France (http://genoplante-info.infobiogen.fr).

tools allowed the analysis of cluster distribution among libraries, and made it possible to infer the likelihood of tissue-specific expression. Interactive tools provide ways of data mining by using refined searches. Other tools, such as 'virtual northern', are available and allow the estimation of gene expression levels between different tissues and organs, or within the endosperm, at distinct developmental stages. The statistical significance of the digital analysis is tested as described by Audic and Claverie (1997). Direct access to the database is achieved through the 'database query' tool which implements a default SQL interface that improves the capabilities for complex data mining. The combined MAS set represents a large and diverse collection of transcripts from genes expressed in different maize tissues and also constitutes an endosperm-enriched database for gene discovery and expression analysis. The MAIZEST database contains 64,537 ESTs that were generated from cDNA libraries prepared from endosperm tissue (Table 1; Supplemental Figure 2). A search of the 29,206 MASs showed that 13,457 MASs (~46%) contained at least one EST derived from endosperm cDNA libraries (Table 3). By assuming a redundancy of 17.8% in this set, we estimated that around 11,000 genes expressed in the developing endosperm were identified. This number, which includes genes that are expressed in the endosperm as well as in other tissues, is twice that recently reported by Lai *et al.* (2004). A search for MAS preferentially expressed in developing endosperm revealed a subset of 4032 MASs consisting of ESTs derived exclusively from endosperm libraries (Table 3). Because of the large amount of ESTs originating from developing endosperm and from other vegetative tissues, this value is a good estimate of genes preferentially expressed in endosperm. Of the 4032 endosperm-preferred MASs, 2794 were singletons and 1238 formed contigs. Singletons are genes expressed at a very low level and it is difficult to determine whether they are expressed in other tissues. However, we preferred to maintain these singletons in the class of endosperm-preferred genes because of the large number of endosperm and non-endosperm libraries analyzed. Schmid et al. (2005),

Table 2. MAIZEST EST summary.

Total number of sequences entering database	243,457
Source: PGL-Campinas[a]	41,450
Source: MaizeGDB[b]	160,019
Source: Génoplante[c]	41,988
Total number of validated sequences	227,431
Sequences in contigs	217,665
Total number of singletons	9766
Total number of contigs[d]	19,440
Total number of MASs[e]	29,206
MASs matching GenBank nr entries[f]	19,944
Average size of validated sequences (bp)	511

[a] Plant Genome Laboratory, Brazil (http://est.cbmeg.unicamp.br/pgl).
[b] MaizeGDB, USA (http://www.maizegdb.org/est.php).
[c] Génoplante, France (http://genoplante-info.infobiogen.fr).
[d] ESTs were assembled using CAP3.
[e] MASs, Maize Assembled Sequences, are the combined sets of contigs and singletons from the three sequencing sources.
[f] A BLASTX match cutoff of $E \leq 10^{-5}$ was used to assign similarity.

Table 3. In silico selection of endosperm-preferred MASs.

Tissue	Total ESTs	Singletons	Contigs	ESTs in contigs	MASs
All tissues[a]	227,431	9766	19,440	217,665	29,206
Endosperm-only MASs[b]	–	2794	1238	19,614	4032
At least one EST from endosperm[c]	–	2794	10,600	167,280	13,457

[a] All ESTs entering database, generated from libraries depicted in Table 1, were clusterized using CAP3.

[b] After clustering of all of the ESTs generated from all tissues in the three projects, MASs containing only ESTs from endosperm libraries were selected using the SQL query.

[c] All MASs containing at least one EST generated from endosperm were selected using the SQL query.

analyzing the expression of over 22,000 Arabidopsis genes in 79 different samples, found an average of 92% overlap of transcripts among the tissues. A fraction of 4.4–11.6% of the 22,000 Arabidopsis genes was found as being specific for a particular tissue. This number is in good agreement with the 13.8% of endosperm-preferred genes we found in the MAIZEST database.

Genes expressed at a low level, including regulatory genes, play key roles in tissue development and metabolism. The genes preferentially expressed in endosperm included 118 transcription factors and 76 genes encoding proteins involved in signaling processes (data not shown). The complete set of preferentially expressed endosperm genes is provided as online information (Supplemental Table 3). In order to access the accuracy of the *in silico* identified endosperm-preferred genes, we performed an expression profile analysis of 288 ESTs randomly chosen from 10, 15 and 20 DAP cDNA libraries. These ESTs were hybridized with ^{33}P-labeled RNA from leaf, root and endosperm tissue (Supplemental Figure 3). The 288 ESTs used in the expression profiling analysis correspond to 174 MASs. Among these, there were 47 (27%) classified by the *in silico* analysis as endosperm-preferred. The 47 endosperm-preferred MASs sampled in the filters are formed by 92 ESTs. Among the 288 ESTs spotted in the membranes, 89 presented a significant endosperm-preferred profile. All these ESTs are among those classified as endosperm-preferred in the *in silico* analysis. Only 3 ESTs classified as endosperm-preferred in the *in silico* analysis didn't demonstrate significant tissue expression difference. These 3 ESTs represent singletons. Figure 1 shows few examples of typical endosperm specific genes and other proteins with endosperm-preferred and non-preferred expression profiling as determined by *in silico* analysis and high density membrane array.

Functional annotation of endosperm-preferred MASs

Provisional annotation of the entire endosperm-preferred MASs set was inferred from electronic annotation by searching the Swiss-Prot/TrEMBL database. For those MASs matching the database, GO terms were assigned based on the highest significant similarity score ('best hit') using a cutoff value of $E \leq 10^{-15}$. From the 4032 endosperm-preferred MASs, a sub-set of 2403 MASs was functionally annotated by curators after evaluation through a series of *in silico* comparisons, as described in Material and Methods and illustrated in Figure 1 of supplemental material.

Figures 2 and 3 summarize the assignments of the 2403 MASs to major biological processes and molecular functions, respectively. Examination of the biological processes shown in Figure 2 revealed that a significant portion of the expressed transcripts is involved in cellular and metabolic processes associated with endosperm metabolism, such as cell division and growth, high rates of DNA replication within the cell, and amino acid and sugar transport, the latter being intrinsically linked to the accumulation of storage proteins and starch (Lopes and Larkins, 1993; Olsen, 2004). In addition, key processes in organ development, such as regulation of the cell cycle, partitioning of growth between cell division and cell expansion, regulation of cell expansion and terminal differentiation, cell-to-cell signaling, and determination of cell fate, may be related to significant cellular processes assigned in the functional annotation (Figure 2, outward blue sections). The transcripts related to those processes required for cell survival and growth include transport (5.3%), cell proliferation (2.3%) and cell communication (2.0%). Among physiological processes (Figure 2, green

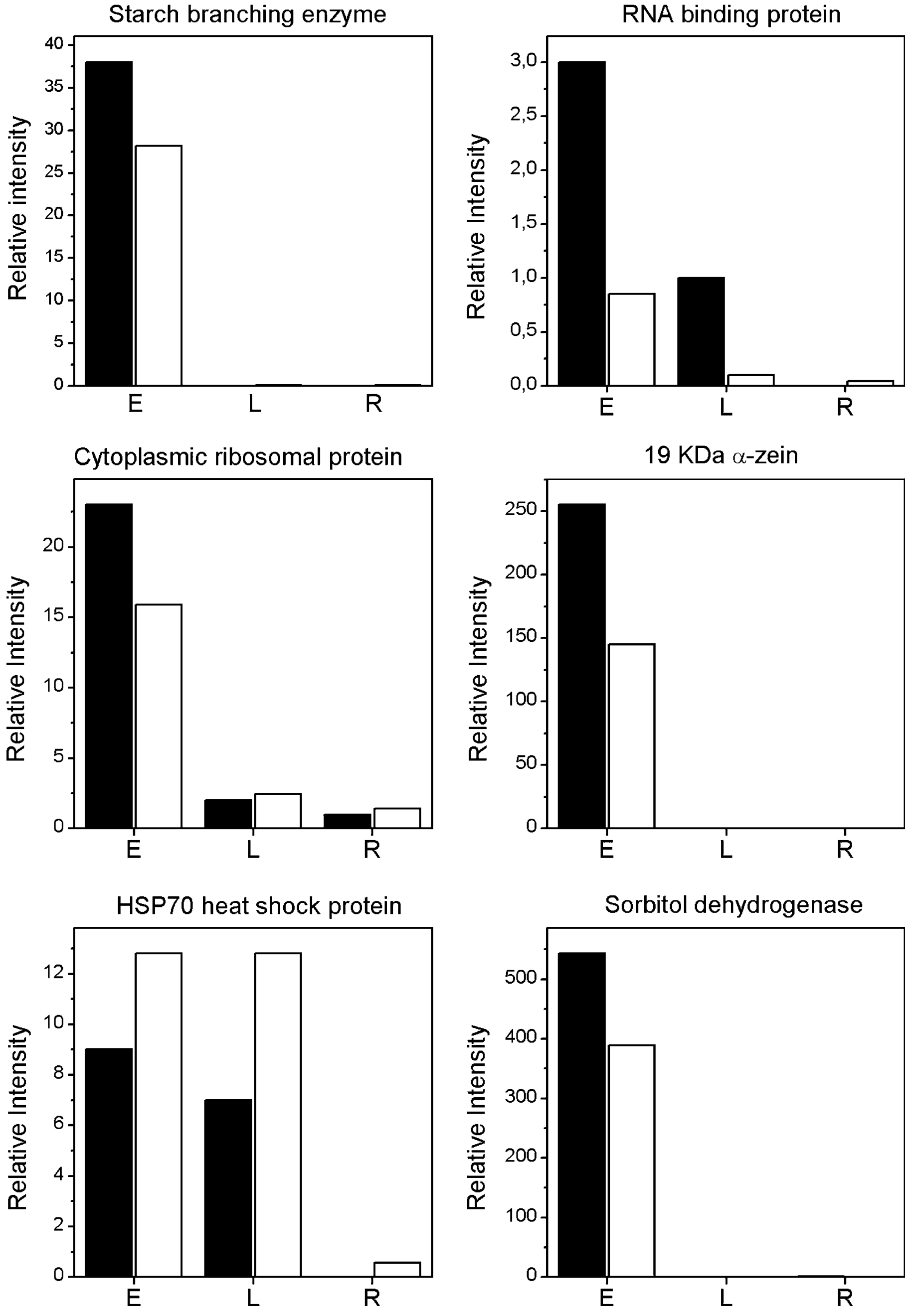

Figure 1. Examples of the expression profile of endosperm-preferred and not-preferred maize ESTs. Virtual Northern (black bars) and expression profile of high-density membrane arrays (open bars). E: Endosperm, L: Leaf and R: Root.

sections), those transcripts implicated in protein metabolism (16.8%), nucleic acid metabolism (11.1%), phosphorus metabolism (3.6%) and carbohydrate metabolism (3.3%), were successfully assigned. Nevertheless, large portions (ca. 36%) of the transcripts remained unassigned.

Annotation of the MASs with respect to molecular function also consistently revealed an array of gene functions most likely involved in endosperm development. As shown in Figure 3, transcripts putatively encoding transporters accounted for 5.6% of the MASs preferentially

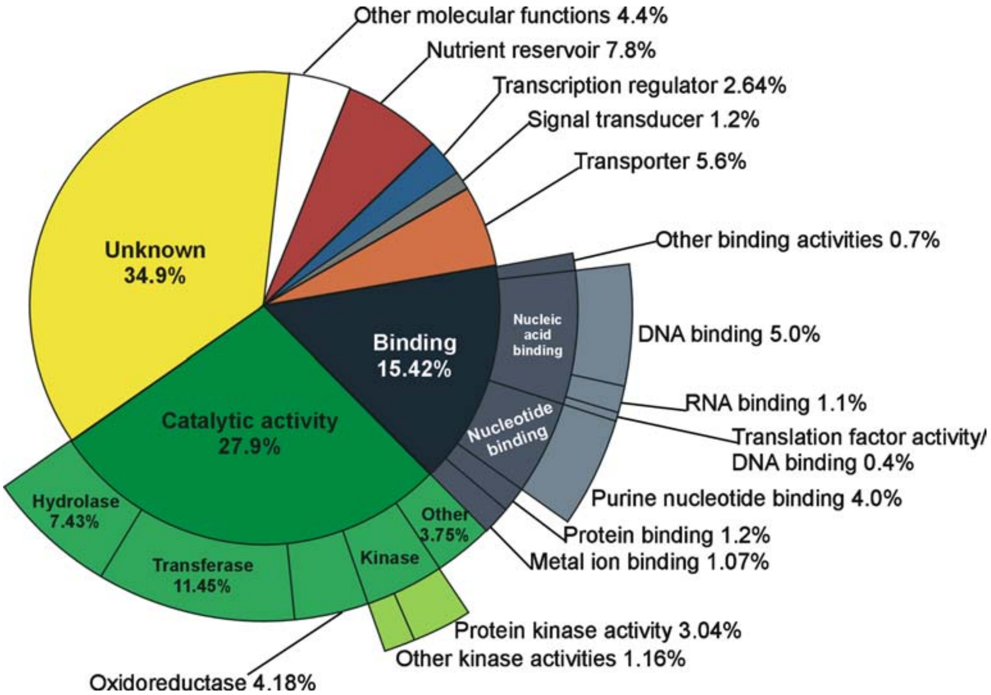

Figure 2. Maize endosperm gene prediction: biological process. Gene Ontology categories were assigned to MASs through curator-revised categorization. Classification is hierarchical, as children categories progress outwards from the inner parental categories. Two thousand four hundred and three endosperm-preferred MASs were classified. Gene Ontology terms (http://www.geneontology.org) were assigned based on similarity to known protein sequences in several databases (GenBank nr, http://www.ncbi.nlm.nih.gov/Genbank/; SwissProt/TrEmbl, http://us.expasy.org/sprot/; TRANSFAC 6.3 and Transpath 3.3, http://www.gene-regulation.com/) using a BLASTX cutoff value of $E \leq 10^{-5}$. The percentage of MASs in each category is indicated next to the corresponding map sector. The 'unknown' category includes MASs that matched to 'unknown protein', 'putative protein' or 'hypothetical protein', with no indication of the corresponding function. The total sum of the percentages did not add to 100% because MASs may be assigned to more than one category or child categories may have more than one parental category (See Gene Ontology Consortium at http://www.geneontology.org/GO.nodes.html).

expressed in endosperm, while nutrient reservoir activity (the zein family of storage proteins) represented 7.8%, and nucleotide and nucleic acid binding accounted for up to 9.0%. The assignment of other important classes of transcripts, such as transcription regulators (2.6%; mostly representing transcription factors) and signal transducers (1.2%) provides new perspectives for data mining and for studies of coordinated gene regulation in developing maize endosperm.

Discussion

By integrating large amounts of EST data generated from developing endosperm cDNA libraries with data generated from cDNA libraries of vegetative tissues, we have obtained a broader view of the possible set of genes expressed in endosperm. *In silico* comparisons uncovered a number of genes that can be specifically targeted in future functional genomic studies. Such an approach should advance our knowledge of the genes and functions underlying maize endosperm development.

In this work, we focused on a comparative analysis of ESTs from endosperm and non-endosperm cDNA libraries. The addition of over 30,000 new cDNA sequence-reads from developing endosperm created one of the largest publicly available databases of endosperm expressed genes. The novelty of this new collection of endosperm ESTs is that of 4032 endosperm-preferred MASs, 1962 were formed exclusively by ESTs from our laboratory while 1637 were from MaizeGDB and 80 were from Génoplante. Another important aspect was the diversity of sequences representing the developing endosperm. Even if mRNAs

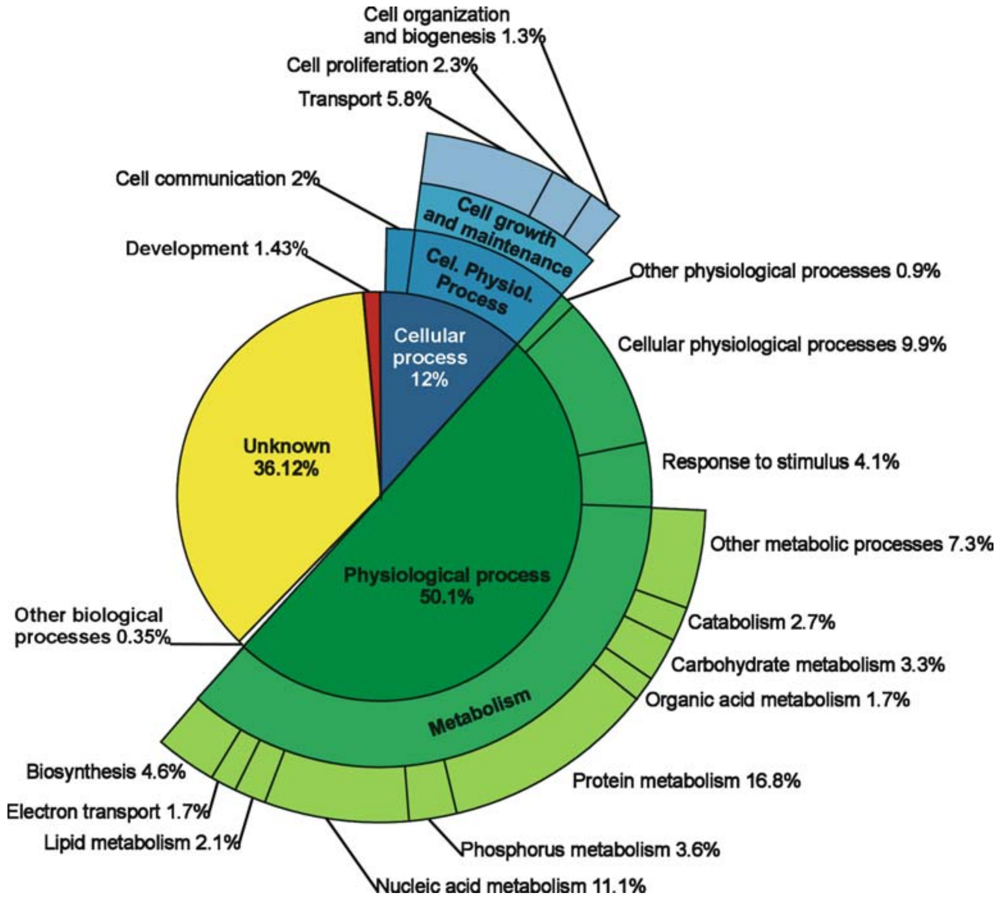

Figure 3. Maize endosperm gene prediction: molecular function. Gene Ontology categories were assigned to MASs through curator-revised categorization. Classification is hierarchical, as children categories progress outwards from the inner parental categories. Two thousand four hundred and three endosperm-preferred MASs were classified. Gene Ontology terms (http://www.geneontology.org) were assigned based on similarity to known protein sequences in several databases (GenBank nr, http://www.ncbi.nlm. nih.gov/Genbank/; SwissProt/TrEmbl, http://us.expasy.org/sprot/; TRANSFAC 6.3 and Transpath 3.3, http://www.gene-regulation.com/) using a BLASTX cutoff value of $E \leq 10$. The percentage of MASs in each category is indicated next to the corresponding map sector. The 'unknown' category includes MASs that matched to 'unknown protein', 'putative protein' or 'hypothetical protein', with no indication of the corresponding function. The total sum of the percentages did not add to 100% because MASs may be assigned to more than one category or child categories may have more than one parental category (See Gene Ontology Consortium at http://www.geneontology.org/GO.nodes.html).

> encoding zein genes account for over 60% of the mRNA pool of the endosperm during periods of high storage protein synthesis (see for example, Woo *et al*., 2001), a large portion of non-zein transcripts is present in the database. In fact, since most of the cDNA sequence-reads from our laboratory came from 10 and 15 DAP cDNA libraries, we have sequenced only 8,468 zein cDNAs out of 30,553 (ca. 27%). As a result, we have contributed considerably to the diversity of this database with respect to genes expressed in the endosperm. On the other hand, the non-endosperm sequences came from a large and diverse set of vegetative tissues and represented nearly two-thirds of the total data set. If the number of genes in maize is similar to that of rice, which is estimated to be around 40,000 genes (Yu *et al*., 2002; Lai *et al*., 2004), then the 24,000 putative genes identified here represent ~60% of the maize genes.

In assembling over 60,000 endosperm sequence-reads, we assumed that we had possibly identified ca. 11,000 genes expressed in the endosperm. This number is in good agreement with a recent report by Lai *et al*. (2004), who assembled ca. 24,000 endosperm sequence-reads into 5326 putative expressed genes. Similarly, the search for

MASs containing at least one EST derived from MaizeGDB libraries revealed 5,887 MASs. Hence, the large amount of information compiled in the MAIZEST database provides a good opportunity for studying the regulatory processes governing endosperm development and metabolism. As an example, a search for MASs encoding regulatory genes revealed that of the 11,000 putative genes expressed in the developing endosperm, 365 represent putative transcription factors, and 118 of these were preferentially expressed in endosperm (Verza, et al., in preparation). This information is of interest if considered along with the studies related to the *opaque-2* maize mutant. The expression profile of an opaque-2 endosperm has revealed that a number of genes encoding enzymes involved in amino acid and carbohydrate metabolism, as well as genes encoding storage proteins are downregulated (Hunter *et al.*, 2002). The *Opaque-2* gene encodes a b-ZIP transcription factor that regulates the expression of a set of enzymes involved in these metabolic pathways (Lohmer *et al.*, 1991; Schmidt *et al.*, 1992; Habben *et al.*, 1993; Giroux *et al.*, 1994; Gallusci *et al.*, 1996; Arruda *et al.*, 2000) and it is supposed to play a central role in endosperm development. Therefore, it will be interesting to clarify the interactions among Opaque-2 and those other 118 putative transcription factors.

Acknowledgments

The authors thank Almir S. Zanca, Ana L. Beraldo and Renato V. dos Santos for technical assistance, and Dr. Edson L. Kemper for technical and scientific advice and critical reading of the manuscript. N.C.V. was supported by a postgraduate fellowship from Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior (CAPES), and T.R.S., F.T.S.N., and V.E.R. Jr were supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Arruda, P., Kemper, E.L., Papes, F. and Leite, A. 2000. Regulation of lysine catabolism in higher plants. Trends Plant Sci. 5: 324–330.

Audic, S. and Claverie, J.M. 1997. The significance of digital gene expression profiles. Genome Res. 7: 986–995.

Bass, H.W., Webster, C., Obrian, G.R., Roberts, J.K.M. and Boston, R.S. 1992. A maize ribosome-inactivating protein is controlled by the transcriptional activator Opaque2. Plant Cell 4: 225–234.

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

Berger, F. 1999. Endosperm development. Curr. Opin. Plant. Biol. 2: 28–32.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. and Schneider, M. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31: 365–370.

Cord Neto, G., Yunes, J.A., Vettore, A.L., Da Silva, M.J., Arruda, P. and Leite, A. 1995. The involvement of Opaque2 on β-prolamin gene regulation in maize and Coix suggests a more general role for this transcriptional activator. Plant Mol. Biol. 27: 1015–1029.

Dong, Q., Roy, L., Freeling, M., Walbot, V. and Brendel, V. 2003. ZmDB, an integrated database for maize genome research. Nucleic Acids Res. 31: 244–247.

Ewing, B., Hillier, L.A., Wendl, M.C. and Green, P. 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Res. 8: 175–185.

Fernandes, J., Brendel, V., Gai, X., Lal, S., Chandler, V.L., Elumalai, R.P., Galbraith, D.W., Pierson, E.A. and Walbot, V. 2002. Comparison of RNA expression profiles based on maize-expressed sequence tag frequency analysis and micro-array hybridization. Plant Physiol. 128: 896–910.

Gai, X., Lal, S., Xing, L., Brendel, V. and Walbot, V. 2000. Gene discovery using the maize genome database ZmDB. Nucleic Acids Res. 28: 94–96.

Gallusci, P., Varott, S., Matsuoko, M., Maddaloni, M. and Thompson, R.D. 1996. Regulation of cytosolic pyruvate, orthophosphate dikinase expression in developing maize endosperm. Plant Mol. Biol. 31: 45–55.

Giroux, M.J., Boyer, C., Feix, G. and Hannah, L.C. 1994. Coordinated transcriptional regulation of storage product genes in the maize endosperm. Plant Physiol. 106: 713–722.

Guo, M., Rupe, M.A., Danilevskaya, O.N., Yang, X. and Hu, Z. 2003. Genome-wide mRNA profiling reveals heterochronic allelic variation and a new imprinted gene in hybrid maize endosperm. Plant J. 36: 30–44.

Habben, J.E., Kirleis, A.W. and Larkins, B.A. 1993. The origin of lysine-containing proteins in the opaque-2 maize endosperm. Plant Mol. Biol. 23: 825–838.

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

Hunter, B.G., Beatty, M.K., Singletary, G.W., Hamaker, B.R., Dilkes, B.P., Larkins, B.A. and Jung, R. 2002. Maize opaque endosperm mutations create extensive changes in patterns of gene expression. Plant Cell. 14: 2591–2612.

International Human Genome Sequencing Consortium 2004. Finishing the euchromatic sequence of the human genome. Nature 431: 931–945.

Job, D. 2002. Génoplante: the French national network in plant genomics. GenomXPress 4: 13–17.

Kemper, E.L., Cord Neto, G., Papes, F., Moraes, K.C.M., Leite, A. and Arruda, P. 1999. The role of Opaque2 in the control of lysine-degrading activities in developing maize endosperm. Plant Cell 11: 1981–1993.

Lai, J., Dey, N., Kim, C.S., Bharti, A.K., Rudd, S., Mayer, K.F., Larkins, B.A., Becraft, P. and Messing, J. 2004. Characterization of the maize endosperm transcriptome and its comparison to the rice genome. Genome Res. 14: 1932–1937.

Lohmer, S., Maddaloni, M., Motto, M., Di Fonzo, N., Hartings, H., Salamini, F. and Thompson, R.D. 1991. The maize regulatory locus Opaque-2 encodes a DNA-binding protein which activates the transcription of the b-32 gene. EMBO J. 10: 617–624.

Lopes, M.A. and Larkins, B.A. 1993. Endosperm origin, development, and function. Plant Cell 5: 1383–1399.

Maizel, A. and Weigel, D. 2004. Temporally and spatially controlled induction of gene expression in Arabidopsis thaliana. Plant J. 38: 164–171.

Manning, K. 1991. Isolation of nucleic acids from plants by differential solvent precipitation. Anal. Biochem. 195: 45–50.

Muller, M., Dues, G., Balconi, C., Salamini, F. and Thompson, R.D. 1997. Nitrogen and hormonal responsiveness of the 22 kDa alpha-zein and b-32 genes in maize endosperm is displayed in the absence of the transcriptional regulator Opaque-2. Plant J. 12: 281–291.

Neuffer, M.G. and Sheridan, W.F. 1980. Defective kernel mutants of maize I Genetic and lethality studies. Genetics 95: 929–944.

Nogueira, F.T.S., De Rosa, Jr., , V.E., Menossi, M., Ulian, E.C. and Arruda, P. 2003. RNA expression profiles and data mining of sugarcane response to low temperature. Plant Physiol. 132: 1811–1824.

Olsen, O-A. 2004. Nuclear endosperm development in cereals and Arabidopsis thaliana. Plant Cell 16: S214–S227.

Samson, D., Legeai, F., Karsenty, E., Reboux, S., Veyrieras, J-B., Just, J. and Barillot, E. 2003. GénoPlante-Info (GPI): a collection of databases and bioinformatics resources for plant genomics. Nucleic Acids Res 31: 179–182.

Scanlon, M.J., Stinard, PS., James, M.G., Myers, A.M. and Robertson, D.S. 1994. Genetic analysis of 63 mutations affecting maize kernel development isolated from Mutator stocks. Genetics 136: 281–294.

Scanlon, M.J. and Myers, A.M. 1998. Phenotypic analysis and molecular cloning of discolored-1 (dsc1), a maize gene required for early kernel development. Plant Mol. Biol. 37: 483–493.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Scholkopf, B., Weigel, D. and Lohmann, J.U. 2005. A gene expression map of Arabidopsis thaliana development. Nat. Gen. 37(5): 501–6.

Schmidt, R.J., Ketudat, M., Aukerman, M.J. and Hoschek, G. 1992. Opaque2 is a transcriptional activator that recognizes a specific target site in 22 kDa zein gene. Plant Cell 4: 689–700.

Stolc, V., Gauhar, Z., Mason, C., Halasz, G., van Batenburg, M.F., Rifkin, S.A., Hua, S., Herreman, T., Tongprasit, W., Barbano, P.E., Bussemaker, H.J. and White, K.P. 2004. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. Science 306: 655–660.

Telles, G.P., Braga, M.D.V., Dias, Z., Lin, T-L., Quitzau, J.A.A., da Silva, F.R. and Meidanis, J. 2001. Bioinformatics of the sugarcane EST project. Genet Mol. Biol. 24: 9–15.

Telles, G.P. and da Silva, F.R. 2001. Trimming and clustering sugarcane ESTs. Genet. Mol. Biol. 24: 17–23.

Tuteja, J.H., Clough, S.J., Chan, W.C. and Vodkin, L.O. 2004. Tissue-specific gene silencing mediated by a naturally occurring chalcone synthase gene cluster in Glycine max. Plant Cell 16: 819–835.

Velculescu, V.E., Madden, S.L., Zhang, L., Lash, A.E., Yu, J., Rago, C., Lal, A., Wang, C.J., Beaudry, G.A., Ciriello, K.M., Cook, B.P., Dufault, M.R., Ferguson, A.T., Gao, Y., He, T.C., Hermeking, H., Hiraldo, S.K., Hwang, P.M., Lopez, M.A., Luderer, H.F., Mathews, B., Petroziello, J.M., Polyak, K., Zawel, L. and Kinzler, K.W. 1999. Analysis of human transcriptomes. Nat. Genet. 23: 387–388.

Vettore, A.L., da Silva, F.R., Kemper, E.L. and Arruda, P. 2001. The libraries that made SUCEST. Genet. Mol. Biol. 24: 1–7.

Vettore, A.L., da Silva, F.R., Kemper, E.L., Souza, G.M., da Silva, A.M., Ferro, M.I., Henrique-Silva, F., Giglioti, E.A., Lemos, M.V. and Coutinho, L.L. 2003. Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. Genome Res. 13: 2725–2735.

Whitfield, C.W., Band, M.R., Bonaldo, M.F., Kumar, C.G., Liu, L., Pardinas, J.R., Robertson, H.M., Soares, M.B. and Robinson, G.E. 2002. Annotated expressed sequence tags and cDNA microarrays for studies of brain and behavior in the honey bee. Genome Res. 12: 555–566.

Woo, Y.M., Hu, D.W.N., Larkins, B.A. and Jung, R. 2001. Genomics analysis of genes expressed in maize endosperm identifies novel seed proteins and clarifies patterns of zein gene expression. Plant Cell 13: 2297–2317.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y. and Zhang, X. 2002. A draft sequence of the rice genome (*Oryza sativa* L ssp. indica). Science 296: 79–92.

Yu, L. and Setter, T.L. 2003. Comparative transcriptional profiling of placenta and endosperm in developing maize kernels in response to water deficit. Plant physio. 131: 568–582.