UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE BIOLOGIA
DEPARTAMENTO DE GENÉTICA E EVOLUÇÃO

Natalia Cristina Verza Ferreira

"ANÁLISE, CLASSIFICAÇÃO, ANOTAÇÃO E PERFIL DE EXPRESSÃO DE FATORES DE TRANSCRIÇÃO NO ENDOSPERMA DE MILHO (*Zea mays* L.)"

Tese apresentada ao Instituto de Biologia para obtenção do Título de Doutor em Genética e Biologia Molecular na área de Genética Vegetal e Melhoramento.

Orientador: Prof. Dr. Paulo Arruda

CAMPINAS/SP
2006

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DO INSTITUTO DE BIOLOGIA – UNICAMP

F413a

Ferreira, Natalia Cristina Verza
Análise, classificação, anotação e perfil de expressão de fatores de transcrição no endosperma de milho (*Zea mays* L.) / Natalia Cristina Verza Ferreira. -- Campinas, SP: [s.n.], 2006.

Orientador: Paulo Arruda.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. Milho – Genética. 2. Endosperma. 3. Aleurona. 4. Transcriptoma. 5. Fatores de transcrição. I. Arruda, Paulo. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

(rcdt/ib)

Título em inglês: Analysis, classification, annotation and expression pattern of transcription factors in maize (*Zea mays* L.) endosperm.
Palavras-chave em inglês: Maize – Genetics; Endosperm; Aleurone; Transcriptome; Transcription factors.
Área de concentração: Genética Vegetal e Melhoramento.
Titulação: Doutora em Genética e Biologia Molecular.
Banca examinadora: Paulo Arruda, Isabel Rodrigues Gerhardt, Jörg Kobarg, Jesus Aparecido Ferro, Elíbio Leopoldo Rech.
Data da defesa: 12/05/2006.

BANCA EXAMINADORA

Campinas, 12 de maio de 2006.

Prof. Dr. Paulo Arruda (orientador) _____
Assinatura

Prof. Dr. Jörg Kobarg _____
Assinatura

Profa. Dra. Isabel Gerhardt _____
Assinatura

Prof. Dr. Jesus Aparecido Ferro _____
Assinatura

Prof. Dr. Elíbio Leopoldo Rech Filho _____
Assinatura

Profa. Dra. Anete Pereira de Souza _____
Assinatura

Prof. Dr. Michel Georges Albert Vincentz _____
Assinatura

Profa. Dra. Andrea Almeida Carneiro _____
Assinatura

...la Madre del Maíz cambió su forma de paloma y adoptó la humana; le presento al muchacho sus cinco hijas, que simbolizan los cinco colores sagrados del maíz: blanco, rojo, amarillo, moteado y azul. Como el joven tenía hambre, la Madre del Maíz le dio una olla llena de tortillas y una jícara llena de atole; él no creía que eso pudiera saciar su hambre, pero las tortillas y el atole se renovaban mágicamente, de manera que no podía acabárselos. La Madre del Maíz le pidió que escogiera a una de sus hijas y él tomó a la Muchacha del Maíz Azul, la más bella y sagrada de todas...

Lenda huichol que fala sobre a seleção antropogênica realizada por esta nação indígena pré-colombiana com o milho.

(Furst, Peter T. y Nahmad, Salomón. Mitos y arte huicholes. México, Secretaría de Educación Pública (Col. Septentas, 50), 1972).

"Não haverá parto se a semente não for plantada, muito tempo antes... Não haverá borboletas se a vida não passar por longas e silenciosas metamorfoses..."

Rubem Alves

Aos meus pais.

Eu não poderia ter sido mais abençoada...

AGRADECIMENTOS

Ao meu orientador, Paulo Arruda, por ter acreditado que eu seria capaz de desenvolver esse projeto apenas com a bagagem que trouxe da graduação, algumas idéias e muita vontade. Obrigada por ter me mostrado a importância do planejamento, condição para o sucesso do experimento. Obrigada por ter sempre me disponibilizado tudo o que precisei, de reagentes a contatos. Obrigada por estar sempre presente, por telefone, e-mail, nos finais de semana. Aprendi muito, e não apenas em ciência, pela convivência com você.

Aos membros da pré-banca, Prof. Michel Vincentz (meu primeiro orientador!) e Dra. Isabel Gerhardt, pelas idéias. Aos membros da banca, Prof. Jörg Kobarg, Dra. Isabel Gerhardt, Prof. Jesus Ferro, Dr. Elíbio Rech, Profa. Andréa Carneiro, Prof. Michel Vincentz e Profa. Anete Pereira de Souza por terem aceito o meu convite para participar da banca de defesa. Me sinto muito honrada com a participação de todos vocês.

Ao Dr. André Vettore, do Instituto Ludwig, por ter me ensinado muito do que sei de biologia molecular, por ter me orientado durante a iniciação científica, e depois, mesmo de longe, por toda a ajuda que me deu durante o doutorado. Você é o meu exemplo de cientista e de ser humano. Quando crescer quero ser como você.

Aos meus queridos amigos Sylvia Morais de Sousa, Thaís Rezende e Silva e Mário del Giúdice Paniago (Careca), que foram a "massa crítica" que tive para discutir as coisas que deram errado e para comemorar as que deram certo. Obrigada pela companhia, pelas conversas fúteis, pelas "bolinhas" e pelas altas discussões científicas. Aos queridos Vicente Eugênio de Rosa Jr. e Fábio Tebaldi Nogueira, que me iniciaram nos mistérios do *macroarray*, e me ajudaram tanto no tempo em que convivemos no Laboratório de Genômica de Plantas. Desejo a vocês dois muito sucesso.

Ao Eduardo Kiyota (Dudu), e às meninas do Genoma, Daniela, Heidi, Fabi e Elane, sempre quebrando meus galhos. E também ao Márcio José da Silva, meu consultor para assuntos diversos. Obrigada aos amigos que passaram pelo Genoma, e que deixaram muitas saudades, Almir, Ana e Gabriel. À Letícia Bonatelli, minha aluna de iniciação. Espero que você tenha persistência para continuar seu caminho na ciência.

Ao Sr. Bueno, do SENAI de Betel, pela mágica de sempre me arrumar um pouquinho de milho, mesmo quando não era época.

À todos os alunos do Curso de Férias e da BG581 por terem me dado a oportunidade de aprender mais. Obrigada à Silvia Regina Turcinelli, pelas conversas, risadas, e pelo apoio que me deu em alguns momentos bem difíceis ao longo desses quatro anos.

Ao Fábio Papes, ao Germano, ao Andrés, à Adriana Capella, ao Edson Kemper, ao Ivan Maia e à Jaqueline, ao Celso Benedetti, pela convivência e pelos conselhos. À Isabel Gerhardt por ter sido tão amiga desde o começo do meu namoro com o Felipe. Foi muito bom ter convivido, pelo menos um pouquinho, com cada um de vocês, cada qual pelo seu motivo. E à Luciane Gauer, por tudo isso, por ter sido a minha primeira "orientadora", por ter me iniciado na arte da construção gênica, e principalmente, pela sua amizade. Alguns amigos a gente guarda pela vida toda, e eu sei que, mesmo longe, você é um deles.

Ao Paulo Fisch, por toda a ajuda com as milhares de seqüências, banco de dados e blasts. E ao Marcelo Rebello, por ter facilitado muito a minha vida ao me ensinar o pouco que sei de SQL.

Aos amigos do laboratório do Genoma Funcional, Ju, Renato, Jorge, Sandra, Geraldo, Edna, Paulino, Marcelo, Layra, Michele, Agustina, Pedro, Renata e Eduardo. Aos amigos do laboratório do Michel, especialmente à Amanda, ao Juarez e à Aline. Aos amigos da genética Animal, em especial Tati, Mari, Ana Carolina (que me ajudou muito com a papelada da defesa), Rosângela, Ana Cláudia e Ana Maria. À professora Anete Pereira de Souza, pelas conversas, conselhos, e por sempre me receber tão bem.

Às meninas da secretaria do CBMEG, Tânia, Sandra, Andressa e Paula, pela ajuda e pelas conversas. Vocês são muito especiais. Ao seu Chico, por me ajudar com as plantas. Ao pessoal da secretaria da pós graduação por toda a ajuda com a papelada, em especial à Lourdes, sempre atenciosa comigo.

Ao amigo BH, Truminha ou Luiz Gustavo Guedes Corrêa, pelas dicas sobre bZIPs, por me divertir tanto, e por ter sempre me entendido tão bem. Você faz muita falta.

Às queridas amigas Sylvia e Thaís, pelos grandes almoços, pelos cafezinhos no final da tarde, pelos cinemas fora de hora, pelas conversas nos dias ruins, e nos dias felizes também. Estou certa de que um dia alcançaremos aquela vida de luxo e riqueza. Vou sentir muitas saudades de vocês.

Aos meus amigos 98D, que ainda me proporcionam grandes momentos por e-mail, me fazendo sentir menos saudades dos bons tempos.

Aos meus pais Enecilda e Carlos Alberto (ou Nê e Charles), por terem me apoiado sempre e incondicionalmente, por acreditarem em mim, e por sempre me deixarem decidir sozinha o meu caminho, especialmente quando eu quis que decidissem por mim. Isso me deu força, me fez uma pessoa muito melhor. Amo muito vocês. Ao meu irmão Felipe, pelo companheirismo, por me fazer rir quando isso é a última coisa que eu quero, e também à Calol. E ao meu irmão Daniel. Eu ainda estou tentando te entender, mas mesmo sem conseguir, eu te amo muito.

Aos meus tios e meus avós Esther, Nica e Antenor, vocês são muito importantes para mim. Ao vovô Joaquim, à tia Zezé, à vó Chica. Sei que vocês me fazem companhia e me assopram coisas sábias muitas vezes.

Ao Felipe. Não tenho palavras para dizer da sua importância na minha vida e na minha formação como cientista (que um dia eu vou ser). Te amo, muito. Como dizia Vinícius, “E de te amar assim, muito e amiúde, é que um dia, de repente, hei de morrer de amar mais do que pude”. À Lídia e ao Antônio Carlos Rodrigues da Silva, por terem me recebido tão bem em sua família. Sou muito feliz por ter encontrado pessoas tão especiais como vocês.

À American Society of Plant Biologists e ao comitê de organização do Maize Genetics Conference de 2006, pelos prêmios concedidos, e à FAEP, por tornar possível a minha participação em congressos internacionais. À CAPES, pela bolsa de estudos concedida durante o doutorado.

ÍNDICE

ÍNDICE

[α-P^{32}]dCTP	citosina 5' trifosfato marcada com fósforo 32
[α-P^{33}]dCTP	citosina 5' trifosfato marcada com fósforo 33
ABA	ácido abscísico
At	*Arabidopsis thaliana*
bZIP	basic leucine zipper – domínio básico e zíper de leucinas
cDNA	molécula de DNA complementar a um mRNA transcrito
Cluster	conjunto de reads que parecem representar o mesmo transcrito
Contig	contíguo; seqüência de DNA formada pela sobreposição de duas ou mais seqüências
DAP	dias após a polinização
DEPC	dietil pirocarbonato
DNA	ácido desoxirribonucléico
EDTA	ácido etilenodiaminotetracético
EST	expressed sequence tag – seqüência de um cDNA originado de um mRNA transcrito pelas células de tecidos, órgãos ou partes de um organismo
GA(s)	giberelina(s)
GUS	gene codificador da enzima ß-glucuronidase de Escherichia coli
kb	milhar(es) de par(es) de base(s)
kDa	kilodalton(s)
mRNA	RNA mensageiro
µL	microlitro
ORF	open reading frame (seqüência aberta de leitura)
Os	*Oryza sativa*
pb	par(es) de base(s)
PCR	Polimerase Chain Reaction – Reação em Cadeia da Polimerase
Primers	seqüências iniciadoras da síntese de ácidos nucléicos
Read	leitura; seqüência de um clone
RNA	ácido ribonucléico
RT-PCR	Reverse Transcriptase Polimerase Chain Reaction – Reação da Polimerase em Cadeia usando uma transcriptase reversa
SD	desvio padrão
SDS	dodecil sulfato de sódio
Singleton	read que não se sobrepõe a nenhum outro
SSC	sal citrato de sódio
TF(s)	transcription factors; fatores de transcrição
Tris	tris(hidroximetil)aminometano
Zm	*Zea mays*

O seqüenciamento de ESTs (etiquetas de seqüências expressas) e a sua organização em bancos de dados constituem poderosas ferramentas para identificar genes de interesse expressos em determinados tecidos e/ou tipos celulares. Neste trabalho criou-se um banco de seqüências expressas chamado MAIZESTdb, que contém ESTs de diversos tecidos de milho, porém enriquecido com seqüências provenientes do endosperma de milho em desenvolvimento. O MAIZESTdb contém 227.431 ESTs vindos de mais de 30 órgãos e tecidos de milho diferentes, 30.531 seqüenciados em nosso laboratório a partir de bibliotecas construídas com RNA mensageiro de endosperma. Estas seqüências representam uma grande contribuição na identificação de novos genes expressos no endosperma. A análise deste banco de ESTs possibilitou a identificação de 4.032 transcritos preferencialmente expressos no endosperma, e a sua anotação revelou uma ampla variedade de prováveis genes novos envolvidos no desenvolvimento e no metabolismo do endosperma.

O banco MAIZESTdb foi utilizado neste trabalho para a identificação de fatores de transcrição (TFs) expressos no endosperma de milho, e, especialmente, na identificação de fatores preferencialmente expressos no endosperma, que podem desempenhar papéis regulatórios importantes durante a formação da semente. Foram identificados 1.233 TFs expressos em milho, 414 dos quais expressos no endosperma em desenvolvimento. Foram identificados ainda, através de análises *in silico*, 113 TFs preferencialmente expressos no endosperma, conjunto este que representa 9.2% dos TFs expressos identificados em milho, e que possivelmente contém reguladores importantes dos processos de especificação celular e desenvolvimento do endosperma de milho. Esta é a maior coleção de fatores de transcrição já descrita para este tecido, e representa uma fonte de dados importante para identificação de reguladores dos principais processos relacionados ao desenvolvimento do endosperma, como metabolismo de nitrogênio e carboidratos e controle da massa da semente.

Uma das famílias mais representadas entre os TFs preferencialmente expressos no endosperma foi a família NAC de fatores de transcrição. Esta família apresentou 12 membros preferencialmente expressos no endosperma de milho. Um novo membro da família NAC, chamado de EPN-1 (Endosperm Specific NAM 1), teve seu perfil de expressão caracterizado. Sua expressão pode ser detectada desde os 5 DAPs, embora o pico de expressão ocorra entre 20 e 25 DAP, e ele apresenta expressão preferencial no endosperma. O promotor do gene EPN-1 foi clonado, seqüenciado e analisado quanto aos seus possíveis elementos CIS regulatórios; foram encontrados elementos conservados relacionados à endosperma-especificidade, elementos relacionados à regulação por ácido abscísico e giberelinas, e elementos conservados presentes nos promotores de α-amilases, indicando uma possível relação deste gene com o processo de transição entre a maturação e a germinação da semente. Ensaios de expressão transitória com o promotor do gene EPN-1 revelaram que sua expressão está dirigida à camada de aleurona do endosperma de milho, o que constitui mais uma evidência de sua possível função na regulação de genes relacionados aos processos de maturação e germinação da semente.

ABSTRACT

The sequencing of ESTs (expressed sequence tags) and its organization in databases constitute powerful tools to identify genes of interest in certain tissues and/or cell types. In this work we have created MAIZESTdb, a database of ESTs expressed in diverse maize tissues. The importance of this database, however, is that it is enriched with sequences from developing maize endosperm. The MAIZESTdb contains 227,431 ESTs coming from more than 30 different maize tissues and organs, 30,531 of which sequenced from endosperm cDNA libraries constructed in our laboratory. These sequences represent a great contribution for the identification of novel genes expressed in endosperm. The analysis of this ESTs database led to the identification of 4,032 transcripts preferentially expressed in the endosperm, and its annotation revealed a great variety of new genes involved in endosperm metabolism and development.

The MAIZESTdb was then used to identify transcription factors (TFs) expressed in maize endosperm, and, mainly, in the identification of TFs preferentially expressed in the endosperm. We identified 1,233 TFs expressed in diverse maize tissues, 414 of which expressed in developing endosperm. We also identified, through *in silico* comparison of transcript abundance and library source, 113 TFs with preferential expression in endosperm, representing 9,2% of the TFs identified in this work. This dataset probably contains important regulators of cellular specification of the endosperm development. This is the biggest TFs collection reported for this tissue, and represents an important source of data for identification of regulators for main processes related to the endosperm development such as nitrogen and carbohydrate metabolism and control of seed mass.

One of the most represented families among the TFs preferentially expressed in endosperm was the NAC family of transcription factors. This family presented 12 members with preferential expression in the endosperm. A new member of the NAC family, called EPN-1 (Endosperm Specific NAM 1), was characterized. Its expression

can be detected preferentially in the endosperm, beginning early at 5 DAPs, and the peak of expression occurs between 20 and 25 DAP. The EPN-1 promoter was cloned and sequenced, and its sequence was screened for putative CIS-acting regulatory elements. Conserved elements related to endosperm-specific expression were found, as well as elements related to abscisic acid and gibberellins regulation and conserved elements found in the promoters of alpha-amylases, indicating that this gene may have a regulatory role during the transition from the seed maturation to the seed germination process. Transient expression assays were conducted using the EPN-1 promoter driving a reporter gene and its expression was directed to the aleurone layer of the endosperm, what constitutes an additional evidence of its possible role in the regulation of genes related to the maturation and germination processes.

1. Os Cereais

Entre as plantas cultivadas, os cereais merecem grande destaque em relação à área plantada, produção e contribuição para alimentação animal e humana. No ano de 2005, mais de 681 milhões de hectares foram cultivados com cereais em todo o mundo, produzindo pouco mais de 2,2 bilhões de toneladas de grãos. Três espécies contribuíram com 89% deste total: arroz (614 milhões de toneladas), trigo (626 milhões de toneladas) e milho (692 milhões de toneladas) (FAO, 2006; Tabela 1). O grande sucesso no cultivo de cereais deve-se, principalmente, à sua alta produtividade, facilidade de colheita e à capacidade dos cultivares em adaptarem-se a diferentes condições ambientais (Lazzeri and Shewry, 1993).

O principal produto resultante do cultivo de cereais é o grão, apesar de caules e folhas serem bastante utilizados para silagem. Em termos botânicos, o grão é uma cariopse, tipo de fruto em que a parede da semente (testa) encontra-se fundida com a parede do fruto (pericarpo) (Lazzeri e Shewry, 1993). Pesquisas recentes têm mostrado que proteínas vegetais representam 65% da quantidade total de proteínas ingeridas em todo o mundo, e que 47% destas são proteínas de grãos de cereais (Millward, 1999).

Tabela 1. Produção e área cultivada com cereais no mundo

Espécie	Produção (milhões de toneladas)	Área (milhões de hectares)
Milho	692,0	147,0
Trigo	626,5	216,2
Arroz	614,7	153,5
Cevada	138,3	56,5
Sorgo	56,9	42,7
Milheto	27,3	35,9
Aveia	24,6	11,8
Centeio	15,0	6,6
Triticale	13,5	3,5

Fonte: FAOSTAT Database, 2006.

2. Milho: origem, genética e importância econômica

O milho (*Zea mays* L.) é uma gramínea de origem centro e sul-americana, pertencente à família Poaceae e à tribo Andropogoneae, que engloba também o sorgo, o *Trypsacum* e o *Coix* (Claynton, 1973; 1983). O milho é uma das plantas cultivadas mais importantes atualmente, e a espécie mais produzida nos países em desenvolvimento. Seu cultivo pode ser feito na amplitude latitudinal de 50ºN a 50ºS - o que compreende climas tropicais, subtropicais e temperados - e do nível do mar a altitudes superiores a 3000 metros. Devido à sua alta adaptabilidade a diversos ambientes, o milho é o cereal mais cultivado em termos de número de países (cerca de 70). Apenas no ano de 2005 foram produzidas aproximadamente 692 milhões de toneladas de milho em cerca de 147 milhões de hectares. Os Estados Unidos respondem por pouco mais de 41% dessa produção, e os cinco maiores produtores concentram cerca de 71%. O Brasil ocupa a terceira posição no ranking mundial, produzindo 34,8 milhões de toneladas em cerca de 11,4 milhões de hectares.

Além de ser uma das culturas de maior importância econômica no mundo, o milho merece destaque como planta modelo para pesquisa básica em Genética e Bioquímica, sendo o sistema genético mais estudado entre as monocotiledôneas, devido à ampla disponibilidade de mutantes e à facilidade de efetuar-se cruzamentos controlados (Chasan, 1994; MGDb - www.maizegdb.org).

O genoma do milho está organizado em 10 cromossomos (n=10, 2n=20) que contém cerca de 2,5 bilhões de pares de bases, tamanho comparável ao do genoma humano (~3,2 bilhões de pares de bases). Uma porção significativa deste genoma compreende regiões repetitivas (Hake and Walbot, 1980), a maioria delas contendo retroelementos, fragmentos móveis de DNA que se transpõe no genoma através de intermediários de RNA utilizando transcriptases reversas (Bennetzen, 2000).

A família das Gramíneas é formada por cerca de 10.000 espécies, muitas delas com grande importância econômica. O conteúdo haplóide dos genomas é bastante variável entre as espécies, indo desde 0,45 picogramas em arroz até 11,7 picogramas em aveia, e elas apresentam diferentes número de cromossomos (Arumanagathan and Earle, 1991). A construção de mapas genéticos comparativos de várias espécies de gramíneas, tais como milho, trigo e arroz, tem facilitado o conhecimento e a localização de genes

nos genomas deste grupo (Gale and Devos, 1998). Recentes pesquisas têm revelado que os genomas de gramíneas possuem um alto grau de similaridade, não somente em relação aos genes, mas também em relação aos grupos de ligação nos cromossomos (Cook, 1998). A descoberta da colinearidade dos genes nos cereais, os quais possuem uma estreita relação evolutiva, tem permitido uma nova perspectiva no estudo de como os genes e as informações geradas podem ser usados sinergisticamente para o melhoramento de todas as espécies de gramíneas (Bennetzen et al., 1998). Isso representa uma oportunidade para entender como a evolução favoreceu a formação de novos padrões morfológicos e vias metabólicas partindo de um mesmo conjunto inicial de material genético. Uma ampla variedade de espécies de gramíneas tem sido estudada com o intuito de identificar alelos úteis para a engenharia genética e melhoramento da produção de grãos.

3. O endosperma da semente de milho

A semente do milho é composta basicamente de duas partes: o endosperma e o embrião (Figura 1). Pesquisas recentes têm mostrado que o sucesso na formação da semente depende da interação entre seus dois principais componentes, e que a presença de um endosperma intacto é de extrema importância para o desenvolvimento apropriado do embrião (Consonni et al., 2005).

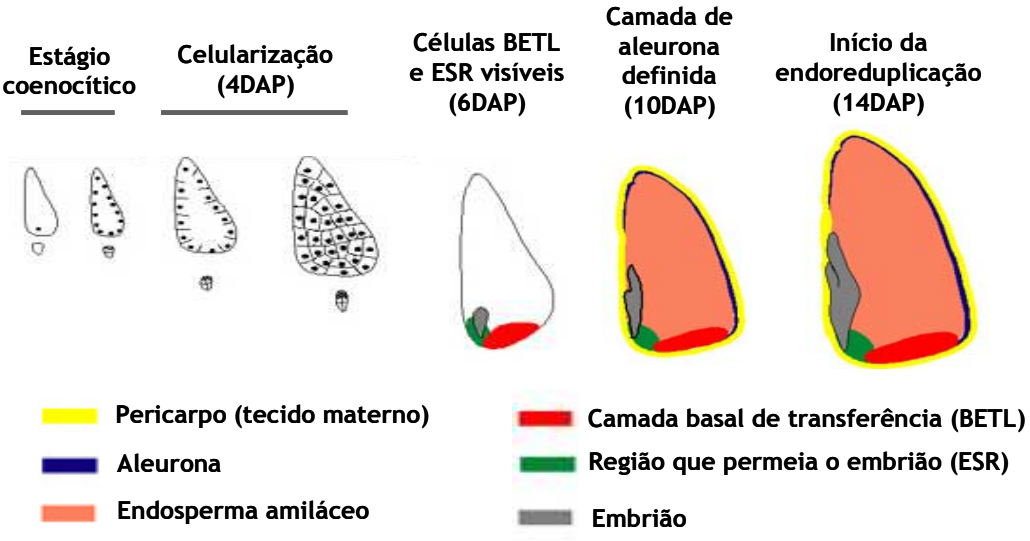


Figura 1. Principais tipos celulares e estágios do desenvolvimento do endosperma de milho (adaptado do original de Matt Evans – www.ciwdpb.stanford.edu/research/research_evans.php)

O endosperma constitui cerca de 80% do peso da semente. Sua função é distinta entre sementes de monocotiledôneas e dicotiledôneas. No primeiro grupo, o endosperma possui função de armazenamento de nutrientes a serem utilizados pelo embrião durante a germinação e no início do crescimento da plântula. Na maioria das dicotiledôneas, ao contrário, o endosperma assiste à embriogênese nutrindo o embrião apenas nos estágios iniciais, sendo completamente assimilado durante esta fase. Os cotilédones, folhas formadas durante a embriogênese, assumem a função de tecido de reserva de nutrientes a serem utilizados durante o processo de germinação (Lopes e Larkins, 1993).

Endosperma e embrião são produzidos por meio de um processo de dupla fertilização único em plantas superiores, no qual um núcleo espermático se funde com a célula-ovo do megagametófito, originando o zigoto diplóide que dará origem ao embrião, e outro núcleo espermático se funde com a célula central binucleada, dando origem ao endosperma triplóide (revisado por Russell, 1992 e Olsen, 2004). Logo após a fertilização, as células centrais começam a se dividir em ciclos repetitivos de mitose, sem a formação de parede celular ou citocinese, formando o endosperma coenócito (Figura 1). Então é iniciado o processo de celularização, e até o quarto dia após a polinização (DAP) o tecido deixa de ser uma única célula multinucleada e assume uma morfologia multicelular uninucleada. Entre 4 DAP e 15 DAP ocorre um rápido crescimento do endosperma, devido tanto à expansão quanto à divisão celular. Aos 12 DAP o endosperma preenche a região central da semente. As divisões celulares cessam nesta região, e os núcleos iniciam um processo de endoreduplicação (duplicação cromossômica sem mitose) que eleva substancialmente o conteúdo de DNA. No milho, entre 10 a 20 DAP, o conteúdo de DNA aumenta de 3 vezes o conteúdo do genoma haplóide para até 600 vezes. Acredita-se que o papel da endoreduplicação seja possibilitar altos níveis de expressão gênica em um tecido que demanda uma intensa atividade gênica e onde existem grandes limitações, tanto em termos de espaço quanto de tempo. Leiva-Neto et al. (2004) propõem ainda que a endoreduplicação funcione como um acúmulo de nucleotídeos para serem usados durante a embriogênese e/ou a germinação.

Entre 8 e 12 DAP se inicia o acúmulo de grandes quantidades de amido e de proteínas de reserva no endosperma amiláceo, e aos 16 DAP inicia-se o processo de maturação, preparando as sementes para dissecação e dormência. Aos 23 DAP o processo

de dissecação já se iniciou, e por volta de 25-30 DAP a quantidade relativa de água no endosperma começa a diminuir, sinal que mantém o desenvolvimento germinativo reprimido (revisado em Olsen, 2001, Lopes e Larkins, 1993 e Olsen, 2004).

O endosperma completamente desenvolvido é formado por 4 tipos celulares principais: o endosperma amiláceo, a camada basal de transferência (BETL, de *basal endosperm transfer layer*), a aleurona, composta por uma única camada de células, e a região que permeia o embrião (ESR, de *embryo surrounding region*) (Olsen, 2001; Figura 1).

O endosperma amiláceo representa a maior parte da massa da semente. As células da região central do endosperma acumulam uma grande quantidade de amido, enquanto as regiões periféricas são mais ricas em proteínas de reserva. A aleurona é a camada celular mais externa do endosperma, e é conservada durante os processos de maturação e dissecação da semente. As células da aleurona são morfológica e funcionalmente distintas dos outros tipos celulares do endosperma. Quando as sementes começam a germinar, estas células, estimuladas por giberelinas produzidas pelo embrião, iniciam a produção de enzimas hidrolíticas. Estas enzimas catalisam a degradação de paredes celulares e macromoléculas de reserva (amido, proteínas e DNA) acumulados durante o desenvolvimento no endosperma amiláceo. Na região chalazal da semente, próximo ao pedicelo, a camada de aleurona é substituída pela camada basal de transferência (BETL), que faz a interface entre o tecido esporofítico e a semente, mediando a entrada de nutrientes maternos (Thompson et al., 2001; Offler et al., 2003). A região que permeia o embrião (ESR) corresponde a uma pequena área localizada no pólo micropilar, circundando a porção basal do embrião. É caracterizada por células pequenas com citoplasma denso que podem desempenhar funções na nutrição do embrião e/ou na formação de uma barreira física entre o embrião e o endosperma durante o desenvolvimento da semente (Opsahl-Ferstad et al., 1997).

4. Fatores reguladores da transcrição

A eficiência dos mecanismos moleculares e bioquímicos que controlam fenômenos biológicos tais como diferenciação, controle celular, desenvolvimento e resposta a estímulos ambientais está estritamente relacionada com a fina regulação da expressão gênica. Esta regulação assegura que uma determinada proteína seja produzida em sua

exata quantidade, no exato momento e no local apropriado para que sua função biológica no desenvolvimento do organismo seja cumprida (Näär et al, 2001).

Em células eucarióticas, a indução da expressão gênica e da atividade de proteínas biologicamente ativas pode ser regulada em diversos níveis (Meshi and Iwabuchi, 1995; Beckett, 2001; Warren, 2002; Wray et al., 2003):

1. **Estrutura da cromatina** - a estrutura física do DNA compactado e a presença de histonas e de ilhas de metilação podem afetar a habilidade das proteínas regulatórias (conhecidas como fatores de transcrição) e da RNA polimerase de acessar genes específicos e iniciar a sua transcrição;

2. **Iniciação da transcrição** - este é o principal ponto de regulação da expressão gênica, que pode ser afetada pela ligação de diferentes reguladores (ativadores ou repressores) ao promotor do gene e das interações entre eles e o complexo basal de transcrição;

3. **Processamento e modificação pós transcricionais** - RNAs mensageiros eucarióticos devem ser poliadenilados, e os íntrons devem ser removidos com precisão; neste ponto podem ocorrer splicings alternativos, que darão origem a diferentes proteínas a partir de um mesmo gene;

4. **Transporte do mRNA** - o mRNA processado deve sair do núcleo e chegar ao citoplasma, onde será traduzido;

5. **Estabilidade do transcrito** - Ao contrário dos mRNAs procarióticos, que possuem uma meia-vida de 1 a 5 minutos, a estabilidade dos mRNAs eucarióticos pode variar bastante. Alguns transcritos instáveis apresentam sinais para rápida degradação (geralmente na porção 3' não traduzida);

6. **Iniciação da tradução** - Muitos mRNAs têm múltiplos códons de iniciação (ATG), e a habilidade dos ribossomos em reconhecer o sítio correto pode afetar a produção da proteína;

7. **Modificações pós-traducionais** - Entre as mais comuns estão a glicosilação, a acetilação, a fosforilação e a formação de pontes dissulfeto;

8. **Transporte da proteína** - Para que possam tornar-se biologicamente ativas após a tradução, as proteínas devem ser transportadas para o seu sítio de ação;

9. **Controle da estabilidade da proteína** - Muitas proteínas são rapidamente degradadas, enquanto outras permanecem estáveis, fato relacionado a seqüências específicas de aminoácidos que levam à rápida degradação.

Exemplos de regulação em cada um desses passos são conhecidos, embora para a maioria dos genes o principal nível de regulação ocorra durante a transcrição do DNA em mRNA através da atuação de proteínas regulatórias, os fatores de transcrição. Fatores de transcrição são proteínas que se ligam a regiões específicas nos promotores dos genes e controlam a produção do RNA mensageiro. Eles podem ser divididos em 2 tipos: (1) Fatores basais, necessários para a formação do complexo de pré-iniciação da transcrição, presentes em todas as células e ativos nas mais diversas condições e (2) Fatores de transcrição sítio-específicos, presentes apenas nos tipos celulares onde atuam e/ou em determinado momento do ciclo de vida do organismo. Fatores sítio-específicos reconhecem e se ligam a seqüências específicas localizadas nos promotores dos genes (elementos CIS), e, associados a outros componentes da maquinaria de transcrição, como fatores basais, cofatores, remodeladores de cromatina e a própria RNA polimerase II, ativam ou reprimem a síntese do mRNA (Figura 2; Kuhlemeier, 1992; Kornberg, 1999; Lee e Young, 2000).

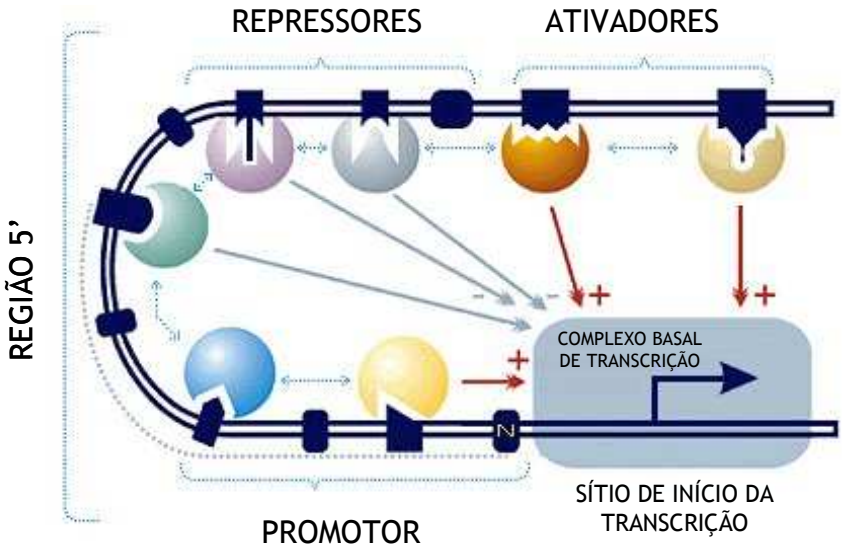


Figura 2. Esquema do complexo de iniciação da transcrição, contendo o complexo basal associado à RNA polimerase, os ativadores e os repressores ligados ao promotor do gene a ser transcrito (adaptado de GeneNetWorks™).

Com base nas similaridades entre seqüências de aminoácidos e entre as estruturas dos domínios de ligação ao DNA e de multimerização, os fatores de regulação da transcrição podem ser classificados em famílias, caracterizadas por motivos conservados, entre as quais podemos citar:

4.1. Família bZIP (*basic-region leucine zipper*)

É caracterizada por dois subdomínios: uma região básica e um zíper de leucinas. A região básica é composta por cerca de 30 aminoácidos básicos, que formam uma estrutura em forma de hélice para interagir com o DNA-alvo. O zíper de leucinas é constituído por repetições de resíduos de leucina a cada sete aminoácidos, numa extensão de 20 a 40 resíduos, com um número de repetições de leucina variando entre três e nove. Esta região é responsável pela dimerização com outras proteínas (Meshi e Iwabuchi, 1995; Landschultz et al., 1998; Pabo e Sauer, 1992; Hurst, 1995).

Análises genéticas, moleculares e bioquímicas indicam que os fatores bZIP são reguladores importantes de processos específicos de angiospermas como o desenvolvimento de órgãos (Walsh et al., 1997; Chuang et al., 1999); elongação celular (Yin et al., 1997; Fukasawa et al., 2000); controle do balanço nitrogênio/carbono (Ciceri et al., 1999); mecanismos de defesa (Niggeweg et al., 2000; Zhang et al., 1999; Despres et al., 2000; Pontier et al., 2001); vias de sinalização de hormônios e da sacarose (Choi et al., 2000; Uno et al., 2000; Niggeweg et al., 2000); resposta à luz (Osterlund et al., 2000; Ulm, et al., 2004); e controle osmótico (Satoh et al., 2004).

O gene Opaco-2 (*O2*) codifica um fator de transcrição desta família bastante estudado. Desde a descoberta do mutante *opaco-2* (*o2*) de milho rico em lisina (Mertz et al.,1964), muitos pesquisadores têm trabalhado com o intuito de desvendar os mecanismos moleculares e bioquímicos que levam ao aumento do conteúdo de lisina no endosperma da semente. Estudos realizados nos últimos 30 anos revelaram que sementes homozigotas *o2o2* apresentam uma redução de aproximadamente 70% no conteúdo de zeínas, proteínas de reserva de milho, devido principalmente a uma drástica redução das α-zeínas de 22 kDa, e que o conteúdo de outras proteínas e enzimas relacionadas ao metabolismo de açúcar e nitrogênio no endosperma está alterado nestas sementes (Giroux et al., 1994; Gallusci et al., 1996; Vettore et al., 1998; Kemper et al., 1999). A clonagem do gene *O2* revelou que ele codifica uma proteína pertencente à classe dos

fatores de transcrição do tipo bZIP (Schmidt et al., 1987; Motto et al., 1988). Mais tarde, foi demonstrado que a proteína O2 controla a transcrição das α-zeínas de 22 kDa e do gene da albumina b-32 de milho através do reconhecimento de uma seqüência específica em seus promotores (Lohmer et al., 1991; Schmidt et al., 1992). Em seguida, foi descoberto que, além do gene de α-prolaminas, a proteína O2 também controla a transcrição de genes de ß-prolaminas de milho e Coix (Cord Neto et al., 1995). Evidências mais recentes sugerem que a proteína O2 está envolvida na regulação coordenada da síntese de proteínas e do metabolismo de açúcar e nitrogênio durante a maturação das sementes de milho (Yunes et al. 1998, Gallusci et al., 1996; Kemper et al., 1999).

Em leveduras, o fator GCN4 é um dos componentes mais importantes no sistema regulatório do metabolismo de nitrogênio. Foi demonstrado, em cevada, que um fator contendo um motivo similar a GCN4 tem um papel importante na indução da síntese de proteínas de reserva por nitrogênio, e, em arroz, que este motivo, que é altamente conservado nos promotores dos genes de proteínas de reserva entre os cereais, tem um papel importante no controle da expressão endosperma-específica destas proteínas. A composição de prolaminas em sementes de milho também parece ser influenciada pela quantidade de nitrogênio. Essas observações, os efeitos da mutação *o2* em enzimas que fazem parte do metabolismo de aminoácidos e carbono e as similaridades funcional e estrutural entre a proteína O2 e o fator GCN-4 sugerem que O2 pode estar envolvido em um controle geral do metabolismo de aminoácidos no endosperma de milho (Kemper et al., 1999; Onodera et al., 2001, Arruda et al., 2000).

4.2. Família helix-loop-helix (HLH)

Fatores de transcrição pertencentes a essa família são componentes regulatórios importantes em muitas vias transcricionais relacionadas ao desenvolvimento de um organismo. São fatores envolvidos em processos como proliferação celular e diferenciação, determinação de linhagem celular e do sexo e até neurogênese e miogênese, e são encontrados desde leveduras até humanos (Atchley and Fitch, 1997).

Proteínas HLH são caracterizadas por possuírem 2 domínios altamente conservados, o de ligação ao DNA e o de interação com outras proteínas. O primeiro, composto principalmente por resíduos básicos, permite a ligação específica a uma

seqüência de 6 nucleotídeos conhecida como E-box (CANNTG). O segundo motivo, formado principalmente por resíduos hidrofóbicos, é chamado de domínio *helix-loop-helix* e permite interações entre proteínas e a formação de homo e heterodímeros. O motivo de dimerização contém cerca de 50 aminoácidos e se dispõe na forma de duas α-hélices anfipáticas separadas por um *loop* de tamanho variável. Algumas proteínas conhecidas como bHLH (*basic helix-loop-helix*) contêm ainda um motivo de dimerização do tipo zíper de leucinas, caracterizado por hepta-repetições de resíduos de leucina que ocorrem imediatamente após o motivo HLH, na porção C-terminal da proteína .

Em milho, duas famílias de reguladores, r e c1, controlam a transcrição de genes do metabolismo de antocianinas em diversos tecidos, como anteras, sementes, folhas e plântulas. Os membros da família r (R, Lc, Sn, B) codificam fatores de transcrição do tipo bHLH (Ludwig et al., 1989; Radicella et al., 1991; Consonni et al., 1993).

4.3. Família Homeobox

O papel das proteínas desta família está relacionado ao controle da determinação genética do desenvolvimento e diferenciação celular (Gehring, 1994). Fatores homeobox foram identificados pela primeira vez como proteínas expressas a partir de regiões de um cromossomo de *Drosophila* que continham seqüências conservadas chamadas homeoboxes. Esta nomenclatura foi adotada porque estes genes foram identificados por mutações que afetavam a morfologia da mosca. Essas mutações são chamadas de homeóticas por muitas vezes envolverem duplicações de partes do corpo (homeose) (Bürglin, 2005).

A seqüência amplamente conservada entre os diversos genes homeóticos é conhecida como homeodomínio, e é composta por cerca de 60 aminoácidos próximos à região C-terminal, cuja estrutura tridimensional apresenta três estruturas α-hélice consecutivas, com a terceira interagindo principalmente com o sulco maior da dupla fita de DNA. O domínio é composto por cerca de 50 aminoácidos, organizados numa estrutura globular que mantém a habilidade de ligação ao DNA. As hélices 2 e 3 interagem formando uma estrutura do tipo *helix-turn-helix* (Meshi e Iwabuchi, 1995; Luscombe et al, 2000).

4.4. Família MYB

O domínio MYB foi originalmente descrito como o domínio de ligação ao DNA do proto-oncogene MYB. Apresenta duas a três cópias de uma seqüência repetitiva composta por 51 a 53 aminoácidos com três resíduos conservados de triptofano, intercalados por intervalos de 18-19 aminoácidos, formando assim uma estrutura hidrofóbica.

Algumas da funções desempenhadas por proteínas MYB são regulação do ciclo celular, proliferação e especificação celular. Alguns membros dessa família em plantas constituem uma subfamília caracterizada pelo domínio MYB tipo R2R3, entre eles os fatores C1, P, PL, Zm1 e Zm38 de milho, que estão envolvidos na regulação da biossíntese de fenilpropanóides (Meshi e Iwabuchi, 1995; Avila et al., 1993).

4.5. Família MADS

Este nome é derivado das iniciais dos quatro membros inicialmente identificados neste grupo (MCM1 de levedura, envolvido na resposta a ferormônios, AGAMOUS de *Arabidopsis*, e DEFA de *Antirrhinum*, envolvidos no desenvolvimento floral, e SRF humano, fator de regulação de genes expressos no início do desenvolvimento). O domínio MADS é composto por 56 aminoácidos, consistindo num par de α-hélices anti-paralelas que formam um *coiled coil* (estrutura protéica muito estável na qual α-hélices sofrem torções helicoidais adicionais) e de uma estrutura antiparalela *β-sheet* dupla fita, envolvida também em interações com outras proteínas acessórias (Meshi e Iwabuchi, 1995; Luscombe et al, 2000).

Os fatores MADS-box mais estudados são aqueles envolvidos na determinação da identidade dos órgãos florais. Análises de mutantes florais resultaram na criação de um modelo genético chamado ABC, que explica como a combinação de três classes de genes (A, B e C) determina a identidade dos 4 órgãos florais (pétala, sépala, estame e carpelo; revisado por Coen e Meyerowitz, 1991).

4.6. Família *Zinc-finger*

Zinc finger é um dos domínios mais encontrados entre as proteínas de ligação ao DNA, e elas podem desempenhar as mais diversas funções. Uma grande variedade de fatores de transcrição contendo zinco foram descritas, nas quais um ou mais íons zinco

estabilizam a estrutura terciária do motivo. O clássico motivo zinc-finger é caracterizado por dois resíduos conservados de cisteína e dois resíduos conservados de histidina que ligam-se a um íon zinco, formando um tetraedro. A porção finger é composta por cerca de 30 aminoácidos que compreendem duas estruturas antiparalelas ß-sheet e uma estrutura em α-hélice (Meshi e Iwabuchi, 1995; ; Luscombe et al, 2000).

4.7. Família NAC

Esta família é formada por proteínas específicas de plantas que apresentam um domínio altamente conservado, definido como NAC. Este domínio foi nomeado com base nas primeiras proteínas identificadas em *Arabidopsis thaliana*: NAM, ATAF1 e 2 e CUC2 (Aida et al., 1997).

O domínio NAC pode ser subdividido em cinco subdomínios (A a E). O domínio como um todo é rico em aminoácidos básicos (R, K e H) , mas a distribuição dos resíduos positivos e negativos entre os domínios é desigual. Os subdomínios C e D são ricos em aminoácidos básicos e pobres em aminoácidos ácidos, enquanto o subdomínio B contém uma alta proporção de aminoácidos ácidos. Sinais de localização nuclear (NLS, de *nuclear localization signal*) putativos foram encontrados nos subdomínios C e D (Kikuchi et al., 2000). O domínio de ligação ao DNA está localizado numa região de 60 aminoácidos localizada nos subdomínios D e E (Duval et al., 2002). O domínio NAC consiste numa estrutura *ß-sheet* antiparalela torcida, que se encontra com uma α-hélice N-terminal de um lado e com uma hélice menor do outro lado. Esta estrutura sugere que o domínio NAC está envolvido na dimerização destas proteínas, e a face do dímero rica em resíduos positivos se liga ao DNA (Ernst et al., 2004).

Membros dessa família podem estar envolvidos em diversos processos celulares, tais como formação do meristema apical (Souer et al., 1996), resposta a patógenos e sinalização para crescimento (Xie et al. 1999), senescência (John et al. 1997; Guo et al., 2004), desenvolvimento de flores, folhas, raízes e sementes (Sablowski e Meyerowitz, 1998; Xie et al., 2000; Ge et al., 2004) e resposta a diferentes estresses (Kikuchi et al., 2000; Collinge e Boller, 2001; Tran et al. 2004). Guo et al. (2003) identificaram o primeiro membro da família NAC expresso especificamente no endosperma de milho. O gene foi chamado de NRP1 (de NAM-related protein 1) e seu pico de expressão ocorre aos 25 DAP. Sua função, no entanto, permanece desconhecida, embora tenha sido

demonstrado que este gene sofre *imprinting* de maneira gene-específica, assim como alguns genes que possuem papéis importantes na regulação do desenvolvimento do endosperma (Alleman e Doctor, 2000; Baroux et al. 2002), como MEA, FIS2 e FIE em *Arabidopsis* (Chaudhury et al., 2001; Grossniklaus et al., 1998; Luo et al., 1999) e FIE1 em milho (Danilevskaya et al., 2002).

Estas são as famílias de fatores de transcrição mais estudadas. Entretanto, muitas proteínas que apresentam capacidade de ligação ao DNA em seqüências específicas não apresentam homologia com domínios já descritos, e outras possuem mais de um motivo atuando conjuntamente na interação com o DNA. Deste modo, à medida que novos fatores de transcrição forem descobertos e caracterizados, essa classificação poderá ser complementada, e até mesmo novas classes poderão ser criadas.

5. O seqüenciamento de *Expressed Sequence Tags* (ESTs) como ferramenta para a descoberta de novos genes

Com o advento da era genômica, a identificação de genes tornou-se um processo mais dinâmico, capaz de gerar um vasto volume de informação em um curto período de tempo (Grivet e Arruda, 2001). Vários projetos têm sido conduzidos em diferentes espécies vegetais com o intuito de estudar o transcriptoma, ou seja, a população de RNAs transcrita de um determinado organismo, tecido, estágio de desenvolvimento, ou mesmo em resposta a tratamentos hormonais ou a estresses bióticos e abióticos (Ewing et al., 1999; White et al., 2000; Dong et al., 2003; Ma et al., 2003). Esses projetos são denominados projetos EST (*Expressed Sequence Tag* ou Etiqueta de Seqüência Expressa), e constituem uma poderosa ferramenta para identificar genes expressos em determinados tecidos e/ou tipos celulares de interesse. Nos projetos EST, bancos de dados contendo pequenas seqüências de DNA são gerados a partir do seqüenciamento de moléculas de cDNA sintetizadas das populações de mRNA com o auxílio de *primers* específicos que se ligam ao vetor (plasmídio) utilizado no processo de clonagem gênica. Essas seqüências são usadas na montagem de *contigs* ou *clusters* que, na maioria das vezes, possuem ORFs (*open reading frames*) representando a região codificadora de diversos genes (Telles et al., 2001). Desta forma, a tradução destas ORFs fornece os primeiros indícios da função da proteína codificada por um determinado clone de cDNA.

Bancos de ESTs contém informações biológicas de centenas de genes de um organismo, além de permitirem a identificação de diferentes isoformas de transcritos (Andrews et al., 2000) e o mapeamento gênico (Schuler, 1997; Wu et al., 2002). Outro aspecto importante dos ESTs é o acesso a informações sobre os genes expressos em organismos que contêm um genoma muito grande ou complexo (Vettore et al., 2001), tais como o milho, a cana-de-açúcar e o homem.

Uma grande quantidade de ESTs obtidos a partir de diferentes populações de mRNA pode fornecer uma estimativa da abundância relativa de transcritos de genes de interesse em diferentes tecidos/órgãos vegetais e também em diversas condições biológicas (Audic e Claverie, 1997). Esse processo de investigação do padrão de expressão de um gene *in silico*, conhecido como "*northern* digital", aliado a metodologias experimentais, possibilita a identificação e a análise de uma ampla gama de genes, os quais podem ser selecionados e utilizados em programas de melhoramento genético via biotecnologia.

Desta forma, o presente trabalho de doutoramento descreve, sob a forma de três artigos científicos, um deles já publicado e dois em processo de submissão:

1. A construção de um banco de ESTs enriquecido com seqüências vindas do endosperma de milho em desenvolvimento e a sua utilização para a identificação de genes preferencialmente expressos no endosperma (Capítulo I);

2. A identificação, a partir do banco de ESTs criado, de fatores de transcrição expressos no endosperma em desenvolvimento, incluindo um subconjunto de fatores preferencialmente expressos no endosperma (Capítulo II); e

3. A caracterização de um novo fator de transcrição preferencialmente expresso na camada de aleurona do endosperma, que pode ser um componente importante para a regulação da transição entre os processos de maturação e germinação da semente de milho (Capítulo III).

Esta tese de doutoramento foi realizada considerando-se os seguintes objetivos:

1. Identificar, classificar e anotar os fatores de transcrição

- Expressos no endosperma
- Preferencialmente expressos no endosperma

2. Avaliar o perfil de expressão e a possível função de alguns desses fatores de transcrição

Os objetivos específicos do Capítulo I foram:

- Identificar genes expressos no endosperma de milho através do seqüenciamento de ESTs (Expressed Sequence Tags);
- Criar um banco de dados com as seqüências geradas;
- Identificar e categorizar os genes tecido-específicos ou com expressão preferencial no endosperma.

Os objetivos específicos do Capítulo II foram:

- Identificar os fatores de transcrição (TFs) expressos no endosperma de milho presentes no banco de seqüências MAIZESTdb;
- Classificar e anotar os TFs identificados;
- Identificar os TFs tecido-específicos ou com expressão preferencial no endosperma, que possivelmente têm um papel fundamental no desenvolvimento deste tecido durante a formação da semente.

Os objetivos específicos do Capítulo III foram:

- Avaliar a seqüência e a estrutura gênica do gene EPN-1 (Endosperm Preferred NAM-1), identificado pela primeira vez em milho, e compará-lo a ortólogos;
- Clonar o promotor do gene EPN-1 e avaliar os possíveis elementos-CIS regulatórios;
- Avaliar seu perfil de expressão e sua possível função.

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

Natalia C. Verza, Thaís R. Silva, Germano Cord Neto, Fábio T.S. Nogueira, Paulo H. Fisch, Vicente E. de Rosa Jr, Marcelo M. Martins, André L. Vettore, Felipe R. da Silva and Paulo Arruda

Plant Molecular Biology 2005 Sep; 59(2):363-74.

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

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.

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 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

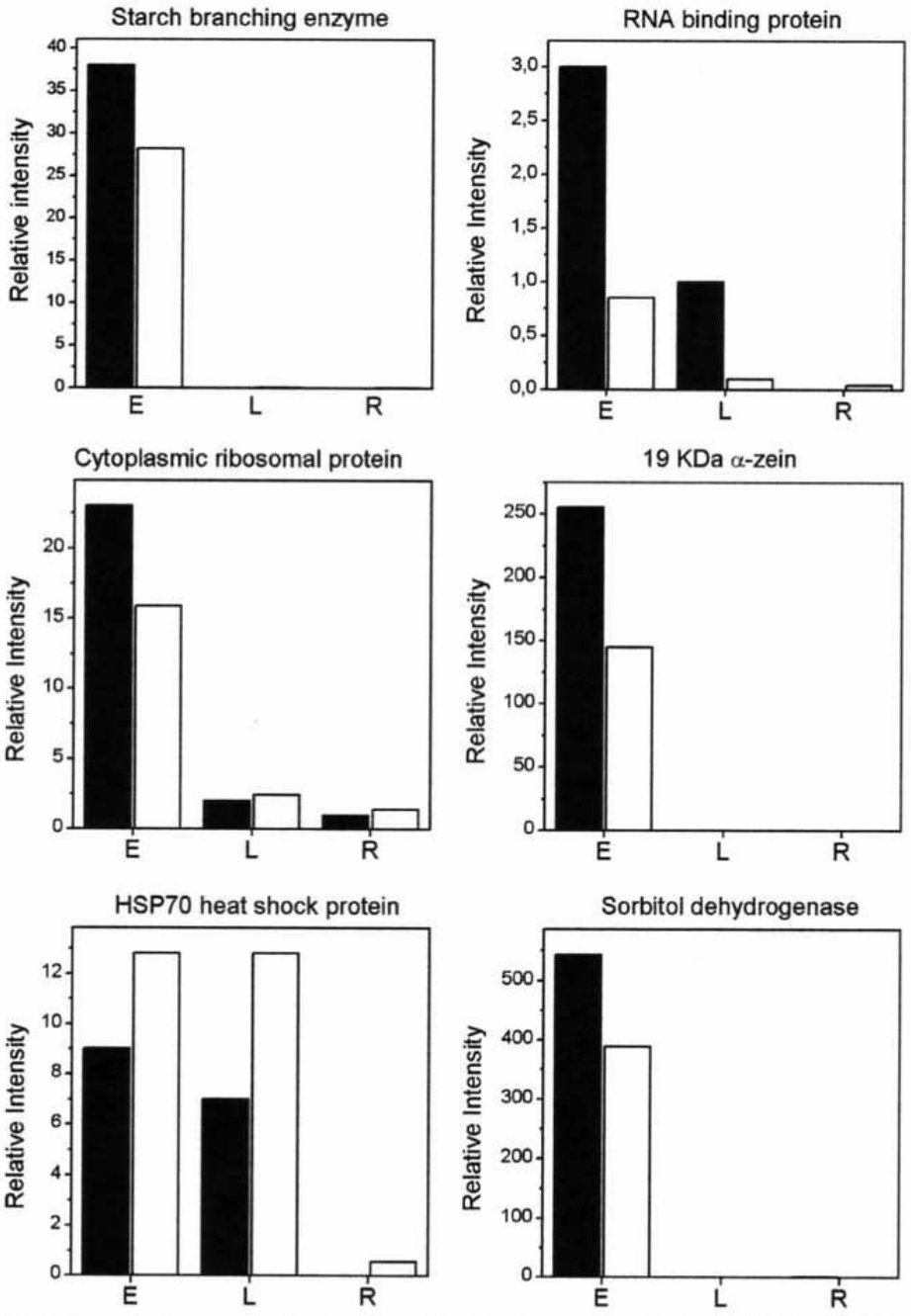


Figure 1. Examples of the expression profile of endosperm-preferred and not-preferred maize ESTs. Virtual Northern (black bars) and expression profile of high-density membrane arrays (open bars). E: Endosperm, L: Leaf and R: Root.

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

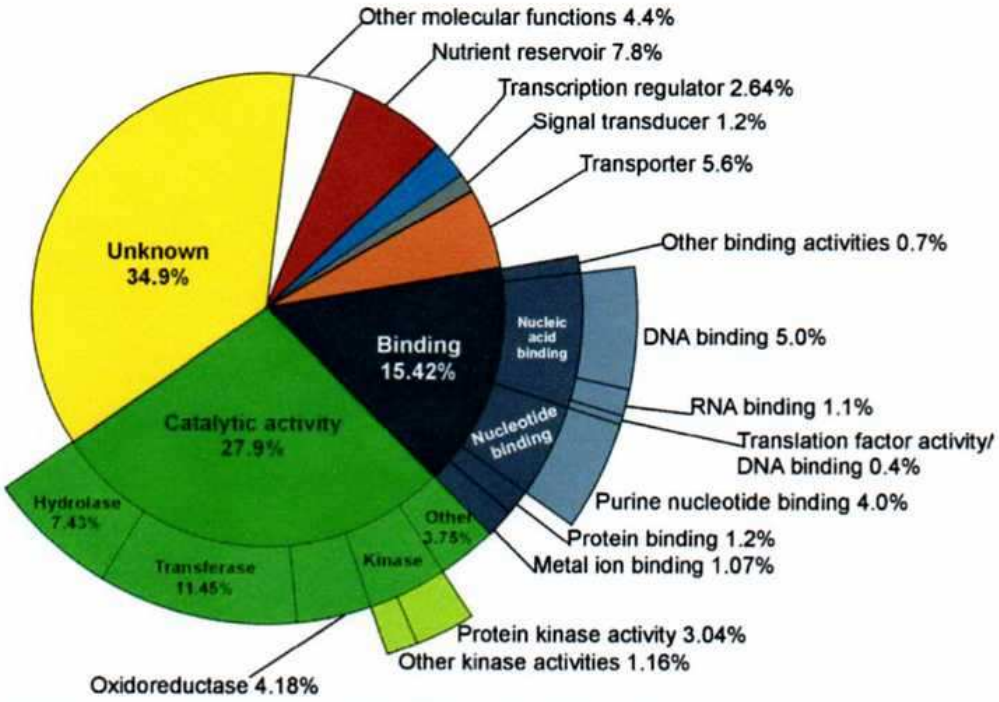


Figure 2. Maize endosperm gene prediction: biological process. Gene Ontology categories were assigned to MASs through curator-revised categorization. Classification is hierarchical, as children categories progress outwards from the inner parental categories. Two thousand four hundred and three endosperm-preferred 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

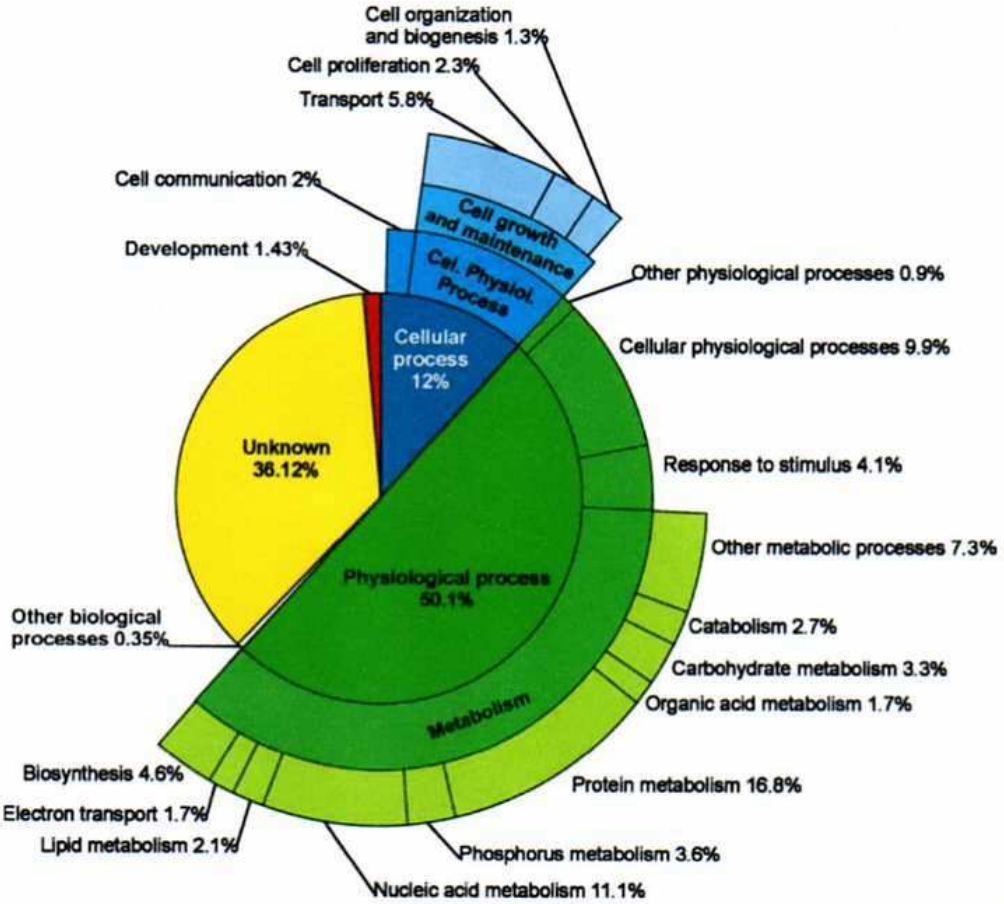


Figure 3. Maize endosperm gene prediction: molecular function. Gene Ontology categories were assigned to MASs through curator-revised categorization. Classification is hierarchical, as children categories progress outwards from the inner parental categories. Two thousand four hundred and three endosperm-preferred 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 microarray 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.

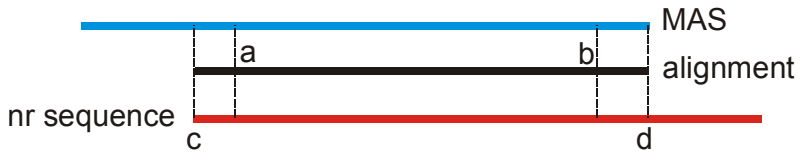
Whitfield, C.W., Band, M.R., Bonaldo, M.F., Kumar, C.G., Liu, L., Pardinas, J.R., Robertson, H.M., Soares, M.B. and Robinson, G.E. 2002. Annotated expressed sequence tags and cDNA microarrays for studies of brain and behavior in the honey bee. Genome Res. 12: 555–566.

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.

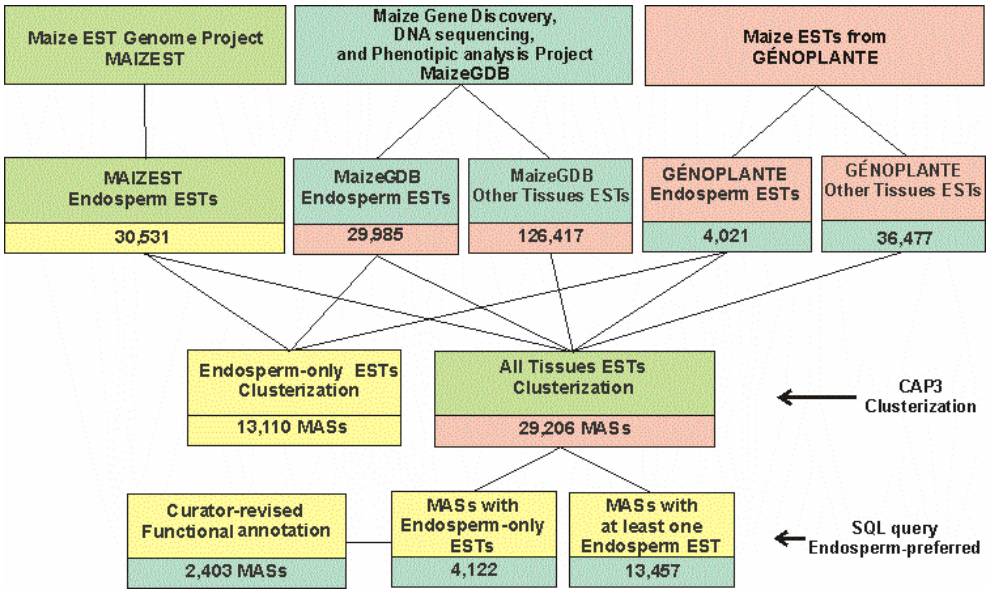
Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y. and Zhang, X. 2002. A draft sequence of the rice genome (*Oryza sativa* L ssp. indica). Science 296: 79–92.

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.

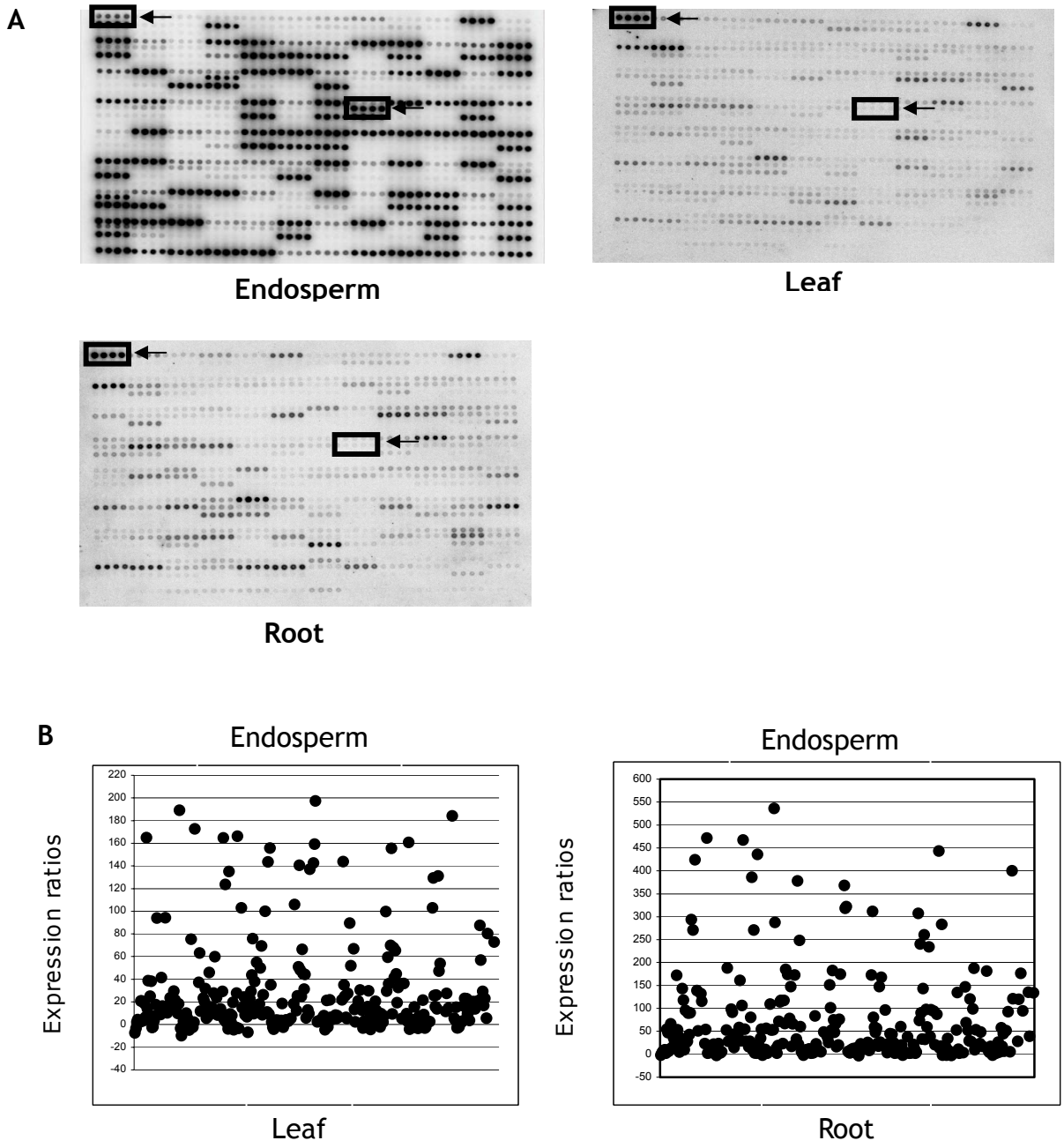
Material suplementar disponível no endereço eletrônico do periódico Plant Molecular Biology



Supplemental figure 1 - Schematic representation of the overlaping parameters used to compare ESTs and MASs with complete cDNA sequences. Curator-revised categorization was performed when BLASTX (cutoff value E = 10-5) alignment length (ab) was at least 50% of maximum overlapping length (cd).



Supplemental figure 2 - EST source project, clustering pipeline and endosperm-preferred MASs selection. EST sequence data were generated from nine endosperm cDNA libraries (MAIZEST project) or retrieved from public maize EST projects (MaizeGDB, http://www.maizegdb.org and Génoplante, http://www.genoplante.com). After trimming procedure, the sets of ESTs were clusterized using CAP3 program to generate Maize Assembled Sequences (MASs). Among those, a subset of endosperm-preferred transcripts was selected for curator-revised functional annotation according to the Gene Ontology consortium (http://www.geneontology.org).



Supplemental figure 3 - Expression profiling of ESTs randomly chosen from 10, 15 and 20 DAP endosperm libraries and hybridized with ^{33}P-labeled RNA from immature endosperm and leaves and roots from young seedlings. 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 endosperm (a mix of 10, 15 and 20 DAP), leaf and root from 7-day-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. (**A)** Typical membranes hybridized with RNA from 10, 15 and 20 DAP endosperm mix, leaf and root. Arrows indicate an endosperm-preferred EST and constitutive EST. (B) Expression ratios between endosperm, leaf and root tissues. Only the normalized expression ratios were used to construct the scatter plots. Expression ratios are plotted against the number of MAIZEST clones analyzed. Ratios below 1 were inverted and multiplied by –1 to aid better interpretation of the scatter plots.

Supplemental table 1: Endosperm specific sequences used for screening of non-endosperm tissue libraries

Accession	Description
AF072725	*Zea mays* starch branching enzyme IIb (ae), complete cds
AF371280	*Zea mays* Hageman factor inhibitor mRNA, complete cds
AF371279	*Zea mays* legumin 1 mRNA, complete cds
AF371278	*Zea mays* alpha globulin mRNA, complete cds
AF371277	*Zea mays* 22kD alpha zein 5 mRNA, complete cds
AF371276	*Zea mays* 22kD alpha zein 4 mRNA, complete cds
AF371275	*Zea mays* 22kD alpha zein 3 mRNA, complete cds
AF371274	*Zea mays* 22kD alpha zein 1 mRNA, complete cds
AF371273	*Zea mays* 19kD alpha zein B4 pseudogene, mRNA sequence
AF371272	*Zea mays* 19kD alpha zein B5 mRNA, partial cds
AF371271	*Zea mays* 19kD alpha zein B3 mRNA, complete cds
AF371270	*Zea mays* 19kD alpha zein B2 mRNA, complete cds
AF371269	*Zea mays* 19kD alpha zein B1 mRNA, complete cds
AF371268	*Zea mays* 19kD alpha zein D2 mRNA, complete cds
AF371267	*Zea mays* 19kD alpha zein D1 mRNA, complete cds
AF371266	*Zea mays* 10kD delta zein mRNA, complete cds
AF371265	*Zea mays* 18kD delta zein mRNA, complete cds
AF371264	*Zea mays* 15kD beta zein mRNA, complete cds
AF371263	*Zea mays* 50kD gamma zein mRNA, complete cds
AF371262	*Zea mays* 16kD gamma zein mRNA, complete cds
AF371261	*Zea mays* 27kD gamma zein mRNA, complete cds
M29411	*Zea mays* DNA binding protein opaque-2 (O2) mRNA, complete cds

Sequences available at http://www.ncbi.nlm.nih.gov/entrez/

Supplemental table 2: Complete maize sequences used to estimate redundancy into the MASs set

Accession	Description
AB112936	ZmpOMT1 mRNA for plastidic 2-oxoglutarate/malate transporter
AB112937	ZmpDCT1 mRNA for plastidic general dicarboxylate transporter
AB112938	ZmpDCT2 mRNA for plastidic general dicarboxylate transporter
AB112939	ZmpDCT3 mRNA for plastidic general dicarboxylate transporter
AB127981	mRNA for MinE
AF330034	isopentenyl pyrophosphate isomerase mRNA
AF330035	ADP-glucose pyrophosphorylase small subunit mRNA
AF330036	geranylgeranyl-diphosphate synthase mRNA
AF450481	DREB-like protein (DREB1A) mRNA
AF545813	SET domain protein 113 (sdg113) mRNA
AF545814	SET domain protein 110 (sdg110) mRNA
AY029312	seven transmembrane protein Mlo1 mRNA
AY122271	cultivar B73 SET domain-containing protein SET118 (SET118) mRNA
AY122272	cultivar B73 SET domain-containing protein SET104 (SET104) mRNA
AY122273	cultivar B73 SET domain-containing protein SET102 (SET102) mRNA
AY172976	SET domain protein 123 mRNA
AY183450	legumin-like protein mRNA
AY187718	SET domain protein SDG111 mRNA
AY187719	SET domain protein SDG117 mRNA
AY195849	fertilization-independent type 1 mRNA
AY195850	fertilization-independent type 2 mRNA
AY211982	transparent leaf area peptide mRNA
AY219173	submergence induced protein SI397 mRNA
AY241545	glyoxalase I (GlxI) mRNA
AY243800	plasma membrane intrinsic protein (PIP1-1) mRNA
AY243801	aquaporin (PIP2-1) mRNA
AY243802	aquaporin (PIP2-5) mRNA
AY291061	photosystem II subunit PsbS precursor, mRNA nuclear gene for chloroplast product.
AY315822	non-photosynthetic NADP-malic enzyme mRNA
AY372244	cellulose synthase catalytic subunit 10 (CesA10) mRNA
AY389497	cultivar Chalqueno ribosomal protein S6 kinase mRNA
AY466159	DEAD box RNA helicase (DRH1) mRNA
AY472082	narrow sheath 2 mRNA
AY485263	adenine phosphoribosyltransferase (apt1) mRNA
AY485529	heat shock protein HSP101 (HSP101) mRNA
AY488135	allene oxide synthase (aos) mRNA
AY488136	allene oxide cyclase (aoc) mRNA
AY496080	tousled-like kinase 2 (TLK2) mRNA
AY501430	rolled leaf1 mRNA
AY505017	Gl1 protein mRNA
AY515607	putative zinc finger protein ZmZf mRNA
AY530961	dTDP-glucose 4,6-dehydratase mRNA
AY530962	26S proteasome regulatory complex ATPase RPT3 mRNA
AY536525	phospholipase C (PLC) mRNA
AF039304	cpSecY (csy1) mRNA chloroplast gene for chloroplast product.
AF390542	ATP synthase subunit 9 (atp9) mRNA mitochondrial gene for mitochondrial product.
AF534133	S male sterility locus ORF355 and ORF17 mRNA mitochondrial genes for mitochondrial products.
AF536187	bicistronic S male sterility locus variant 1 mRNA mitochondrial gene for mitochondrial products.
AF536188	bicistronic S male sterility locus variant 2 mRNA mitochondrial gene for mitochondrial products.
AF536189	bicistronic S male sterility locus variant 3 mRNA mitochondrial gene for mitochondrial products.
AF536190	bicistronic S male sterility locus variant 4 mRNA mitochondrial gene for mitochondrial products.
AF536191	bicistronic S male sterility locus variant 5 mRNA mitochondrial gene for mitochondrial products.
MIZMMSODA	Mn-superoxide dismutase (Sod3.2) mRNA mitochondrial gene for mitochondrial product.
MIZMMSODB	Mn-superoxide dismutase (Sod3.3) mRNA mitochondrial gene for mitochondrial product.
MIZMMSODC	Mn-superoxide dismutase (Sod3.4) mRNA mitochondrial gene for mitochondrial product.
(…)	

Sequences retrieved from EMBLdatabase (http://www.ebi.ac.uk/embl/) on March 18, 2004.

Observação: Esta tabela contém 747 linhas, mas apenas as 55 primeiras linhas e a nota de rodapé estão representadas aqui. A tabela completa pode ser obtida através do endereço eletrônico http://www.springerlink.com/(qxyuv3bcuhbtevygc1ldcum3)/app/home/contribution.asp?referrer=parent&backto=issue,10,10;journal,9,339;linkingpublicationresults,1:100330,1/, no link Electronic Supplementary Material – OPEN ESM.

Supplemental Table 3. List containing the 4,032 endosperm-preferred MASs

MAS	Best hit BlastX	E-value	Accession number[a]
MZCCL10001A03.g	dbj|BAC64996.1| P0443G08.20 [Oryza sativa]	5,0E-22	CO439028
MZCCL10001D01.g	ref|XP_467965.1| remorin protein-like [Oryza sativa]	2,0E-48	CO439052
MZCCL10001D04.g	gb|AAL58889.1|AF461049_1 11 kDa methionine-rich protein [*Zea mays*]	6,0E-39	CO442451
MZCCL10001D06.g	gb|AAA33537.1| gamma zein [*Zea mays*]	1,0E-26	CO468283
MZCCL10001D08.g	ref|NP_909861.1| putative transposase [Oryza sativa]	3,0E-53	CO439058
MZCCL10001E03.g	ref|XP_464346.1| putative protein kinase 2 [Oryza sativa]	1,0E-142	CO439063
MZCCL10001F11.g	ref|XP_476301.1| unknown protein [Oryza sativa]	2,0E-66	CO439077
MZCCL10001G09.g	ref|NP_910634.1| P0534A03.14 [Oryza sativa]	1,0E-56	CO449900
MZCCL10001G10.g	gb|AAP32017.1| gamma zein [*Zea mays*]	1,0E-26	CO454286
MZCCL10002A07.g	no hits	-	CO439101
MZCCL10002A08.g	dbj|BAD01240.1| AP2 domain-containing protein AP29-like [Oryza sativa]	1,0E-114	CO439102
MZCCL10002B09.g	dbj|BAD09117.1| putative uridine kinase/uracil phosphoribosyltransferase [Oryza sativa]	5,0E-47	CO439112
MZCCL10002C05.g	ref|NP_908624.1| putative sarcosine oxidase [Oryza sativa]	4,0E-50	CO451769
MZCCL10002C07.g	gb|AAU44042.1| putative circadian clock coupling factor ZGT [Oryza sativa]	2,0E-23	CO439119
MZCCL10002C10.g	ref|NP_917194.1| P0707D10.11 [Oryza sativa]	4,0E-16	CO453469
MZCCL10002E04.g	ref|XP_507054.1| PREDICTED OJ1202_E07.24 gene product [Oryza sativa]	1,0E-54	CO439135
MZCCL10002F01.g	ref|NP_914460.1| gigantea-like protein [Oryza sativa]	2,0E-46	CO439143
MZCCL10002F05.g	emb|CAE03144.2| OSJNBa0081L15.6 [Oryza sativa]	1,0E-145	CO439146
MZCCL10002F07.g	gb|AAL16995.1| Hageman factor inhibitor [*Zea mays*]	1,0E-86	CO441034
MZCCL10002G09.g	no hits	-	CO439155
MZCCL10003A12.g	gb|AAL16980.1| 15kD beta zein [*Zea mays*]	4,0E-45	CO452643
MZCCL10003C02.g	ref|XP_478367.1| chorismate mutase/prephenate dehydratase-like protein [Oryza sativa]	9,0E-61	CO439182
MZCCL10003C04.g	ref|XP_468052.1| kinesin motor protein 1-like [Oryza sativa]	2,0E-25	CO439184
MZCCL10003C05.g	dbj|BAD45504.1| phospholipase -like [Oryza sativa]	2,0E-38	CO439185
MZCCL10003D05.g	emb|CAA69075.1| S-adenosylmethionine decarboxylase [*Zea mays*]	2,0E-26	CO439192
MZCCL10003E12.g	pir||JQ1005 glucose-1-phosphate adenylyltransferase (EC 2.7.7.27) [*Zea mays*]	0	CO467438
MZCCL10003H02.g	ref|XP_470218.1| Putative retroelement [Oryza sativa]	1,0E-155	CO439222
MZCCL10003H04.g	no hits	-	CO445753
MZCCL10004A07.g	ref|XP_462835.1| B1085F09.24 [Oryza sativa]	3,0E-24	CO443552
MZCCL10004A11.g	no hits	-	CO439238
MZCCL10004D01.g	no hits	-	CO439257
MZCCL10004F07.g	ref|NP_917778.1| P0006C01.15 [Oryza sativa]	1,0E-18	CO454844
MZCCL10004G12.g	no hits	-	CO439289
(...)			

a The Accession Number corresponds to the longest sequence-read of the respective MAS

Observação: Esta tabela contém 4.034 linhas, mas apenas as 35 primeiras linhas e a nota de rodapé estão representadas aqui. A tabela completa pode ser obtida através do endereço eletrônico http://www.springerlink.com/(qxyuv3bcuhbtevygc1ldcum3)/app/home/contribution.asp?referrer=parent&backto=issue,10,10;journal,9,339;linkingpublicationresults,1:100330,1/, no link Electronic Supplementary Material – OPEN ESM.

Supplemental Table 4. GO terms for the 2,403 anotatted endosperm-preferred MAS

MAS	Molecular function ontology	Biological process ontology
MZCCL10142D02.g	GO:0003700	GO:0045449
MZCCL10075F08.g	GO:0005554	GO:0000004
ZMZZEN6067F10.g	GO:0005524	GO:0007046
MZCCL10093F02.g	GO:0005554	GO:0000004
MZCCL15026A08.g	GO:0004826	GO:0006432
MZCCL10125C02.g	GO:0004497, GO:0016491	GO:0006118, GO:0006725
MZCCL15009G01.g	GO:0005554	GO:0000004
ZMZZEN5004F09.g	GO:0005554	GO:0000004
ZMZZEN7041H11.g	GO:0003723, GO:0003735	GO:0042254
MZCCS20027A10.g	GO:0005554	GO:0000004
ZMZZEN6096H04.g	GO:0005554	GO:0000004
MZCCL10112A11.g	GO:0003743	GO:0007275
ZMZZEN6044H12.g	GO:0004386	GO:0006268
MZCCL10094H09.g	GO:0003676, GO:0003723	
ZMZZEN1034G12.g	GO:0004553, GO:0016787	GO:0005975
MZCCL10004H06.g	GO:0045735	GO:0019538
MZCCL15026F12.g	GO:0045544	GO:0045487
MZCCL10214E10.g	GO:0005554	GO:0000004
MZCCL10005F02.g	GO:0003999, GO:0016757	GO:0006168, GO:0009116
MZCCL10011G05.g	GO:0003676, GO:0004474	GO:0006097, GO:0006099
ZMZZEN6071H03.g	GO:0005554	GO:0000004
MZCCL10016E08.g	GO:0005554	GO:0000910
MZCCL10006H10.g	GO:0004553	GO:0005975
ZMZZEN7035A09.g	GO:0008415	GO:0000004
MZCCL20041C08.g	GO:0004672, GO:0005524	GO:0006468
MZCCL10054F01.g	GO:0005488	GO:0006839
ZMZZEN6088A07.g	GO:0003872	GO:0006096
ZMZZEN6004E05.g	GO:0005554	GO:0000004
ZMZZEN5044C05.g	GO:0004867	GO:0009611
MZCCS20013E07.g	GO:0005554	GO:0000004
MZCCL10081A09.g	GO:0005554	GO:0000004
MZCCL10032E03.g	GO:0005554	GO:0000004
MZCCL20005G01.g	GO:0045735	GO:0019538
MZCCL10147H04.g	GO:0016758	GO:0005975, GO:0030259
ZMZZEN5019B07.g	GO:0005489	GO:0006118, GO:0016070

Observação: Esta tabela contém 2.045 linhas, mas apenas as 37 primeiras linhas estão representadas aqui. A tabela completa pode ser obtida através do endereço eletrônico http://www.springerlink.com/(qxyuv3bcuhbtevygc1ldcum3)/app/home/contribution.asp?referrer=parent&backto=issue,10,10;journal,9,339;linkingpublicationresults,1:100330,1/, no link Electronic Supplementary Material – OPEN ESM.

Endosperm-preferred transcription factors involved in maize seed development

Natalia C. Verza, Sylvia M. Sousa, Paulo H. Fisch, Thaís R. Silva, Marcelo M. Rebello and Paulo Arruda

Endosperm-preferred transcription factors involved in maize seed development

Natalia C. Verza[1], Sylvia M Sousa[1], Paulo H. Fisch[1], Thaís R. Silva[1], Marcelo M. Rebello[1] and Paulo Arruda[1,2*]

[1]Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas (UNICAMP), 13.083-970, Campinas, SP, Brazil.

[2]Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas (UNICAMP), 13.083-970, Campinas, SP, Brazil.

* Corresponding author

E-mail addresses:

NCV: verza@unicamp.br

SMS: smsousa@unicamp.br

PHF: fisch@unicamp.br

TRS: trsilva@unicamp.br

MMR: rebello@unicamp.br

PA: parruda@unicamp.br

Keywords: maize endosperm, development, transcription factors.

Abstract

- Background

We have recently created a database of maize ESTs called MAIZEST (www.maizest.unicamp.br), that focuses on genes expressed in developing endosperm. The MAIZEST compiles over 227,000 maize ESTs, 64,537 of which coming from developing endosperm. This database contains over 80% of the genes expressed in the endosperm, and is a powerful tool for genome-wide approaches of data mining and gene discovery. In this work we describe the identification of maize TFs preferentially expressed in developing endosperm.

- Results

We identified 1,233 maize TFs, 414 of which coming from endosperm libraries. We also identified 113 putative endosperm-preferred TFs, represented by 326 ESTs from developing endosperm. This endosperm-preferred set accounts for 9.2% of all identified maize TFs, and may represent part of the regulators involved in endosperm specification and development. The most represented TF family among the endosperm-preferred TFs was the Zinc-finger domain family (13,2%) followed by the NAM family (10,6%) and the bZIP family (17%). This indicates that these are probably important regulators of the endosperm development.

- Conclusion

This paper describes the identification of an extensive collection of maize endosperm-preferred transcription factors, which represents an important source of potential candidates for regulators of major aspects of the endosperm development, such as nitrogen and carbohydrate metabolism and control of seed mass.

Background

Transcription factors (TFs) are sequence-specific DNA binding proteins that are capable of activating and/or repressing transcription. They are mainly responsible for the selectivity in gene regulation, and are often expressed in a tissue-specific, developmental-stage-specific, or stimulus-dependent manner [1]. TF genes constitute a considerable proportion of the eukaryotic genome. The programmed and regulated interactions between TFs and genomic DNA bring a genome to its life and define many of its functional features [2]. Many mutants impaired in their development or metabolic processes have been associated with altered expression of TF genes (for reviews: [1], [3]).

TFs can be grouped into gene families according to the type of DNA-binding domain they encode. Functional redundancy is not unusual within TF families; therefore the proper characterization of particular transcription-factor genes often requires their study in the context of a whole family. The largest transcription factor family found in the eukaryotic genomes contains a DNA-binding motif known as zinc finger. Based on the structure and spatial arrangement of this domain, the zinc finger gene family can be further subdivided into several classes, including plant-specific ones, such as WRKY [4], YABBY [5] and Dof [6]. Plant zinc finger TFs have been associated to the regulation of several biological processes, such as flower development, leaf and lateral shoot initiation, salt tolerance, carbon and nitrogen metabolism and seed development (for reviews: [4], [7]). Another important TF class in plants is the bZIP family. Its estimated that in *Arabidopsis* and rice 5,28% and 5,74% of their TF content, respectively, are represented by bZIPs, which are about four times as many bZIP genes as yeast, worm and human ([8], [9]). bZIPs regulate diverse biological processes such as pathogen defense, light and stress signaling, seed maturation and flower development (reviewed in [10]). Much of what is known about the genetic and molecular mechanisms regulating seed storage compounds comes from studies of a bZIP protein originally isolated from maize, the Opaque 2 (O2; [11], [12]). The O2 protein binds to and activates transcription from diverse motifs in promoters of genes encoding storage proteins, and enzymes of carbohydrate and amino acid metabolism. O2 is involved in the coordinated regulation of protein synthesis, nitrogen and sugar metabolism during the maturation of maize seeds [13].

The necessity of using genomic approaches becomes clear when it is considered that less than 10% of the *Arabidopsis* transcription factors have been genetically characterized [3], and even a smaller fraction of the TF content is known in maize. We have created a large database enriched in genes expressed in developing maize endosperm called MAIZESTdb (www.maizest.unicamp.br; [14]). The MAIZESTdb compiles over 227,000 maize ESTs, 64,537 coming from developing endosperm, clustered into 29,206 maize assembled sequences (MAS). This database contains over 80% of the genes expressed in the endosperm, and is a powerful tool for genome-wide approaches for data mining and gene discovery. In this work we focused on the identification of maize TFs expressed in developing endosperm. The results are discussed in the context of the regulatory components of the complex network underlying seed development.

Results and discussion

Transcription factors expressed in maize endosperm

The cereal endosperm is a suitable model for gene regulation studies [15]. Maize endosperm begins as a triploid tissue with the union of two polar nuclei and one sperm nucleus. For the first 4 days after pollination (DAP), the endosperm nuclei divide synchronously without cell wall formation. The process of cellularization of the endosperm coenocyte is completed up to 4 DAP in maize, when the tissue changes from a multinucleate single cell to a uninucleate multicellular morphology. Most of the endosperm cells are produced between 4 and 12 DAP, with the mitotic index peaking between 6 and 8 DAP. Around 12 DAP, the endosperm begins to accumulate large amounts of starch and storage proteins. By 16 DAP, the maturation program has initiated, preparing the seeds for desiccation and dormancy, and by 23 DAP desiccation has begun. At around 25-30 DAP, the relative water content of the endosperm begins to decrease, and the seed desiccation is a signal to arrest germinative development (reviewed in: [15]; [16]; [17]).

The regulation of gene expression in a particular cell type depends on the activity of different types of TFs: those expressed in most cells/tissues, probably regulating the basic cell metabolism, and those expressed specifically or preferentially in that particular cell/tissue type. The cell-specific TFs are, most probably, the responsible for the cell specification and development. To identify these two types of TFs expressed in

the developing maize endosperm, the 29,206 Maize Assembled Sequences (MASs) of MAIZEST database [14] were compared to the TRANSFAC, the Pfam and the GenBank databases. These searches resulted in the identification of 1,233 (4,2% of the MAS set) MASs representing TFs, 414 of which coming from endosperm libraries (Table 1).

The frequency of ESTs for individual genes in diverse cDNA libraries can be used to estimate the expression patterns of these genes [18]. By searching for MASs composed only by ESTs originated from endosperm libraries, we identified 113 putative endosperm-preferred TFs, distributed among 53 contigs and 60 singletons (Table 1). These 113 MASs are composed by 326 ESTS from developing endosperm, and may represent part of the regulators involved in endosperm specification and development.

The *Arabidopsis* genome codes for ~1,533 transcriptional regulators, which account for ~5.9% of its estimated total number of genes [8]. If the number of genes in maize is similar to that of rice, which is estimated to be around 40,000 genes [19, 20], and maize contains TFs in a proportion similar to that of *Arabidopsis*, one could estimate that the maize genome codes for ~2,300 TFs. We analyzed a collection of 227,000 ESTs corresponding to 24,000 putative genes (~60% of the maize genome [14]) and obtained 1,233 TFs expressed in all maize tissues, which is in good agreement with the expected number. The 414 TFs expressed in developing endosperm accounts for 33% of the identified maize TFs, and the endosperm-preferred set of TFs identified accounts for 9.2% of all identified maize TFs.

To estimate the level of redundancy among the MAS sequences representing endosperm expressed TFs, the 414 TF MASs were compared with each other using BLASTN [21]. Two sequences were considered as originating from the same transcript when they had 98% nucleotide identity over a minimum of 100 bp. The comparison was made within each TF family, and the average redundancy found was 10,4%, indicating that we have identified at least 369 endosperm-expressed TFs. Information about sequences, library contribution and annotation for all of the 414 MAS can be accessed through the MAIZEST database (www.maizest.unicamp.br; [14]).

Classification of the maize transcription factors

In accordance with the structural features of the DNA-binding domains that they encode, TFs can be grouped into distinct families. We used the TRANSFAC classification

[21] for the distribution of the endosperm-preferred TFs identified, including those identified in the Pfam [22] and in the GenBank [23] databases. When classification was not possible, TFs were placed in the "Other" group. The distribution of these endosperm-preferred TFs among the main families is shown in Table 2.

The most represented TF family expressed in the developing maize endosperm was the Zinc-finger domain, with 50 MAS (12,1% of the TFs), followed by the Homeodomain family, with 38 MASs (9,2%) and the bZIP family, with 28 MASs (6,7%) (Figure 1). A different distribution was found among the endosperm-preferred TFs (Figure 1). The Zinc-finger domain family remained as the most represented one, with 13,2% of the endosperm-preferred TFs (15 MASs), while the NAM family was the second, with 10,6% (12 MASs), and the bZIP family had 9,7% (Figure 1). This indicates that these families of TFs are probably more important for the regulation of endosperm development.

Functional annotation of endosperm-preferred-transcription factors

Nitrogen and carbohydrate metabolism

The nitrogen and carbohydrate metabolisms in developing endosperm require a strikingly coordination of complex processes ([25], [26]), involving a regulatory network of TFs among other regulatory processes. The bZIP Opaque-2 (O2) gene is the most studied maize endosperm-preferred TF involved in nitrogen and carbohydrate metabolism. The recessive opaque-2 (o2) mutation gives an opaque character to the mature seed, and produces a very marked decrease in the prolamin storage protein content, mainly the 22-kD [alpha]-zein, while the proportions of lysine and tryptophan are increased, producing grains with improved nutritional quality. Various aspects of endosperm metabolism are modified in o2 seeds: RNase activity is higher in o2 than in wild-type [27], amino acid metabolism, especially aspartate metabolism, appeared to be altered [13] as well as the expression of various enzymes related to nitrogen and sugar metabolism ([25]; [28]). Finally, mutant kernels are more susceptible to plant pathogens and yield is decreased [29]. Studies have shown that key enzymes involved in amino acid and carbon metabolism are altered in o2 mutants. The activity of Aspartate kinase (AK), an important enzyme involved in the synthesis of several amino acids, including Thr, Lys, Met, and Leu, is up-regulated by o2 [30]. However, the effect of o2

on AK must be indirect, as there is no evidence that O2 inhibits the expression of the gene. The o2 mutation also affects the activity of the bifunctional lysine ketoglutarate reductase-saccaropine dehydrogenase (LKR-SDH), which regulates Lys degradation in maize endosperm. Kemper et al. [31] showed that LKR-SDH is down-regulated by the o2 mutation as a consequence of reduced levels of mRNA.

Other TFs have recently been associated with sugar and nitrogen metabolism. Maize Dof1 is a member of the Dof TF family unique to plants that has been shown to regulate several genes involved in carbohydrate metabolism. Overexpression of Dof1 in transgenic *Arabidopsis* caused a remarkable rise in amino acid concentrations, especially in the glutamine level, and an elevation in the nitrogen content [6]. In addition, a transcription profiling in response to sugar using microarray in *Arabidopsis* showed that glucose treatments affected several families of TFs, including bHLH, MYB, AP2, and various zinc finger–containing factors [32].

In the present study we identified 15 Zinc-domain containing endosperm-preferred TFs, including one Dof, 6 MYB family members, 6 bHLH family members and 5 AP2 factors, totalizing 32 endosperm-preferred as potential candidates for regulating nitrogen and carbohydrate metabolism in developing endosperm.

Control of seed mass

In angiosperms, seed development depends on the interaction between the triploid endosperm and the diploid sporophytic and embryonic genomes to orchestrate morphogenesis and the deposition of seed reserves in the developing seed [33]. Because the maternal plant contributes with two genome equivalents to the triploid tissue, the endosperm has been implicated to serve as the site of parent-of-origin effects on seed mass through the imprinting of genes thought to be involved in enhancing or suppressing endosperm size and, therefore, seed and embryo mass [34]. It was shown that, in *Arabidopsis* and maize, the endosperm plays a central role in the control of seed size ([35], [36]).

Recent studies in *Arabidopsis* reported that a member of AP2/EREBP family, Apetala2, controls seed mass, in part through its activity in the maternal sporophyte and endosperm ([37], [38]). Members of this plant-specific family of TFs play a variety of roles throughout the plant life cycle, from being key regulators of several

developmental processes, like floral organ identity determination, control of leaf epidermal cell identity and germination, to forming part of the mechanisms used by plants to respond to various types of biotic and environmental stress [39]. AP2/EREBP genes can be found expressed not only in flowers, but also in leaves, stems, seedlings and seeds, suggesting that they might be involved in a range of functions. Three Apetala2-like genes strongly expressed in endosperm were identified in Petunia [40]. Their expression pattern resembles that of prolamin seed storage proteins of maize, which start to be expressed coordinately in the maize endosperm at 8 to 12 days after pollination.

We have found 26 MAS belonging to the AP2/EREBP family of TFs, and five of these presented endosperm-preferred expression. These evidences make these transcription factors, specially the endosperm-preferred ones, good candidates for controlling seed mass in maize.

Regulation of endosperm development

In plants, insects and mammals, Polycomb group (PcG) proteins are involved in the regulation of various developmental processes ([41], [42], [43], [44]). Examples of PcGs are the Fertilization-Independent Endosperm gene (FIE), that regulates endosperm and embryo development and represses flowering during embryo and seedling development, and MEDEA (MEA), that functions as a suppressor of endosperm development [45]. FIE and MEA form a PcG complex that regulates endosperm and embryo development [46]. FIE and MEA, and Fertilization-Independent Seed2 (FIS2), a zinc finger protein, were shown to control expression of Pheres1 (PHE1), a MADS-box gene which regulates seed development [47]. In addition to control MADS-box genes in plants, PcG proteins control expression of homeobox genes in *Arabidopsis* [48], and this function seems to be conserved, as in mammals and insects PcG proteins also control the expression of homeotic genes .

This intricate and complex regulatory network involved in endosperm development involves TFs from four different families. We have identified five PcGs, two of them with endosperm-preferred expression. In addition, the new TF collection has 15 Zinc-domain MASs, 4 MADS-box MASs and 7 Homeobox MASs, all of them

presenting endosperm-preferred expression, representing possible targets for the PcG regulatory complex.

Stage-specific TFs

Large sets of ESTs from redundant non-normalized cDNA libraries can be used to evidence differential expression of individual genes in distinct tissues and/or developmental stages. In our study, since some endosperm libraries were constructed from endosperm RNA extracted at distinct and defined stages of endosperm development, we were able to identify TFs preferentially expressed at different stages (Table 3). The endosperm-preferred TFs were searched for sequences coming from endosperms at 10 days after pollination (DAP), 15 DAP and 20 DAP. Of the 113 endosperm-preferred TFs, information about endosperm developmental stage was available for 81 TFs. The relative expression of each one was calculated as the number of ESTs from a given stage in its MASs, divided by the total number of ESTs available for this stage. We found 36 MASs expressed only at 10 DAP, 12 MASs expressed only at 15 DAP and 18 MASs expressed only at 20 DAP (Table 3).

TFs presenting preferred early expression, that are probably involved in specification of cell fate, include members of the Polycomb Group, such as FIE1 and Enhancer of zeste-like protein 2, that are known to maintain homeotic gene repression to control cell identity and differentiation and Homeobox family members, related to the regulation of growth patterns and cell-fate acquisition, and that was shown to be regulated by PcG genes [49]. Some NAC-family genes also presented an early expression pattern in the endosperm, and we couldn't find previous report of that. These genes represent new candidates involved in the early stages of the endosperm developmental pathway.

bZIP family members can be found expressed at all stages during development, although members of the NAC and the Zinc domains families usually presented a late expression pattern, concentrating the expression at 15 DAP and 20 DAP.

NAC is a multigenic family of TFs specific to plants and are found to play roles in a diverse set of developmental processes, including developmental programmes, defense and abiotic stress responses (reviewed in [50]). A member of NAC family, AtNAM was found to be up-regulated in *Arabidopsis* developing seeds [51], and a NAM-related

protein (NRP1) of maize was already reported as an endosperm-specific member of this family [52]. We have found 24 MASs corresponding to NAM-family TFs, 12 of them having endosperm-preferred expression. The majority of the NAM members had a late expression pattern, in special the endosperm-preferred ones. After 20 DAP, when the NAM transcripts seems to accumulate in endosperm, the relative water content of the endosperm begins to decrease, and the initiation of seed desiccation provides a signal to arrest germinative development. It is possible that these TFs are involved in the regulation of the desiccation process, in response to the hydric stress accompanying seed desiccation, and in the transition from the maturation to the germination process.

The expression pattern of a subset of endosperm-preferred-transcription factors corroborates the in silico findings

In order to access the accuracy of the in silico approaches to identify endosperm-preferred genes, we used RT-PCR to perform an expression profile analysis of five TFs selected among the 113 endosperm-preferred group. We used samples from 15 DAP endosperm, young leaves and roots, coleoptiles and 30 DAP embryos. As shown in Figure 2, all of the tested genes presented an endosperm-preferred expression, including four novel maize genes, an EREBP-family like TF, a NAM-family like TF, a PHD finger protein-related TF and a Zinc finger family PCP-1 like protein. Two genes were used as controls: the endosperm-specific Opaque-2 and the housekeeping α-tubulin. The experimental procedure, with the presence of a gene known to be preferentially expressed in maize seeds, demonstrates that the computational-based procedure identified genes specifically, or, at least, predominantly expressed in developing endosperm, and constitutes a valuable tool for gene discovery.

Conclusion

We reported here the identification of a collection of 414 maize endosperm-expressed transcription factors, 113 of which being preferentially expressed in developing endosperm. This is the most extensive collection of endosperm-preferred transcription factors reported, and represents an important source of potential candidates for main regulators of important aspects of endosperm development, such as

nitrogen and carbohydrate metabolism and control of seed mass. These endosperm-preferred genes are also good candidates for studies of gene-function relationships by screening populations of maize mutants. A better understanding of the complex mechanisms involved in regulation of the endosperm development can provide tools for genetically engineered plants with improved seeds.

Methods

Plant material

Maize (*Zea mays* L.) plants from the Oh43 inbred line were grown in the greenhouse. Immature ears were harvested before self pollination. The upper third of the endosperms, containing only endosperm, aleurone and pericarp tissues were harvested at 10, 15 and 20 days after pollination (DAP). Embryos were dissected manually. Roots, leaves and coleoptiles were harvested from 5-day-old seedlings germinated under controlled conditions. All the tissues were immediately frozen in liquid nitrogen and stored at -80 ºC.

RNA extraction and RNA-blot analysis

Total RNA was isolated from frozen material as described by Manning (1991; [53]) for endosperm and embryo tissues and using the TRIzol reagent (Invitrogen, Carlsbad, CA) for roots, leaves and coleoptiles. The purity and integrity of the RNA were assessed by the absorbance at 260/280 nm and agarose gel electrophoresis.

Ten micrograms of total RNA were electrophoresed in a 1% (w/v) agarose gel containing formaldehyde and transferred to a Hybond-N^+ filter (Amersham Biosciences) as described by Sambrook et al. (1989). The filters were hybridized with the cDNA inserts of transcription factors labeled with [α-32P]dCTP, and hybridization was done at 42°C (Sambrook et al.,1989). The blots were then washed at high stringency and exposed to imaging plates. Images were obtained using the Image Gauge software (Fujifilm).

Transcription factors identification

Putative transcription factors were identified by sequence homology based on the screening of the entire MAIZESTdb maize ESTs database (http://www.maizest.unicamp.br; [14]) using the TRANSFAC Professional data set (release 8.2 Professional; [22]), the Genbank [24] and a subset of transcription factor domains from Pfam database (release 7.0; [23]). We combined automated search and manual curation to generate a collection of endosperm-expressed transcription factors as complete as possible.

First, the 5,597 transcription factors in the TRANSFAC Professional data set were used to perform BLASTX [21] searches against the 29,206 MASs from MAIZESTdb. Only matches with an E value ≤ 1.E- 15 were included, and the TF classification from TRANSFAC was maintained. In order to eliminate false positives, the nucleotide sequences of the MASs retrieved from the first search were then compared to the complete set of proteins from all organisms available from GenBank, using BLASTX. All results from BLAST searches were manually inspected, and it were removed some false positive matches including proteases, splicing factors, kinases, translation factors and many others that are not transcription factors.

In a parallel protein domain and motif analysis, the entire complement of MASs was searched with a subset of TF motifs selected from Pfam 7.0, using the default settings and the Pfam gathering threshold [23], and the matches with transcription factor motifs not identified yet were included. These MASs were allocated in one of the TRANSFAC TF classes when it was possible, or included in the "Other" group.

In silico endosperm-preferred TFs identification

The MAIZESTdb [14] MASs 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. Thus, the MAIZESTdb tools allowed the analysis of ESTs distribution among MASs, and made it possible to infer the likelihood of tissue-specific expression. MASs consisting of ESTs derived exclusively from endosperm libraries were considered endosperm-preferred transcripts.

Author's contribution

NCV performed the in silico analysis and the experimental procedures, and wrote the manuscript jointly with PA. SMS helped with the experimental procedures. PHF, MMR and TRS helped with the database searches and the in silico analysis. PA supervised the study.

Acknowledgements

NCV was supported by a postgraduate fellowship from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), TRS was supported by a postgraduate fellowship from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), and SMS was supported by a postgraduate fellowship from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References

1. Zhang JZ: **Overexpression analysis of plant transcription factors.** *Current Opinion in Plant Biology* 2003, **6:** 430-440.
2. Gong W, Shen YP, Ma LG, Pan Y, Du YL, Wang DH *et al.*: **Genome-wide ORFeome cloning and analysis of *Arabidopsis* transcription factor genes.** *Plant Physiology* 2004, **135:** 773-782.
3. Riechmann JL, Ratcliffe OJ: **A genomic perspective on plant transcription factors.** *Curr Opin Plant Biol* 2000, **3:** 423-434.
4. Takatsuji H: **Zinc-finger transcription factors in plants.** *Cellular and Molecular Life Sciences* 1998, **54:** 582-596.
5. Bowman JL: **The YABBY gene family and abaxial cell fate.** *Current Opinion in Plant Biology* 2000, **3:** 17-22.
6. Yanagisawa S: **Dof DNA-binding proteins contain a novel zinc finger motif.** *Trends in Plant Science* 1996, **1:** 213-214.
7. Takatsuji H: **Zinc-finger proteins: the classical zinc finger emerges in contemporary plant science.** *Plant Molecular Biology* 1999, **39:** 1073-1078.
8. Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J *et al.*: ***Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes.** *Science* 2000, **290:** 2105-2110.
9. Yu J, Hu SN, Wang J, Wong GKS, Li SG, Liu B *et al.*: **A draft sequence of the rice genome (Oryza sativa L. ssp indica).** *Science* 2002, **296:** 79-92.

10. Jakoby M, Weisshaar B, Droge-Laser W, Vicente-Carbajosa J, Tiedemann J, Kroj T *et al.*: **bZIP transcription factors in *Arabidopsis*.** *Trends in Plant Science* 2002, **7:** 106-111.

11. Hartings H, Maddaloni M, Lazzaroni N, Difonzo N, Motto M, Salamini F *et al.*: **The O2 Gene Which Regulates Zein Deposition in Maize Endosperm Encodes A Protein with Structural Homologies to Transcriptional Activators.** *Embo Journal* 1989, **8:** 2795-2801.

12. Schmidt RJ, Burr FA, Aukerman MJ, Burr B: **Maize regulatory gene opaque-2 encodes a protein with a "leucine-zipper" motif that binds to zein DNA.** *Proc Natl Acad Sci U S A* 1990, **87:** 46-50.

13. Yunes JA, Cord NG, Leite A, Ottoboni LM, Arruda P: **The role of the Opaque2 transcriptional factor in the regulation of protein accumulation and amino acid metabolism in maize seeds.** *An Acad Bras Cienc* 1994, **66 Su 1 (Pt 2):** 227-237.

14. Verza NC, Silva TR, Cord-Neto G, Nogueira FTS, De Rosa Jr VE, Fisch PH *et al.*: **Endosperm-preferred expression of maize genes as revealed by transcriptome-wide analysis of expressed sequence tags.** *Plant Molecular Biology* 2005, **59:** 361-372.

15. Olsen OA: **Nuclear endosperm development in cereals and *Arabidopsis thaliana*.** *Plant Cell* 2004, **16:** S214-S227.

16. Olsen OA: **Endosperm development: Cellularization and cell fate specification.** *Annual Review of Plant Physiology and Plant Molecular Biology* 2001, **52:** 233-+.

17. Lopes MA, Larkins BA: **Endosperm origin, development, and function.** *Plant Cell* 1993, **5:** 1383-1399.

18. Fernandes J, Brendel V, Gai X, Lal S, Chandler VL, Elumalai RP *et al.*: **Comparison of RNA expression profiles based on maize expressed sequence tag frequency analysis and micro-array hybridization.** *Plant Physiol* 2002, **128:** 896-910.

19. Yu J, Hu SN, Wang J, Wong GKS, Li SG, Liu B *et al.*: **A draft sequence of the rice genome (Oryza sativa L. ssp indica).** *Science* 2002, **296:** 79-92.

20. Lai JS, Dey N, Kim CS, Bharti AK, Rudd S, Mayer KFX *et al.*: **Characterization of the maize endosperm transcriptome and its comparison to the rice genome.** *Genome Research* 2004, **14:** 1932-1937.

21. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W *et al.*: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:** 3389-3402.

22. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R *et al.*: **TRANSFAC (R): transcriptional regulation, from patterns to profiles.** *Nucleic Acids Research* 2003, **31:** 374-378.

23. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S *et al.*: **The Pfam protein families database.** *Nucleic Acids Research* 2004, **32:** D138-D141.

24. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucleic Acids Research* 2005, **33:** D34-D38.

25. Giroux MJ, Boyer C, Feix G, Hannah LC: **Coordinated Transcriptional Regulation of Storage Product Genes in the Maize Endosperm.** *Plant Physiol* 1994, **106:** 713-722.

26. Ho S-L, Chao Y-C, Tong W-F and Yu S-M: **Sugar coordinately and differentially regulates growth- and stress-regulated gene expression via a complex signal transduction network and multiple control mechanisms.** *Plant Physiol. 2001*, **125:** 877-890.

27. Dalby A, Davies II: **Ribonuclease activity in the developing seeds of normal and opaque-2 maize.** *Science. 1967*, 155(769):1573–1575.

28. Lodha ML, Mali PC, Agarwal AK, Mehta SL: **Changes in soluble protein and isoenzymes in normal and opaque-2 Zea mays endosperm during grain development**. *Phytochemistry. 1974,* 13:539-542.

29. Loesch Jr. PJ, Foley DC, COX DF: **Comparative resistance of Opaque-2 an Normal inbred lines of maize to ear-rotting pathogens.** Crop *Science. 1976,* 16:841-842.

30. Brennecke K, Souza Neto AJ, Lugli J, Lea, PJ, Azevedo RA: **Aspartate kinase in the maize mutants ask1-lt19 and opaque-2.** *Phytochemistry. 1996,* 41:707-712.

31. Kemper EL, Cord Neto G, Papes F, Martinez Moraes KC, Leite A, Arruda P: **The role of Opaque2 in the control of lysine-degrading activities in developing maize endosperm.** *Plant Cell. 1999,* 11:1981-1993.

32. Price J, Laxmi A, St Martin SK, Jang JC: **Global transcription profiling reveals multiple sugar signal transduction mechanisms in *Arabidopsis*.** *Plant Cell* 2004, **16:** 2128-2150.

33. Alonso-Blanco C, Blankestijn-de Vries H, Hanhart CJ, Koornneef M: **Natural allelic variation at seed size loci in relation to other life history traits of *Arabidopsis thaliana*.** *Proceedings of the National Academy of Sciences of the United States of America* 1999, **96:** 4710-4717.

34. Gehring M, Choi Y, Fischer RL: **Imprinting and seed development.** *Plant Cell* 2004, **16:** S203-S213.

35. Lin BY: **Ploidy Barrier to Endosperm Development in Maize.** *Genetics* 1984, **107:** 103-115.

36. Scott RJ, Spielman M, Bailey J, Dickinson HG: **Parent-of-origin effects on seed development in *Arabidopsis thaliana*.** *Development* 1998, **125:** 3329-3341.

37. Jofuku KD, Omidyar PK, Gee Z, Okamuro JK: **Control of seed mass and seed yield by the floral homeotic gene APETALA2.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102:** 3117-3122.

38. Ohto M, Fischer RL, Goldberg RB, Nakamura K, Harada JJ: **Control of seed mass by APETALA2.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102:** 3123-3128.

39. Riechmann JL, Meyerowitz EM: **The AP2/EREBP family of plant transcription factors.** *Biol Chem* 1998, **379:** 633-646.

40. Maes T, Van de Steene N, Zethof J, Karimi M, D'Hauw M, Mares G *et al.*: **Petunia Ap2-like genes and their role in flower and seed development.** *Plant Cell* 2001, **13:** 229-244.

41. Kohler C, Grossniklaus U: **Epigenetic inheritance of expression states in plant development: the role of Polycomb group proteins.** *Current Opinion in Cell Biology* 2002, **14:** 773-779.

42. Berger F, Gaudin V: **Chromatin dynamics and *Arabidopsis* development.** *Chromosome Research* 2003, **11:** 277-304.

43. Sung ZR, Chen LJ, Moon YH, Lertpiriyapong K: **Mechanisms of floral repression in *Arabidopsis*.** *Current Opinion in Plant Biology* 2003, **6:** 29-35.

44. Wagner D: **Chromatin regulation of plant development.** *Current Opinion in Plant Biology* 2003, **6:** 20-28.

45. Kiyosue T, Ohad N, Yadegari R, Hannon M, Dinneny J, Wells D *et al.*: **Control of fertilization-independent endosperm development by the MEDEA polycomb gene *Arabidopsis*.** *Proceedings of the National Academy of Sciences of the United States of America* 1999, **96:** 4186-4191.

46. Kohler C, Hennig L, Bouveret R, Gheyselinck J, Grossniklaus U, Gruissem W: ***Arabidopsis* MSI1 is a component of the MEA/FIE Polycomb group complex and required for seed development.** *Embo Journal* 2003, **22:** 4804-4814.

47. Kohler C, Hennig L, Spillane C, Pien S, Gruissen W, Grossniklaus U: **The Polycomb-group protein MEDEA regulates seed development by controlling expression of the MADS-box gene PHERES1.** *Genes & Development* 2003, **17:** 1540-1553.

48. Katz A, Oliva M, Mosquna A, Hakim O, Ohad N: **FIE and CURLY LEAF polycomb proteins interact in the regulation of homeobox gene expression during sporophyte development.** *Plant Journal* 2004, **37:** 707-719.

49. Scheres B: Plant **Cell Identity. The Role of Position and Lineage.** *Plant Physiol. 2001,* 125:112-114.

50. Olsen AN, Ernst HA, Lo Leggio L, Skriver K: **NAC transcription factors: structurally distinct, functionally diverse.** *Trends in Plant Science* 2005, **10:** 79-87.

51. Duval M, Hsieh TF, Kim SY, Thomas TL: **Molecular characterization of AtNAM: a member of the *Arabidopsis* NAC domain superfamily.** *Plant Mol Biol* 2002, **50:** 237-248.

52. Guo M, Rupe MA, Danilevskaya ON, Yang XF, Hut ZH: **Genome-wide mRNA profiling reveals heterochronic allelic variation and a new imprinted gene in hybrid maize endosperm.** *Plant Journal* 2003, **36:** 30-44.

53. Manning K: **Isolation of nucleic acids from plants by differential solvent precipitation.** *Analytical Biochemistry* 1991, **195:** 45-50.

Tables

Table1

Table 1. Transcription factors (TFs) expressed in developing maize endosperm

Number of sequences analized[1]	227,431
Number of sequences from endosperm[2]	64,537
Number of MAS[3] analized	29,206
Number of MAS representing TFs[4]	1,233
Number of TF MAS expressed in endosperm	414
Endosperm-preferred MAS[5]	113

[1]Total sequences from MAIZEST database [17]

[2]Total sequences derived from developing endosperm cDNA libraries.

[3]MASs, Maize Assembled Sequences [17]

[4]Transcription factor (TF) sequences were identified by comparing the MAS set with TRANSFAC Professional 8.2 (Biobase), GenBank, and with a set of TF domains from the Pfam database

[5]Sequences that appear only in the endosperm libraries

Table 2

Table 2. Distribution of endosperm-preferred transcription factors among the main families

MAS[1]	No. of sequences[2]	Class	Highest identity[3]	e-value[4]
Zinc Domains				
MZCCL10172D03.g	8	Zinc Finger	Zinc finger PCP1-like	0
MZCCL10209H12.g	5	Zinc Finger	Zinc finger PCP1-like	2.E-75
MZCCL10107E04.g	3	PHD-finger	PHD finger protein-related	1.E-174
MZCCL15009C05.g	3	Zinc Finger	Zinc finger transcription factor-like protein	6.E-45
MZCCL10126H06.g	2	PHD-finger	PHD finger protein-related	1.E-111
MZCCS15001B10.g	1	RING finger	COP1	1.E-119
ZMZZEN7040A01.g	1	Zinc Finger	Trithorax 1-like protein	4.E-80
MZCCL10112F02.g	1	WRKY	WRKY3-like protein	1.E-69
ZMZZEN5056A01.g	1	YABBY	Yabby10 protein	1.E-56
ZMZZEN1038F05.g	1	Zinc Finger	Putative Zinc finger transcription factor	5.E-34
ZMZZEN1059B06.g	1	GATA	GATA-1 zinc finger protein	3.E-23
MZCCL10156H03.g	1	GATA	Zinc finger (GATA type) family protein	9.E-23
ZMZZEN2006H01.g	1	Dof	Prolamin-box binding factor	1.E-21
MZCCL10174G03.g	1	WRKY	WRKY7-like protein	3.E-15
MZCCL10127A03.g	1	Zinc Finger	ZFP2-like protein	5.E-14
bZIP family				
MZCCL15028H02.g	22	bZIP	Opaque-2	0
MZCCL10006F06.g	12	bZIP	Opaque-2	0
MZCCL10016E07.g	6	bZIP	Rice seed b-Zipper 4 (RISBZ4)-like	4.E-56
MZCCL20023E03.g	4	bZIP	Putative bZIP transcription factor	8.E-33
MZCCL10186G08.g	3	bZIP	bZIP family transcription factor	1.E-82
ZMZZEN1054G11.g	3	bZIP	Opaque-2	4.E-63
MZCCL10013F06.g	1	bZIP	TRAB1-like	7.E-36
MZCCL10125H06.g	1	bZIP	Putative bZIP transcription factor	2.E-40
MZCCL20021D04.g	1	bZIP	OSE2-like protein	4.E-25

MZCCL20017C12.g	1	bZIP	Putative bZIP transcription factor	2.E-27
MZCCL20028B04.g	1	bZIP	Putative bZIP transcription factor	1.E-130
MADS family				
MZCCL10095D11.g	1	MADS	ZAG2	1.E-109
MZCCL10057C06.g	1	MADS	ZAG2	1.E-105
MZCCL20034F06.g	1	MADS	MADS box protein 1	3.E-92
MZCCL10013G09.g	1	MADS	MADS-box transcription factor-like	8.E-46
Homeodomain family				
MZCCL10121F06.g	3	Homeobox	Hox7-like protein	2.00E-38
MZCCL10056F10.g	3	Homeo-Zip	Putative Hox4 protein	7.00E-30
MZCCL10216G12.g	1	Homeo-Zip	OCL3 protein	5.00E-97
ZMZZEN6061C10.g	1	Homeo-Zip	OCL5 protein	1.00E-56
MZCCL15029D11.g	1	Homeobox	Putative WUSCHEL homeobox protein 2	2.00E-26
ZMZZEN6070H06.g	1	Homeobox	Putative WUSCHEL homeobox protein 11	3.00E-23
MZCCL10079H04.g	1	Homeobox	Putative homeodomain protein	1.00E-14
Helix-loop-helix family				
MZCCS20044E10.g	11	bHLH	bHLH protein family	6.00E-16
ZMZZEN5008D08.g	2	bHLH	Putative bHLH transcription factor	7.00E-74
ZMZZEN7010B07.g	1	bHLH	Putative transcription factor PCF6	5.00E-68
MZCCL10202D08.g	1	bHLH	bHLH protein family	5.00E-46
MZCCL10075H11.g	1	bHLH	bHLH protein family	7.00E-21
MZCCS20019G07.g	1	bHLH	Putative bHLH transcription factor	1.00E-11
NAC family				
ZMZZEN3009C11.g	30	NAC	NAM-related protein 1-like	1.E-117
MZCCL10018G09.g	23	NAC	NAM-related protein 1-like	1.E-111
MZCCS15005C05.g	7	NAC	NAM-related protein 1-like	4.E-85
MZCCL20006E06.g	4	NAC	NAM-related protein 1-like	6.E-74
MZCCL10127E04.g	4	NAC	OsNAC3 protein-like	3.E-32
ZMZZEN6071D10.g	3	NAC	NAM-related protein 1-like	2.E-85
MZCCL10058G04.g	2	NAC	OsNAC2 protein-like	4.E-88
ZMZZEN5053D04.g	2	NAC	NAM-related protein 1-like	5.E-67
ZMZZEN7014B11.g	2	NAC	Putative NAM (no apical meristem) protein	2.E-29
MZCCL20010G12.g	1	NAC	OsNAC1 protein-like	6.E-61
ZMZZEN7015E08.g	1	NAC	NAC2 protein-like	3.E-19
MZCCL10055H06.g	1	NAC	NAM-like protein	1.E-13
APETALA2/EREBP family				
MZCCL20025G09.g	13	ERF	Putative transcription factor EREBP1	4.E-28
MZCCL10005A11.g	2	AP2	AP2 domain-containing transcription factor	8.E-24
MZCCL20042F02.g	1	AP2	AP2 domain-containing transcription factor	6.E-36
MZCCL10193B06.g	1	AP2	WRINKLED1-like protein	3.E-19
MZCCL20022H07.g	1	ERF	ERF1-like transcription factor	3.E-11
HMG (high-mobility group) family				
MZCCS15031G06.g	1	HMG	Putative SSRP1 protein	1.E-15
Myb family				
MZCCS20011C07.g	4	MYB	Putative typical P-type R2R3 Myb protein	4.E-39
MZCCL10097G05.g1	1	MYB	Putative typical P-type R2R3 Myb protein	1.E-75
MZCCL10142D02.g	1	MYB	Putative Myb-family transcription factor	2.E-64
MZCCL15026F02.g	1	MYB	Putative Myb-family transcription factor	3.E-28
MZCCL10068G01.g	1	MYB	Circadian clock associated protein LHY-like	9.E-28
ZMZZEN6076B05.g	1	MYB	Putative c-Myb-like transcription factor	4.E-27

Heat shock factor family				
MZCCL10023G08.g	3	HSF	Heat shock factor RHSF2-like	8.E-66
MZCCL10084F08.g	3	HSF	Heat shock factor RHSF4-like	1.E-51
MZCCL10160C11.g	1	HSF	Heat shock factor RHSF6-like	7.E-58
MZCCL20049C08.g	1	HSF	Heat shock factor RHSF4-like	7.E-17
GRAS family				
ZMZZEN6040H08.g	1	GRAS	Scarecrow transcriptional regulator-like protein	6.E-16
Other				
MZCCL10045B04.g	10	-	Fertilization-independent endosperm protein 1	1.E-162
MZCCL10186E06.g	6	-	Heat shock protein HSP90	1.E-136
MZCCL10038G04.g	6	-	Putative VIP1 transcription factor	5.E-27
MZCCL10026E03.g	5	-	Heat shock protein HSP82	1.E-139
MZCCS20012A11.g	4	-	Putative VIP2 transcription factor	1.E-161
ZMZZEN5005A11.g	4	-	Putative transcription factor X1	2.E-54
MZCCL10039B11.g	3	-	TATA box binding protein-associated factor	1.E-138
MZCCL15012H09.g	3	-	Transcription initiation factor IIB	2.E-88
MZCCM15002E07.g	3	-	Putative co-repressor protein	9.E-49
ZMZZEN5055B11.g	3	-	Transcription initiation factor IIA small subunit	8.E-46
ZMZZEN5026G12.g	3	-	Putative CCAAT-binding transcription factor	2.E-33
MZCCL20014D07.g	2	-	Putative auxin response factor 10 (ARF10)-like	1.E-139
MZCCL20034D01.g	2	-	Heat shock protein HSP82	1.E-120
ZMZZEN6065G08.g	2	-	Transcription initiation factor IIH	1.E-71
MZCCL20004F04.g	2	-	Squamosa-promoter binding-like protein	3.E-60
MZCCL10075G04.g	2	-	TATA box binding protein-associated factor	6.E-50
MZCCS10002E02.g	2	-	TATA box binding protein-associated factor	3.E-48
MZCCL10084A05.g	2	-	Hd1-like protein	5.E-41
MZCCL10200C10.g	2	-	Transcriptional regulator FUSCA3	7.E-33
ZMZZEN6050A03.g	2	-	Putative auxin-regulated IAA22	2.E-32
MZCCL10161E11.g	2	-	Enhancer of zeste-like protein 2	6.E-27
ZMZZEN6049A09.g	2	-	Putative auxin-regulated IAA8	3.E-25
ZMZZEN6039F09.g	2	-	Putative CCAAT-binding transcription factor	2.E-21
ZMZZEN6002G01.g	1	-	Response regulator 10	1.E-132
MZCCS20026C03.g	1	-	Putative auxin response factor 7 (ARF7)-like	1.E-131
MZCCS15007G06.g	1	-	Putative auxin response factor 7 (ARF7)-like	1.E-121
MZCCS15013B03.g	1	-	Putative auxin response factor 10 (ARF10)-like	1.E-111
ZMZZEN5030B10.g	1	-	Putative co-repressor protein	1.E-102
ZMZZEN6061H10.g	1	-	Putative auxin response factor 1 (ARF1)-like	3.E-80
MZCCL20022G06.g	1	-	Transcription initiation factor IIB	2.E-74
MZCCL15035F11.g	1	-	Putative transcription regulatory protein SNF2	3.E-74
MZCCL15009F05.g	1	-	Putative transcriptional regulatory protein	4.E-62
MZCCL10077B06.g	1	-	Putative ethylene-insensitive protein EIL3	2.E-44
ZMZZEN6097G06.g	1	-	Putative transcriptional corepressor LEUNIG	2.E-41
MZCCL10178A09.g	1	-	Putative HAP3 transcriptional-activator	1.E-40
ZMZZEN1009H11.g	1	-	Putative transcription factor IIIB 70 KD subunit	1.E-39
ZMZZEN5062G11.g	1	-	VIP1/ABI3-like transcription factor	3.E-38
MZCCL10212F09.g	1	-	Putative transcriptional regulatory protein	3.E-23
MZCCL10210H11.g	1	-	ABA response element binding factor	2.E-18
ZMZZEN7004F11.g	1	-	Putative trancription factor	3.E-17

[1]MAS, Maize Assembled Sequences, are the sets of contigs and singletons (Verza et al 2005).
[2]Number of endosperm-preferred sequences in contigs
[3]Best GenBank hit
[4]E-value correspondent to the best hit of the blastX of the MAS consensus sequence against the GenBank

Table 3

Table 3. Relative expression of endosperm-preferred transcription factors during endosperm development

Highest identity	MAS	Relative expression[1]		
		10 DAP	15 DAP	20 DAP
NAM-related protein 1-like	MZCCL10018G09.g	0.6	1.6	17.4
Fertilization-independent endosperm protein 1	MZCCL10045B04.g	3.1	1.6	0
Zinc finger PCP1-like	MZCCL10172D03.g	0.6	3.1	1.2
NAM-related protein 1-like	MZCCL20006E06.g	0	1.6	3.7
NAM-related protein 1-like	MZCCS15005C05.g	0	6.3	0
Transcription initiation factor IIB	MZCCL15012H09.g	0	4.7	0
Putative bZIP transcription factor	MZCCL20023E03.g	0	0	3.7
Zinc finger transcription factor-like protein	MZCCL15009C05.g	0	3.1	0
Putative transcription factor EREBP1	MZCCL20025G09.g	0	0	2.5
Heat shock protein HSP82	MZCCL20034D01.g	0	0	2.5
Opaque-2	MZCCL10006F06.g	4.3	1.6	3.7
Putative typical P-type R2R3 Myb protein	MZCCS20011C07.g	0	0	2.5
Putative VIP2 transcription factor	MZCCS20012A11.g	0	0	2.5
Putative Hox4 protein	MZCCL10056F10.g	1.9	0	0
Heat shock factor RHSF4-like	MZCCL10084F08.g	1.9	0	0
Putative VIP1 transcription factor	MZCCL10038G04.g	1.2	1.6	0
Heat shock protein HSP82	MZCCL10026E03.g	1.2	0	2.5
AP2 domain-containing transcription factor	MZCCL10005A11.g	1.2	0	0
OsNAC2 protein-like	MZCCL10058G04.g	1.2	0	0
Hd1-like protein	MZCCL10084A05.g	1.2	0	0
OsNAC3 protein-like	MZCCL10127E04.g	1.2	0	0
Enhancer of zeste-like protein 2	MZCCL10161E11.g	1.2	0	0
TATA box binding protein-associated factor	MZCCS10002E02.g	1.2	0	0
PHD finger protein-related	MZCCL10107E04.g	0.6	1.6	1.2
Rice seed b-Zipper 4 (RISBZ4)-like	MZCCL10016E07.g	0.6	0	1.2
TATA box binding protein-associated factor	MZCCL10075G04.g	0.6	0	1.2
PHD finger protein-related	MZCCL10126H06.g	0.6	0	1.2
bZIP family transcription factor	MZCCL10186G08.g	0.6	0	1.2
Zinc finger PCP1-like	MZCCL10209H12.g	0.6	0	1.2
Opaque-2	MZCCL15028H02.g	0	1.6	1.2
Putative auxin response factor 10 (ARF10)-like	MZCCL20014D07.g	0	1.6	1.2
Putative transcriptional regulatory protein	MZCCL15009F05.g	0	1.6	0
Putative co-repressor protein	MZCCM15002E07.g	0	1.6	0
Putative Myb-family transcription factor	MZCCL15026F02.g	0	1.6	0
Putative WUSCHEL homeobox protein 2	MZCCL15029D11.g	0	1.6	0
Putative transcription regulatory protein SNF2	MZCCL15035F11.g	0	1.6	0
COP1	MZCCS15001B10.g	0	1.6	0
Putative auxin response factor 7 (ARF7)-like	MZCCS15007G06.g	0	1.6	0
Putative auxin response factor 10 (ARF10)-like	MZCCS15013B03.g	0	1.6	0
Putative SSRP1 protein	MZCCS15031G06.g	0	1.6	0
bHLH protein family	MZCCS20044E10.g	0	0	1.2
Squamosa-promoter binding-like protein	MZCCL20004F04.g	0	0	1.2
OsNAC1 protein-like	MZCCL20010G12.g	0	0	1.2

Putative bZIP transcription factor	MZCCL20017C12.g	0	0	1.2
OSE2-like protein	MZCCL20021D04.g	0	0	1.2
Transcription initiation factor IIB	MZCCL20022G06.g	0	0	1.2
ERF1-like transcription factor	MZCCL20022H07.g	0	0	1.2
Putative bZIP transcription factor	MZCCL20028B04.g	0	0	1.2
MADS box protein 1	MZCCL20034F06.g	0	0	1.2
AP2 domain-containing transcription factor	MZCCL20042F02.g	0	0	1.2
Heat shock factor RHSF4-like	MZCCL20049C08.g	0	0	1.2
Putative bHLH transcription factor	MZCCS20019G07.g	0	0	1.2
Putative auxin response factor 7 (ARF7)-like	MZCCS20026C03.g	0	0	1.2
Hox7-like protein	MZCCL10121F06.g	0.6	0	0
TRAB1-like	MZCCL10013F06.g	0.6	0	0
MADS-box transcription factor-like	MZCCL10013G09.g	0.6	0	0
Heat shock factor RHSF2-like	MZCCL10023G08.g	0.6	0	0
TATA box binding protein-associated factor	MZCCL10039B11.g	0.6	0	0
NAM-like protein	MZCCL10055H06.g	0.6	0	0
ZAG2	MZCCL10057C06.g	0.6	0	0
Circadian clock associated protein LHY-like	MZCCL10068G01.g	0.6	0	0
bHLH protein family	MZCCL10075H11.g	0.6	0	0
Putative ethylene-insensitive protein EIL3	MZCCL10077B06.g	0.6	0	0
Putative homeodomain protein	MZCCL10079H04.g	0.6	0	0
ZAG2	MZCCL10095D11.g	0.6	0	0
Putative typical P-type R2R3 Myb protein	MZCCL10097G05.g1	0.6	0	0
WRKY3-like protein	MZCCL10112F02.g	0.6	0	0
Putative bZIP transcription factor	MZCCL10125H06.g	0.6	0	0
ZFP2-like protein	MZCCL10127A03.g	0.6	0	0
Putative Myb-family transcription factor	MZCCL10142D02.g	0.6	0	0
Zinc finger (GATA type) family protein	MZCCL10156H03.g	0.6	0	0
Heat shock factor RHSF6-like	MZCCL10160C11.g	0.6	0	0
WRKY7-like protein	MZCCL10174G03.g	0.6	0	0
Putative HAP3 transcriptional-activator	MZCCL10178A09.g	0.6	0	0
Heat shock protein HSP90	MZCCL10186E06.g	0.6	0	0
WRINKLED1-like protein	MZCCL10193B06.g	0.6	0	0
Transcriptional regulator FUSCA3	MZCCL10200C10.g	0.6	0	0
bHLH protein family	MZCCL10202D08.g	0.6	0	0
ABA response element binding factor	MZCCL10210H11.g	0.6	0	0
Putative transcriptional regulatory protein	MZCCL10212F09.g	0.6	0	0
OCL3 protein	MZCCL10216G12.g	0.6	0	0

[1]The relative abundance of ESTs for each MAS was calculated as the number of ESTs present in a given library pool (10 DAP pool, 15 DAP pool and 20 DAP pool) divided by the total number of ESTs in that pool. The values were multiplied by 10^4 to aid better interpretation.

Figures

Figure 1.

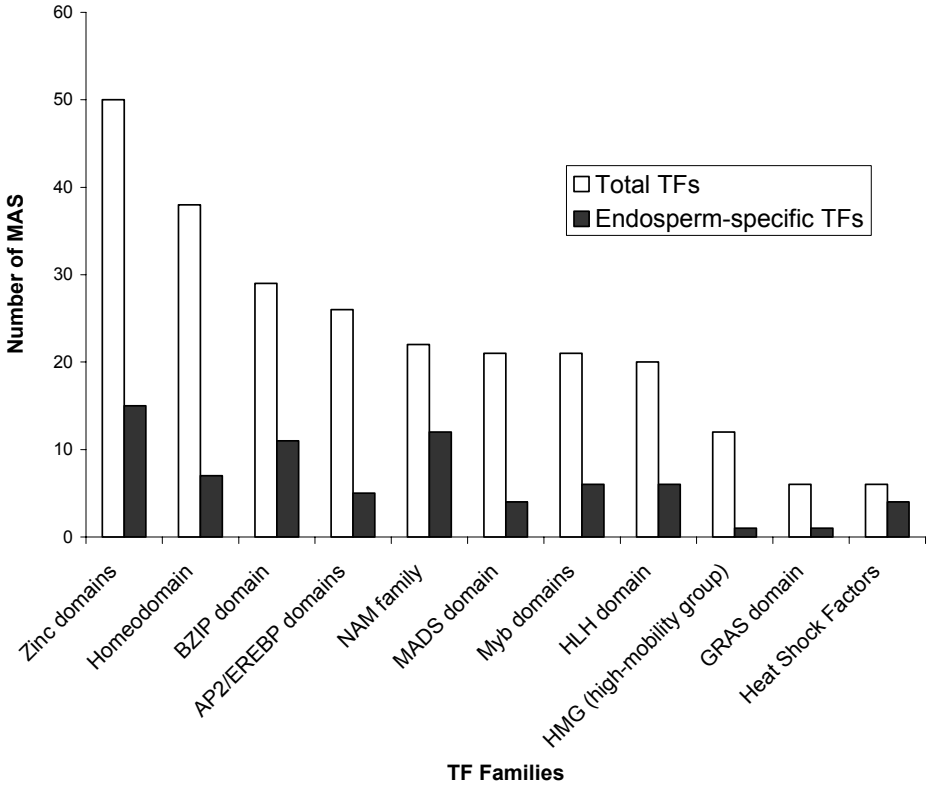


Figure 1. Distribution of the transcription factors (TFs) among the main families; 161 out of 414 TF MASs and 41 out of 113 endosperm-preferred TF MASs were unable to classify.

Figure 2.

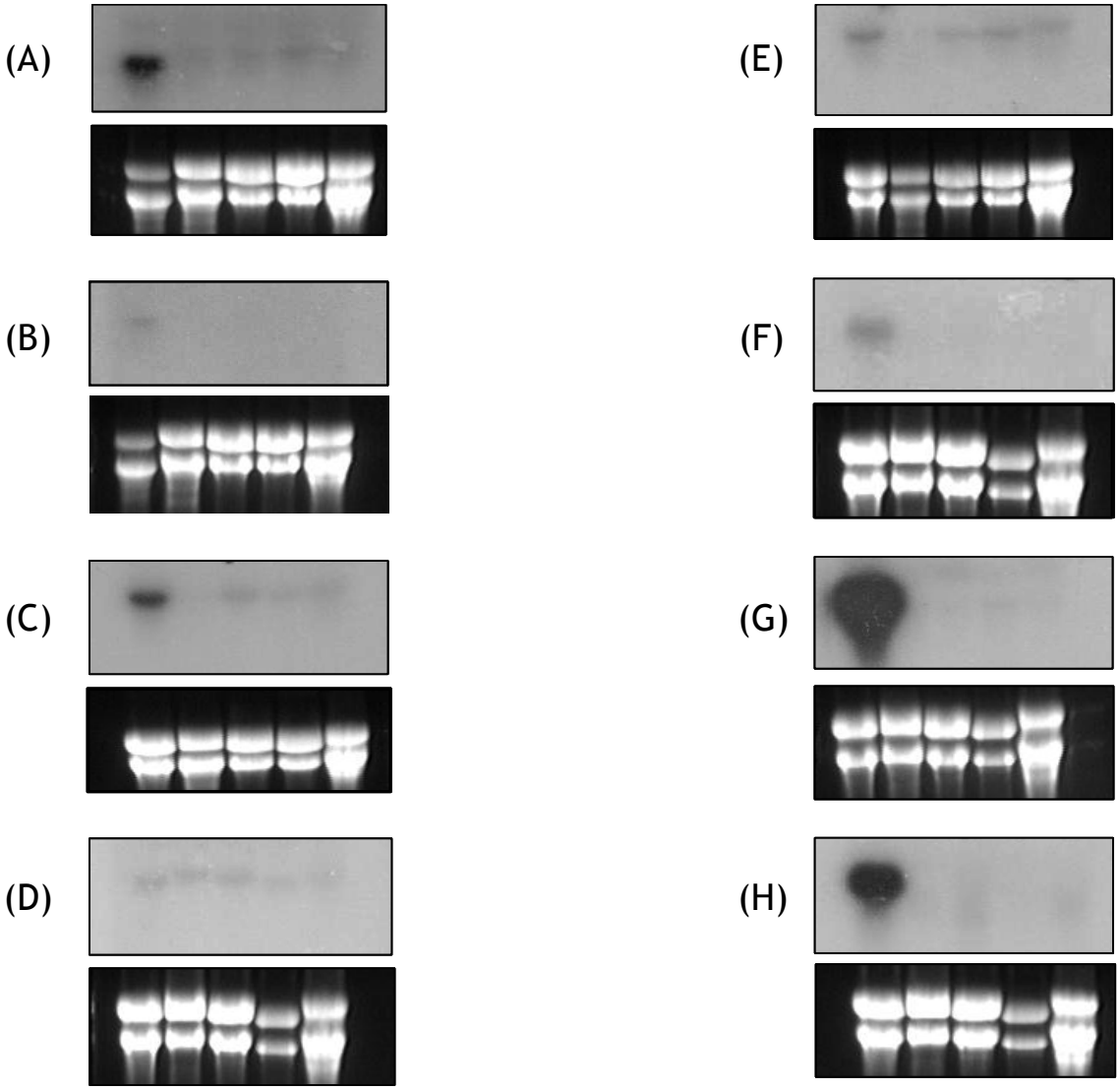


Figure 2. RT-PCR analysis of the expression profiles of endosperm-preferred genes selected by *in silico* approaches. (A) Opaco-2 (MAS MZCCL10006F06.g); (B) NAM-family like TF (MAS MZCCL10018G09.g); (C) EREBP-family like TF (MAS MZCCL20025G09.g); (D) PHD finger protein-related / SET domain-containing protein (MAS MZCCL10107E04.g); (E) Zinc finger family PCP-1 like protein (MAS MZCCL10172D03.g); (F) Alpha-tubulin gene. En: endosperm; L: leaf; R: root; Co: coleoptile; Em: embryo.

Transcriptome analysis of maize endosperm identifies an aleurone-specific transcription factor of the NAC family

Natalia Cristina Verza, Thaís Rezende e Silva, Sylvia Morais de Sousa, Paulo Henrique Fisch, Marcelo Martins Rebello and Paulo Arruda

TRANSCRIPTOME ANALYSIS OF MAIZE ENDOSPERM IDENTIFIES AN ALEURONE-SPECIFIC TRANSCRIPTION FACTOR OF THE NAC FAMILY

Natalia Cristina Verza[1], Thaís Rezende e Silva[1], Sylvia Morais de Sousa[1], Paulo Henrique Fisch[1], Marcelo Martins Rebello[1] and Paulo Arruda*[1],[2]

1 Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas (UNICAMP), 13.083-970, Campinas, SP, Brazil.

2 Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas (UNICAMP), 13.083-970, Campinas, SP, Brazil.

* Corresponding author

Footnotes

Financial source: FAPESP, CAPES and CNPq

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Paulo Arruda (parruda@unicamp.br)

Abstract

The NAC (NAM/ATAF1/2/CUC) domain protein family is widely distributed in plants, and some of their members are involved in biotic and abiotic stress responses. Screening a large maize ESTs database enriched in endosperm sequences, we have identified 12 members of the NAC-family that are preferentially expressed in developing endosperm. One of these, called EPN-1, was found to be preferentially expressed in the aleurone layer. EPN-1 expression can be detected early at 5 days after pollination and peaks at 20-25 DAP. The analysis of the promoter sequence of EPN-1 revealed the presence of CIS-elements related to endosperm-specificity and ABA and GA signaling. We discuss here the possible role of EPN-1 in maize seeds in late embryogenesis and seed maturation processes.

Keywords: NAC domain protein; *Zea mays*; Endosperm; Aleurone; Transcription factor; CIS-acting elements; Seed maturation.

Introduction

The maize seed is composed by two main parts: the embryo and the endosperm. It has been shown that the success of their development depends on the interaction between these two seed components, and the presence of an intact endosperm is required for the proper embryo development (Consonni et al., 2005). The maize endosperm is a highly specialized nutritive tissue consumed by the embryo as energy supply during embryogenesis. The developed endosperm consists of the starchy endosperm, a central mass of cells that accumulate starch and storage proteins, a basal layer of transfer cells (BETL) that mediates the entering of maternal nutrients into the seeds, and a one-cell-thick layer of aleurone cells that surround the starchy endosperm. During germination, over gibberellins stimulus coming from the germinating embryo, the aleurone produces a range of hydrolytic enzymes that digest cell walls, storage proteins and starch present in the endosperm.

Maize endosperm is a triploid tissue formed by the fusion of two polar nuclei and one sperm nucleus. Until the fourth day after pollination (DAP), the endosperm nuclei divide synchronously without cell wall formation. Then the tissue changes from a multinucleate single cell to a uninucleate multicellular morphology. Most of the

endosperm cells are produced up to 12 DAPs, when it begins to accumulate large amounts of starch and storage proteins. By 16 DAPs the maturation program has initiated, preparing the seeds for desiccation and dormancy, and by 23 DAP desiccation has begun. At around 25-30 DAP, the relative water content of the endosperm initiates to decrease, and the seed desiccation, controlled by hormone signaling, maintain the germinative development arrested (reviewed in: Lopes and Larkins, 1993; Olsen, 2001 and Olsen, 2004).

The hormone abscisic acid (ABA) plays a central role in suppressing precocious germination in developing maize seeds and regulates the expression of diverse genes during the seed maturation process. In developing seeds, ABA is synthesized by embryo tissues, and is also transferred from maternal tissues to the seed during water stress (Ober and Setter, 1992). Maize kernels deficient in ABA synthesis are viviparous, germinating on the ear midway through kernel development (Robertson, 1955; Neill et al., 1986.) While the ABA levels are important to prevent precocious seed germination, GAs play a crucial role in promoting germination of many types of mature seeds. In wheat and barley, GAs promote the expression of hydrolytic enzyme genes, leading to the mobilization of endosperm reserves for the embryo nutrition (for review, see Jacobsen et al., 1995). In maize, GAs and ABA play antagonistic roles in controlling vivipary, and GA1 and GA3 levels in maize embryos decline prior to the peak of ABA concentration (White et al., 2000).

Transcription factors are largely responsible for the selectivity in gene regulation, and are often expressed in a tissue-specific, developmental-stage-specific, or stimulus-dependent manner (Zhang, 2003). The regulatory mechanisms underlying mid and late-embryogenesis events remain largely unknown. There are a few known regulators of seed maturation identified in studies from maize and *Arabidopsis*. The maize Vp1 (Viviparous-1) gene is required for ABA induction of maturation-specific genes, contributing to desiccation tolerance acquisition and arrest in embryo growth. VP1 also inhibits the expression of germination-specific alpha-amylase genes in aleurone cells and seems to be involved in preventing precocious hydrolysis of storage compounds accumulated in the endosperm (Hoecker et al., 1995). The VP1 gene is weakly expressed in the starchy endosperm, while highly expressed in the aleurone and embryo during the maturation phase. The VP1 protein is involved in the regulation of a number of diverse

genes like the rice bZIP TRAB1, that interacts with both VP1 and CIS-ABA-responsive elements (ABREs) and mediates ABA signals (Hobo et al., 1999). Other key factors participate in maturation programmes in the cereal endosperm, like GAMYB, BPBF and SAD (Gubler et al., 1995; Isabel-Lamoneda et al., 2003), mainly regulating the post-germination phase.

NAC proteins (from (petunia NAM and *Arabidopsis* ATAF1,2 and CUC2; Souer et al, 1996; Aida et al, 1997) constitute one of the largest families of plant-specific transcription factors. NAC family members are involved in developmental processes, including formation of the shoot apical meristem, floral organs and lateral shoots, as well as in stress responses and plant defense (Olsen et al., 2005). Several NAC genes are found to be induced by hormones like the abscisic acid-responsive NAC gene (ANAC) from *Arabidopsis thaliana* (Greve et al., 2003) and the HSINAC (from HvSPY-interacting NAC protein), that has been shown to be a negative regulator of GA response in barley aleurone, inhibiting GA3 up-regulation of alpha-amylase expression (Robertson, 2004).

In the present study, we have cloned and characterized a NAC-family transcription factor, EPN-1 (Endosperm-Preferred NAM-1), which has an endosperm-preferred pattern of expression and is preferentially expressed in maize aleurone. We have isolated the EPN-1 promoter and used it to drive the b-glucuronidase gene in transient expression assays. We found that this novel NAC-family gene may be involved in the regulation of maturation and germination pathways during maize seed development.

Results and Discussion

Identification of NAC-family transcription factors expressed in maize endosperm

We have created a large database enriched in genes expressed in developing maize endosperm (Verza et al., 2005). Screening the database for transcription factors sequences, we identified over 1,200 TFs; 414 of which expressed in the endosperm. From the set of TFs expressed in maize endosperm, 113 of were found to be preferentially expressed in maize endosperm (Verza et al., in preparation), and may play important roles in the regulation of endosperm development.

Interestingly, the plant-specific NAC family was one of the most represented TF families within the 113 TFs set, with 12 TFs preferentially expressed in the maize

endosperm (Table 1). These 12 TFs correspond to 10 non-redundant NAC-family members preferentially expressed in maize endosperm (see Material and Methods). Since NAC is a multigenic family of TFs that are found to play roles in a diverse set of developmental processes, including developmental programmes, defense and abiotic stress responses, and the majority of the endosperm-preferred NAC members had a late expression pattern (Verza et al., in prep), we believe that these TFs may be involved in the regulation of late endosperm developmental processes, such as the response to the hydric stress accompanying seed desiccation. We decided to characterize one of these endosperm preferred NAC-family TFs, that we named Endosperm-Preferred NAM-1 (EPN-1).

The EPN-1 gene, its sequence features and genomic organization

The complete sequence of the EPN-1 cDNA was already available in the MAIZEST database (www.maizest.unicamp.br), as the consensus sequence of the ESTs cluster MZCCL10018G09.g (Verza et al., 2005). We used this consensus sequence to design primers to clone and re-sequence the complete EPN-1 cDNA. The coding sequence is 1,074 bp long and is 73% identical to the petunia NAM gene (Souer et al., 1996; accession X92205) and 84% identical to the *Zea mays* NAM-related protein 1 (NRP1), an endosperm-specific NAC-family member (Guo et al., 2003; accession AY325313) .

The EPN-1 sequence encodes a 357 amino acid protein. The alignment of the EPN-1 predicted protein sequence with related NAC-family sequences is shown in Figure 1. The NAC domain, represented by the 5 underlined sub-domains in the alignment, is located at the N-terminal portion, and is strictly conserved among all the sequences. The C-terminal region of EPN-1 is highly specific, sharing 65% of identity with the related maize NRP-1.

The comparison of the EPN-1 cDNA sequence with the genomic maize sequences available through the TIGR database revealed that the gene is composed by four exons and three introns, a structure distinct from that of the petunia NAM gene (Figure 2), but very similar to that of the maize NRP-1. The translation start codon is located within the second exon (Figure 2).

The EPN-1 gene is preferentially expressed in the maize endosperm

To access the expression pattern of the EPN-1 gene and confirm the *in silico* findings, we carried out an RT-PCR analysis using cDNAs from maize endosperm, root, leaf, coleoptile and embryo tissues. The EPN-1 transcripts can be found preferentially in the endosperm sample (Figure 3). The transcripts can be found in the endosperm at 5 days after pollination (DAP), rising to a peak around 25 DAP (Figure 3). This expression pattern suggests a late role of this gene during the endosperm development, although the few transcripts found at early stages may perform a regulatory function since the beginning of the developmental process.

The EPN-1 promoter has conserved endosperm-specificity, ABA and GA-binding CIS-elements

The promoter sequence of EPN-1 was retrieved using the coding sequence to perform a BLASTN (Altschul et al., 1997) analysis against the TIGR maize genomic sequences. The resulting sequence was used to design primers to clone and sequence a 1,900 pb DNA fragment upstream the translation start. As CIS-regulatory elements are major controllers of gene expression located within the 5' upstream sequence (Haberer et al., 2004), we used the PlantCare (Lescot et al., 2002) and the Place (Higo et al., 1999) tools to identify possible conserved motifs for gene expression and regulation (Figure 4). Two TATA boxes were identified, one located within the first intron, at position -113 from the initial ATG, and the second located at position -352. The EPN-1 promoter revealed endosperm-specificity related elements like the Prolamin-box, conserved in many cereal seed storage protein genes, the GCN-4-motif, that plays a central role in controlling endosperm-specific expression, and the RY/Sph motif, that is involved in high-level expression of several seed-specific genes, as well as functioning as a negative element repressing expression in non-seed tissues. The RY/Sph motif is also involved in response to Abscisic Acid (ABA) signaling through the maize VP1 binding. VP1 is specifically required for properly regulation of the maturation program in maize seed development, and the Sph element is an enriched sequence motif in promoters of genes co-activated by ABA and VP1.

Interestingly, a number of hormone-related elements were also found in the EPN-1 promoter, like eight ACGT-containing ABA response elements (ABREs), three amylase box

(also called Amy Box and Box I), conserved sequences found in 5'-upstream region of alpha-amylase genes that are related to a gibberellin (GA)-induced expression, one Pyrimidine box, an accessory motif for the transcriptional response to GA found in the promoter of barley alpha-amylase genes (Amy2/32b) (Mena et al., 2002) and two TATC-boxes, that have been shown to be related to GA responsiveness, being part, together with the Pyrimidine box, of a gibberellin response complex that give a high level of GA-regulated expression. These findings suggest that EPN-1 may have a regulatory role during late embryogenesis, possibly in the transition from seed maturation to germination processes, and might be regulated by ABA and GA.

Transient assays show that EPN-1 promoter drives aleurone-specific expression

To assay the pattern of expression driven by the EPN-1 promoter, we conducted a transient expression analysis in which a 1,9kb fragment of the EPN-1 promoter region was cloned into the promoter-less pRT103GUS vector driving the β-glucuronidase gene (pEPN-GUS; Figure 5a). The plasmid pRT103GUS containing the bacterial GUS gene driven by the constitutive CaMV35S promoter was used as control. Immature maize seeds were sectioned transversally, and the caps were prepared by peeling back the entire pericarp and removing the aleurone layer from half of the cap area, remaining a portion of the intact aleurone layer. The caps, as well as the longitudinally sectioned seeds, were bombarded with DNA-coated microprojectiles and the GUS activity was evaluated by counting the blue spots (Figure 5d). Figure 5c shows that the EPN-1 promoter directed the expression preferentially in the aleurone layer of the seed, in contrast to the wide spread pattern given by the CaMV35S:GUS (Figure 5b). The EPN-1 gene, thus, is the first NAC-family transcription factor shown to be preferentially expressed in the aleurone layer. This pattern of expression reinforces its possible role in regulating the maturation to germination transition process in maize seeds.

Conclusion

In the present study, we have cloned and characterized an endosperm-specific member of the NAC-family of transcription factors, EPN-1. The results show that EPN-1 is expressed preferentially in the aleurone layer of the maize endosperm, and its promoter

has conserved CIS-elements related to ABA- and GA-regulated transcription, likewise conserved sequences found in alpha-amylase promoters and the Sph element, bound by the VP1 transcription factor. Since VP1 is known to be expressed only in seed tissues, and it regulates maturation and dormancy in plant seeds by activating genes responsive to the stress hormone abscisic acid (ABA), it may be possible that EPN-1 expression could be regulated by VIP1, being part of the ABA- and GA-signaling pathways in maize seeds.

Although several NAC-family transcription factors have been reported as candidates for stress and hormone responses, none of them have been demonstrated to be preferentially expressed in the aleurone layer. Further investigation needs to be done to unravel the regulatory effects of ABA and GA in the EPN-1 expression, and to clarify if there is any interaction between VP1 and EPN-1 during seed development.

Material and Methods

Screening of Databases and Sequence Alignments

The identification of the maize endosperm-preferred NAC-family transcription factors was conducted as described in Verza et al. (in preparation).

To identify gene redundancy among sequences retrieved from the MAIZEST database (Verza et al., 2005), the consensus sequences (MASs) of all endosperm-preferred NAC TFs were aligned using the CLUSTALW program (Thompson et al., 1994). Those MASs whose DNA sequences showed more than 95% of identity were considered redundant.

The predicted protein sequences were aligned using ClustalX and colored using the BoxShade 3.21 tool (http://www.ch.embnet.org/software/BOX_form.html).

Plant Growth

Maize (*Zea mays* L.) plants from the Oh43 inbred line were grown in the greenhouse. Ears were self pollinated and harvested at 10 days after pollination (DAP) and 30 DAP for the hormone assay and at 25 DAP for the transient expression assay. Embryos were dissected manually. Roots, leaves and coleoptiles were harvested from 5-day-old seedlings germinated under controlled conditions. The 25 DAP seeds for the

transient assay were immediately used, and the other tissues were frozen in liquid nitrogen and stored at -80 ºC.

RNA Extraction and RT-PCR analysis

Endosperm and embryo total RNA free from genomic DNA were extracted according to Manning (1991), and roots, leaves and coleoptiles total RNA were extracted using the Trizol reagent (Invitrogen, USA) as described by the manufacturer. The RT-PCR reactions were performed with 500ng of total RNA using the one-step AccessQuick™ RT-PCR System (Promega). The products were separated by electrophoresis in agarose gel and visualized by UV excitation of ethidium bromide-stained DNA. The primers used to amplify the EPN-1 gene were ZmESN1fwd (5'-CATGGCGGCGGACC-3') and ZmESN1rev (5'-GATGGCGTGTGGAAGTACTGA-3').

EPN-1 promoter isolation, vector construction and transient expression assays in immature maize endosperm

The 1,9 kb fragment of the EPN-1 promoter was amplified from maize Oh43 genomic DNA using the primers ZmNAMprofwd (5'- CCAGTCAACATAGCCCAACT-3') and ZmNAMprorev (5'- GAGGTCAGTCCTCGAGTCAGAGA -3'). The single PCR product was isolated from agarose gel using the Concert™ rapid gel extraction system (Invitrogen, USA) and then subcloned into the pGEM-T EASY vector (Promega, USA). The vector was subsequently digested at the HincII and the XhoI restriction sites included at the 3′- and 5′-ends and ligated into the promoter-less pRT103GUS vector digested with the same restriction enzymes.

Ears were harvested at 20 days after pollination, surface sterilized for 15 min with 5% commercial bleach, and rinsed four times in distilled water. Seeds were dissected from the cob and sectioned longitudinally and transversally. The caps were dissected by removing the the entire pericarp and half of the aleurone layer area. 9 transversal sections and 6 longitudinal sections were flattened on 100-mm-diameter Petri dishes containing 20 mL of MS medium (Murashige and Skoog, 1962), with the sliced surface facing upward. Five micrograms of column-purified DNA was used to coat 3 mg of 1- to 3-pm-diameter gold particles as reported by Yunes et al. (1998). The endosperms were bombarded twice with 0.5 pg DNA using a high-pressure helium-driven particle

acceleration device (Sanford et al., 1991). After bombardment, the samples were incubated for 24 hr in the dark at room temperature. The endosperms were then stained for GUS activity according to the method of Jefferson (1987). To minimize experimental errors, all constructs were analyzed using seeds of the same ear.

Acknowledgments

NCV was supported by a postgraduate fellowship from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and TRS was supported by a postgraduate fellowship from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

Literature Cited

Aida M, Ishida T, Fukaki H, Fujisawa H, Tasaka M (1997) Genes involved in organ separation in *Arabidopsis*: An analysis of the cup-shaped cotyledon mutant. Plant Cell **9**:841-857

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**(17):3389-402

Consonni G, Gavazzi G, Dolfini S (2005) Genetic analysis as a tool to investigate the molecular mechanisms underlying seed development in maize. Ann Bot (Lond) **96**:353-62

Greve K, La Cour T, Jensen MK, Poulsen FM, Skriver K (2003) Interactions between plant RING-H2 and plant-specific NAC (NAM/ATAF1/1/CUC2) proteins: RING-H2 molecular specificity and cellular localization. Biochem J **371**: 97–108

Gubler F, Kalla R, Roberts JK, Jacobsen JV (1995) Gibberellin-regulated expression of a myb gene in barley aleurone cells: evidence for Myb transactivation of a high-pI alpha-amylase gene promoter. Plant Cell **7**: 1879-1891

Guo M, Rupe MA, Danilevskaya ON, Yang X, Hu Z (2003) Genome-wide mRNA profiling reveals heterochronic allelic variation and a new imprinted gene in hybrid maize endosperm. Plant Journal **36**(1):30-44

Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res **27:** 297–300

Hoecker U, Vasil IK, McCarty DR (1995) Integrated control of seed maturation and germination programs by activator and repressor functions of Viviparous-1 of maize. Genes Dev **9**: 2459-2469

Hobo T, Kowyama Y, Hattori T (1999) A bZIP factor, TRAB1, interacts with VP1 and mediates abscisic acid-induced transcription. Proc Natl Acad Sci USA **96**:15348-15353

Isabel-LaMoneda I, Diaz I, Martinez M, Mena M, Carbonero P (2003) Plant J. **33**: 329–340

Jacobsen JV, Gubler F, Chandler PM (1995) Gibberellin action in germinated cereal grains. In Plant Hormones: Physiology, Biochemistry and Molecular Biology, P.J. Davies, ed (Dordrecht, The Netherlands: Kluwer Academic Publishers), pp. 246–271.

Jefferson RA (1987) Assaying chimeric genes in plants: the GUS gene fusion system. Plant Mol Biol Reporter **5**, 387–405

Lescot M, Déhais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, Rouzé P, Rombauts S (2002) PlantCARE, a database of plant *cis*-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. Nucleic Acids Res **30:** 325–327

Lopes MA, Larkins BA (1993) Endosperm origin, development, and function. Plant Cell **5**: 1383-1399

Manning K (1991) Isolation of nucleic acids from plants by differential solvent precipitation. Anal Biochem **195**:45-50.

Mena M, Cejudo FJ, Isabel-Lamoneda I, Carbonero P (2002) A Role for the DOF Transcription Factor BPBF in the Regulation of Gibberellin-Responsive Genes in Barley Aleurone. Plant Physiol. **130**: 111-119

Murashige T and Skoog F (1962) A revised medium for rapid growth and bio-assays with tobacco tissue cultures. Physiologia Plantarum **15**:473-497

Neill SJ, Horgan R, Parry AD (1986) The carotenoid and abscisic-acid content of viviparous kernels and seedlings of *Zea mays*-L. Planta **169**: 87-96

Ober ES, Setter TL (1992) Water deficit induces abscisic acid accumulation in endosperm of maize viviparous mutants. *Plant Physiology* **98**: 353–356

Olsen OA (2001) Endosperm development: Cellularization and cell fate specification. Annu Rev Plant Physiol Plant Mol Biol. 2001, **52**: 233-267

Olsen OA (2004) Nuclear endosperm development in cereals and *Arabidopsis thaliana*. Plant Cell **16**: S214-S227

Olsen AN, Ernst HA, Leggio LL, Skriver K (2005) NAC transcription factors: structurally distinct, functionally diverse. Trends Plant Sci. **10**:79-87

Robertson D (1955) The genetics of vivipary in maize. Genetics **40:** 745-760

Robertson M (2004) Two transcription factors are negative regulators of gibberellin response in the HvSPY-signaling pathway in barley aleurone. Plant Physiol **136**(1):2747-61

Sandford JC, Devit MJ, Russell JA, Smith FD, Harpending PR, Roy MK, Johnston SA (1991) An improved, helium-driven biolistic device. Technique-A Journal of Methods in Cell and Molecular Biology **3**: 3-16

Souer E, van Houwelingen A, Kloos D, Mol J, Koes R (1996) The No Apical Meristem gene of petunia is required for pattern formation in embryos and flowers and is expressed as meristem and primordia boundaries. Cell **85**:159-170

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22:** 4673–4680

Verza NC, Silva TR, Neto GC, Nogueira FT, Fisch PH, de Rosa Jr VE, Rebello MM, Vettore AL, da Silva FR, Arruda P (2005) Endosperm-preferred expression of maize genes as revealed by transcriptome-wide analysis of expressed sequence tags. Plant Mol Biol. **59**(2):363-74

White CN, Proebsting WM, Hedden P, Rivin CJ (2000) Gibberellins and Seed Development in Maize. I. Evidence That Gibberellin/Abscisic Acid Balance Governs Germination versus Maturation Pathways. Plant Physiol **122**:1081-1088

Yunes JA, Vettore AL, Silva MJ, Leite A, Arruda P (1998) Cooperative DNA Binding and Sequence Discrimination by the Opaque2 bZIP Factor. Plant Cell **10**:1941-1956

Zhang JZ (2003) Overexpression analysis of plant transcription factors. Cur Op Plant Biol **6**:430-440

Tables with brief and concise titles and legends

Table1

Table1. Endosperm-preferred NAC-family transcription factors

MAS[1]	No. of sequences[2]	Highest identity[3]	e-value[4]
ZMZZEN3009C11.g	30	NAM-related protein 1-like	1.E-117
MZCCL10018G09.g[5]	***23***	***NAM-related protein 1-like***	***1.E-111***
MZCCS15005C05.g	7	NAM-related protein 1-like	4.E-85
MZCCL20006E06.g	4	NAM-related protein 1-like	6.E-74
MZCCL10127E04.g	4	OsNAC3 protein-like	3.E-32
ZMZZEN6071D10.g	3	NAM-related protein 1-like	2.E-85
MZCCL10058G04.g	2	OsNAC2 protein-like	4.E-88
ZMZZEN5053D04.g	2	NAM-related protein 1-like	5.E-67
ZMZZEN7014B11.g	2	Putative NAM (no apical meristem) protein	2.E-29
MZCCL20010G12.g	1	OsNAC1 protein-like	6.E-61
ZMZZEN7015E08.g	1	NAC2 protein-like	3.E-19
MZCCL10055H06.g	1	NAM-like protein	1.E-13

[1]MAS, Maize Assembled Sequences, are the sets of contigs and singletons (Verza et al, 2005)
[2]Number of endosperm-preferred ESTs in the cluster
[3]Best GenBank hit
[4]E-value correspondent to the best hit of the blastX of the MAS consensus sequence against the GenBank
[5]MAS corresponding to the EPN-1 gene

Original figures

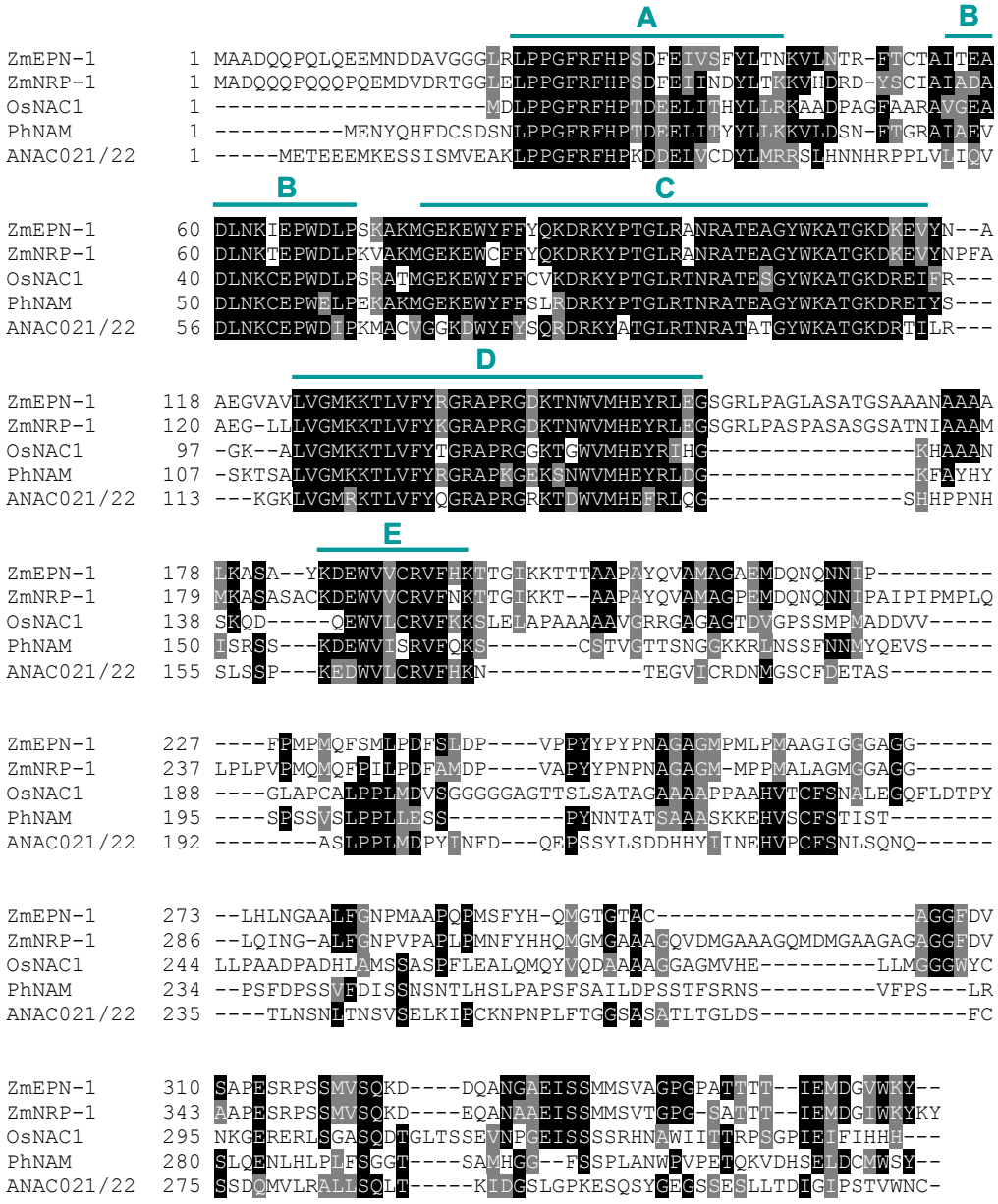


Figure 1. Alignment of five NAC-family proteins: maize EPN-1 and NRP-1 (AAP86221); rice OsNAC1 (BAC53810); petunia NAM (CAA63101) and *Arabidopsis* ANAC021/22 (Q84TE6); Subdomains A to E are shown by lines above the sequences. Amino acid identification: white on black, identical residues and white on light grey, conserved residues.

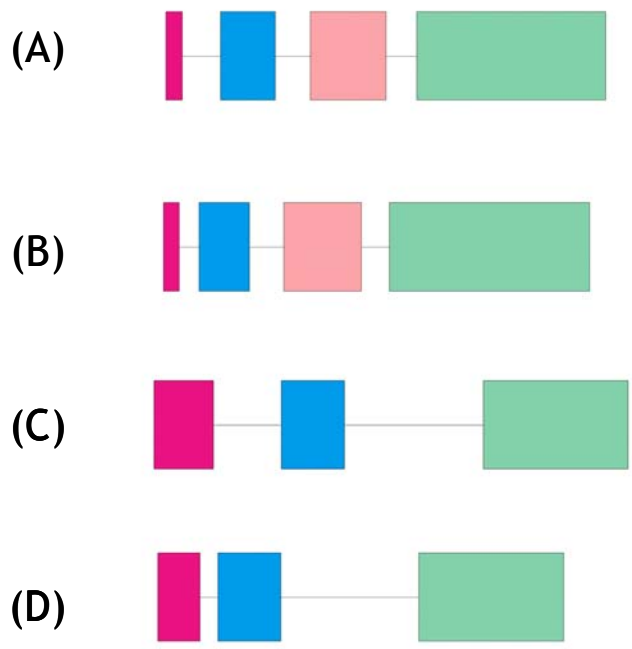


Figure 2. Schematic representation of gene structures of NAC-family members. Boxes represent exons and lines introns. Lengths are scaled up. (A) maize EPN-1; (B) maize NRP-1 (AAP86221); (C) petunia NAM (CAA63101) and (D) *Arabidopsis* CUC1 (AB049069)

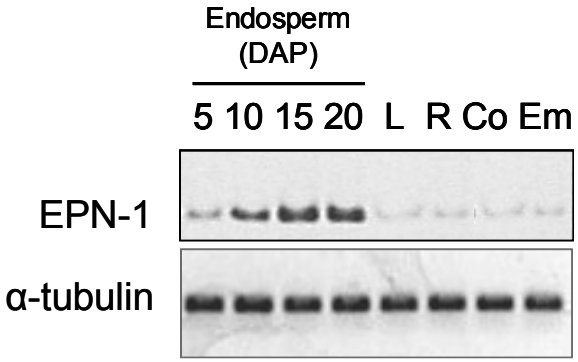


Figure 3. RT-PCR showing the endosperm-preferred expression of EPN-1, and its expression pattern during endosperm development; En: endosperm; L: leaf; R: root; Co: coleoptile; Em: embryo; DAP: days after pollination.

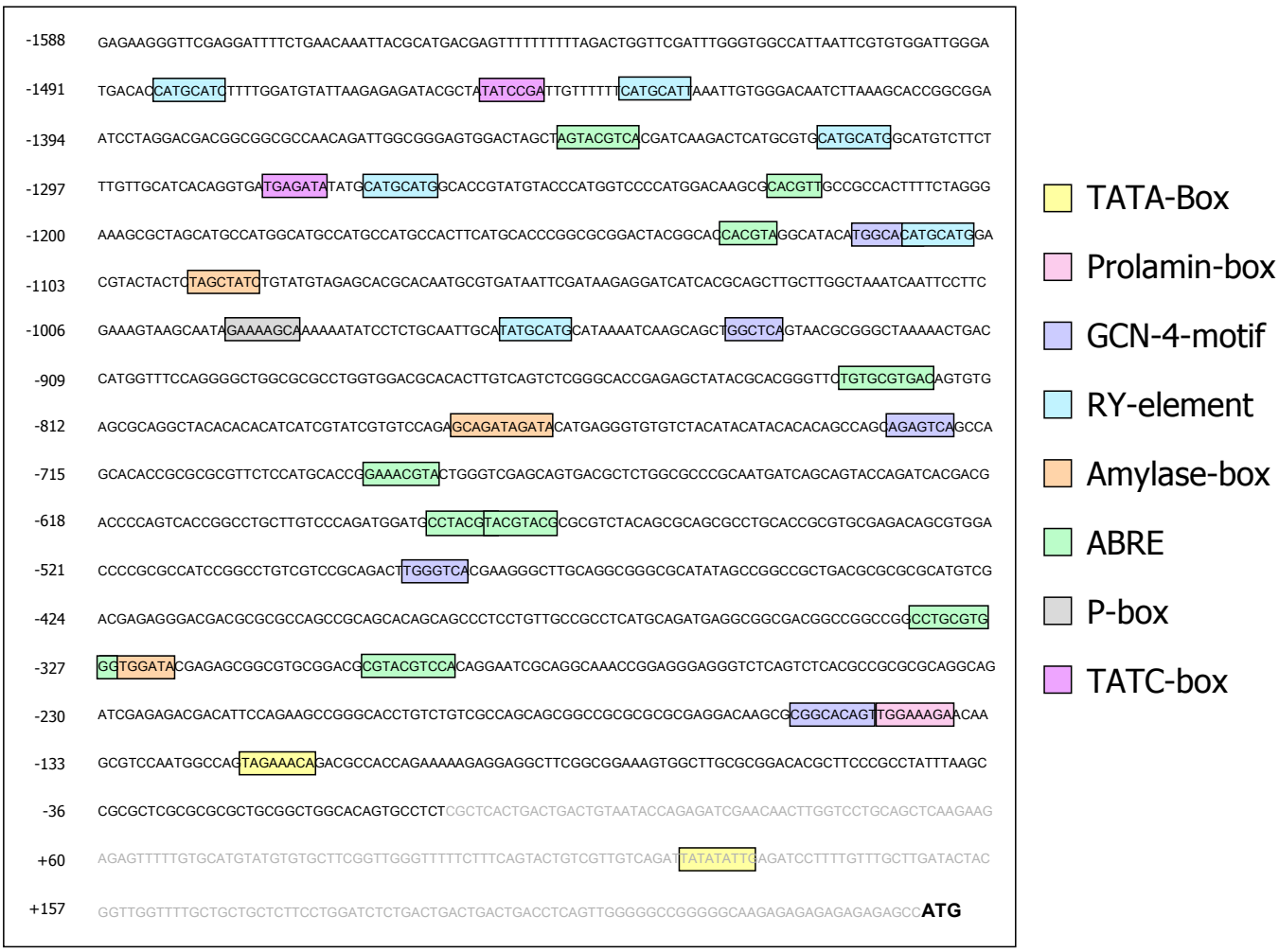
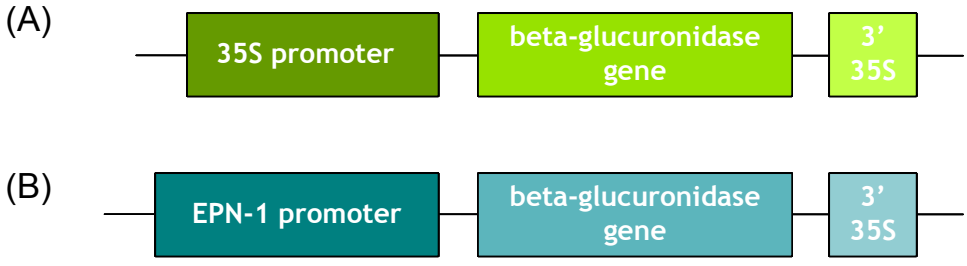
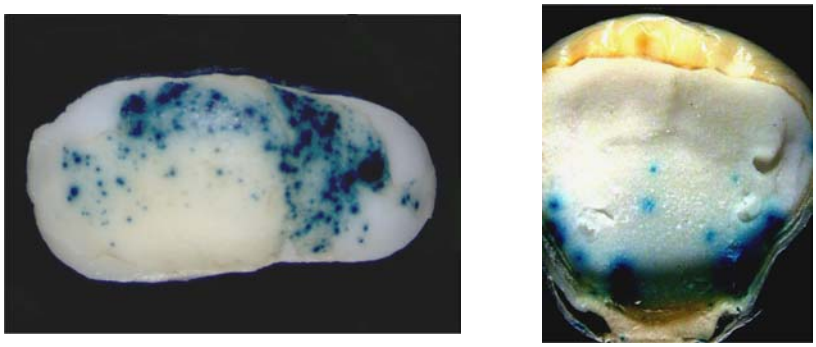


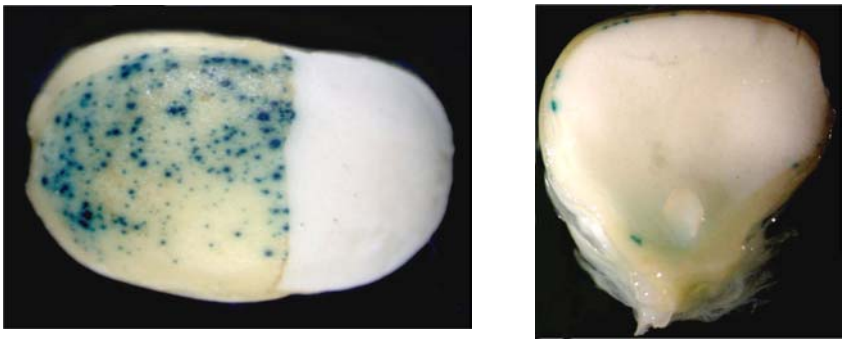
Figure 4. Conserved CIS-elements found in EPN-1 promoter; 1,9kb upstream from the initial ATG were screened using PlantCare and Place.



(C)



(D)



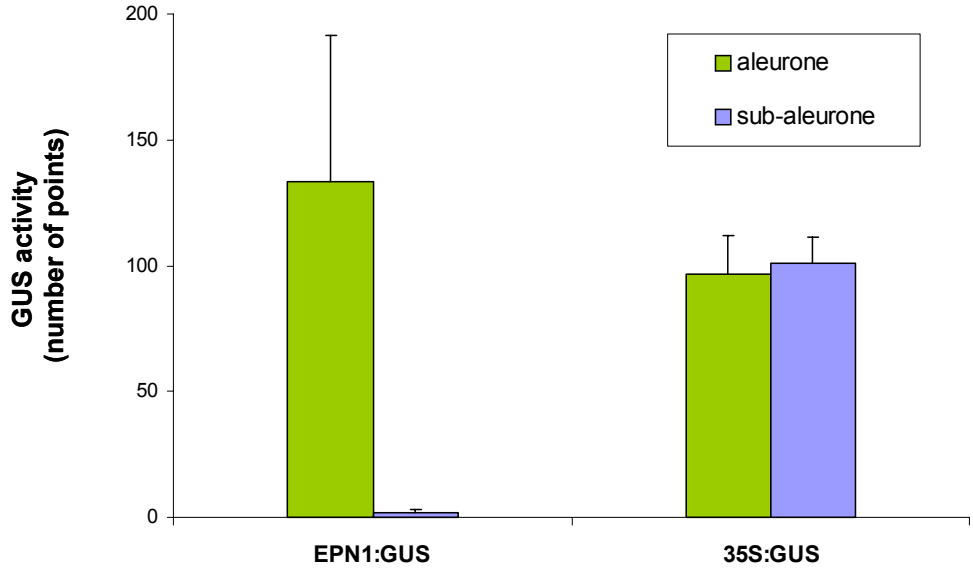


Figure 5. Spatial distribution of GUS activity driven by EPN-1 promotor in 20 DAP maize endosperm caps. Half of the aleurone layers were removed from the caps just before microprojectile bombardment with the ß-glucuronidase (GUS) gene under the control of the EPN-1 or 35S promoters. (A) Schematic representation of the pRT103GUS construct; (B) Schematic representation of the pEPN1:GUS construct; (C) p35S:GUS bombarded caps showing expression in the aleurone and sub-aleurone cell layers; (D) pEPN-1:GUS bombarded caps showing expression exclusively in the aleurone cell layer; (E) GUS activity measured as the number of blue spots within the cap area (error bars represents the SD among three biological repeats).

- As 30.531 seqüências expressas de endosperma em desenvolvimento obtidas neste trabalho, compiladas com 196.900 seqüências expressas de milho disponíveis em bancos de dados públicos, possibilitaram a construção de um banco (MAIZESTdb) contendo 227.431 ESTs provenientes dos mais diversos órgãos e tecidos de milho e representando aproximadamente 24.000 genes, o que constitui uma boa ferramenta para a prospecção e descoberta de novos genes;
- O MAIZESTdb é um banco de ESTs enriquecido com seqüências vindas de bibliotecas de cDNA construídas a partir de endosperma em desenvolvimento, o que possibilitou a identificação de mais de 80% dos genes expressos neste tecido;
- As 30.531 seqüências expressas de endosperma em desenvolvimento obtidas neste trabalho tiveram grande contribuição na descoberta de novos genes, já que a maioria dos cDNAs seqüenciados vieram de bibliotecas construídas com mRNAs extraídos no início do desenvolvimento, aos 10 e 15 DAPs, quando a expressão de genes de proteínas de reserva ainda se mantém baixa;
- A análise do banco de ESTs de diferentes órgãos e tecidos de milho possibilitou a identificação de 4.032 transcritos preferencialmente expressos no endosperma, e a sua anotação revelou uma ampla variedade de prováveis genes novos envolvidos no desenvolvimento e no metabolismo do endosperma;
- Considerando o número de genes em milho similar ao número de genes estimado em arroz, que é cerca de 40.000, o MAIZESTdb contém cerca de 60% dos genes de milho;
- A disponibilidade de grandes coleções de ESTs provenientes de diferentes tecidos de uma planta constitui uma boa ferramenta para identificação de genes órgão/tecido-específicos ou preferencialmente expressos em um órgão ou tecido através da comparação das seqüências provenientes de diferentes bibliotecas de cDNA;
- Foram identificados neste trabalho 1.233 fatores de transcrição expressos em milho, 414 dos quais expressos no endosperma em desenvolvimento;

- Foram identificados ainda, através de análises in silico, 113 fatores de transcrição preferencialmente expressos no endosperma. Este conjunto representa 9.2% dos fatores de transcrição expressos em milho identificados neste trabalho, e possivelmente contém reguladores importantes dos processos de especificação celular e desenvolvimento do endosperma de milho;
- O valor médio de redundancia encontrado entre os 414 fatores de transcrição expressos no endosperma foi de 10,4%, o que significa que nós identificamos pelo menos 369 fatores de transcrição expressos no endosperma;
- Esta é a maior coleção de fatores de transcrição já descrita para este tecido, e representa uma importante fonte de dados para identificação de reguladores dos principais processos relacionados ao desenvolvimento do endosperma, como metabolismo de nitrgênio e carboidratos e controle da massa da semente;
- Utilizando análises in silico do MAIZESTdb, nós identificamos 12 membros da família NAC de fatores de transcrição que são preferenciamente expressos no endosperma de milho;
- Um novo membro da família NAC de fatores de transcrição, chamado de EPN-1 (Endosperm Preferred NAM 1), teve seu perfil de expressão caracterizado. Sua expressão pode ser detectada desde os 5 DAPs, embora o pico de expressão ocorra entre 20 e 25 DAP, e ele apresenta expressão preferencial no endoserma em relação a outros tecidos de milho;
- O promotor do gene EPN-1 foi clonado, seqüenciado e analisado quanto aos seus possíveis elementos CIS regulatórios; foram encontrados elementos conservados relacionados a endosperma-especificidade, elementos relacionados à regulação por ácido abscisico e giberelinas, bem como elementos conservados presentes nos promotores de α-amilases, indicando uma possível relação deste gene com o processo de transição entre a maturação e a germinação da semente;
- Ensaios de expressão transitória com o promotor do gene EPN-1 revelaram que sua expressão está dirigida à camada de aleurona do endosperma de milho, o que constitui mais uma evidência de sua possível função na regulação de genes relacionados aos processos de maturação e germinação da semente.

Aida,M., Ishida,T., Fukaki,H., Fujisawa,H. & Tasaka,M. (1997). Genes involved in organ separation in *Arabidopsis*: An analysis of the cup-shaped cotyledon mutant. Plant Cell 9, 841-857.

Alleman,M. & Doctor,J. Genomic imprinting in plants: observations and evolutionary implications. Plant Molecular Biology 43, 147-161 (2000).

Andrews, J.; Bouffard, G.G.; Cheadle, C.; Lü, J.; Becker, K.G.; Oliver, B. (2000). Gene discovery using computational and microarray analysis of transcription in the Drosophila melanogaster testis. Genome Res. 10:2030-2043.

Arruda, P., Kemper, E. L., Papes, F., and Leite, A. (2000) Regulation oh lysine catabolism in higher plants. Trend. Plant Sci. 5[8]: 324-330.

Arumanagathan, K., Earle, E.D. (1991). Nuclear DNA content of some important plant species. Plant Mol Biol Rep 9:208 218.

Atchley, W.R. and Fitch, W.M. (1997). A natural classification of the basic helix–loop–helix class of transcription factors. PNAS, USA 94:5172–5176.

Audic, S.; Claverie, J.M. (1997). The significance of digital gene expression profiles. Genome Res. 7:986-995.

Avila, J., Nieto, C., Cañas, L., Benito, M.J. and Paz-Ares, J. (1993). Petunia hybrida genes related to the maize regulatory C1 gene and to animal myb proto-oncogenes. Plant J. 3(4):553–562.

Baroux,C., Spillane,C. & Grossniklaus,U. (2002).Genomic imprinting during seed development. Advanced Genetics , 164-214.

Bennetzen, J.L.; San Miguel, P.; Chen, M.; Tikhonov, A.; Francki, M. and Avramova, Z. (1998). Grass genomes. PNAS, USA 95:1975-1978.

Bürglin, T.R. (2005). Homeodomain Proteins. In Meyers, R.A. (ed.), Encyclopedia of Molecular Cell Biology and Molecular Medicine, Wiley-VCH Verlag GmbH & Co., Weinheim, 179-222.

Chasan, R. (1994) A meeting of the minds on maize. Plant Cell 6:920-925.

Chaudhury,A.M. et al. (2001). Control of early seed development. Annual Review of Cell and Developmental Biology 17, 677-699.

Choi, H., Hong, J., Ha, J., Kang, J. and Kim, S.Y. (2000). ABFs a family of ABA-responsive element binding factors. J. Biol. Chem., 275(3):1723-1730.

Chuang, C.F., Running, M.P., Williams, R.W., Meyerowitz, E. (1999). The Perianthia gene encodes a bZIP protein involved in the determination of floral organ number in Arabidopsis thaliana. Genes and Development, 13: 334-344.

Ciceri, P., Locatelli, F., Genga, A., Viotti, A. and Schmidt, R.J. (1999). The activity of the maize Opaque-2 transcriptional activation is regulated diurnally. Plant Physiology, 121(4): 1321 - 1327.

Coen, E.S., and Meyerowitz, E.M. (1991). The war of the whorls: genetic interactions controlling flower development. Nature 353, 31-37.

Collinge,M. & Boller,T. (2001). Differential induction of two potato genes, Stprx2 and StNAC, in response to infection by Phytophthora infestans and to wounding. Plant Molecular Biology 46, 521-529.

Consonni G, Geuna F, Gavazzi G, Tonelli C. Molecular homology among members of the R gene family in maize. Plant J. 1993;3:335–346.

Consonni, G., Gavazzi, G. and Dolfini, S. (2005). Genetic Analysis as a Tool to Investigate the Molecular Mechanisms Underlying Seed Development in Maize. Annals of Botany 96(3):353-362.

Cook, R.J. (1998). Towards a successful multinational crop plant genome initiative. PNAS, USA 95:1993-1995.

Cord-Neto, G., Yunes,J.A., da Silva,M.J., Vettore,A.L., Arruda,P., and Leite,A. (1995) The involvement of Opaque 2 on beta-prolamin gene regulation in maize and Coix suggests a more general role for this transcriptional activator. Plant Mol.Biol. 27:1015-1029.

Danilevskaya,O.N. et al. (2003). Duplicated fie Genes in Maize: Expression Pattern and Imprinting Suggest Distinct Functions. The Plant Cell 15, 425-438.

Despres, C., DeLong, C., Glaze, S., Liu, E. and Fobert, P.R. (2000). The Arabidopsis NPR1/NIM1 protein enhances the DNA binding activity of subgroup of the TGA family of bZIP transcription factors. Plant Cell 12(2): 179-81.

Dong, Q.; Roy, L.; Freeling, M.; Walbot, V.; Brendel, V. (2003). ZmDB, an integrated database for maize genome research. Nuclei. Aci. Res. 31:244–247.

Duval,M., Hsieh,T.F., Kim,S.Y. & Thomas,T.L. (2002). Molecular characterization of AtNAM: a member of the *Arabidopsis* NAC domain superfamily. Plant Mol. Biol. 50, 237-248.

Ernst, H. A., Olsen, A. N., Skriver, K., Larsen, S., and Lo Leggio, L. (2004) Structure of the conserved domain of ANAC, a member of the NAC family of plant specific transcription factors. EMBO Rep. 5, 297-303.

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. v.9, p.950-959.

Fukasawa, J., Sakai, T., Ishida, S., Yamaguchi, I., Kamiya, Y. and Takahashi, Y. (2000). Repression of shoot growth, a bZIP transcriptional activator, regulates all elongation by controlling the level of gibberellins. Plant Cell, 12: 901-915.

Gale, M.D. and Devos, K.M. (1998) Comparative genetics in the grasses. PNAS, USA 95(5):1971-1974.

Gallusci,P., Varotto,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.

Ge,L. et al. (2004). Overexpression of OsRAA1 causes pleiotropic phenotypes in transgenic rice plants, including altered leaf, flower, and root development and root response to gravity. Plant Physiol. 135, 1502-1513.

Gehring, W.J., Affolter, M. and Bürglin, T.R. (1994) Homeodomain proteins. Annu. Rev. Biochem., 63, 487-526.

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.

Grivet, L. and Arruda, P. (2001). Sugarcane genomics: depicting the complex genome of an important tropical crop. Cur. Opi. Plant Biol. v.5, p.122-127.

Grossniklaus,U., Vielle-Calzada,J.P., Hoeppner,M.A. & Gagliano,W.B. (1998). Maternal control of embryogenesis by medea, a Polycomb group gene in *Arabidopsis*. Science 280, 446-450.

Guo,M., Rupe,M.A., Danilevskaya,O.N., Yang,X.F. & Hut,Z.H. (2003). Genome-wide mRNA profiling reveals heterochronic allelic variation and a new imprinted gene in hybrid maize endosperm. Plant Journal 36, 30-44.

Guo,Y., Cai,Z. & Gan,S. (2004). Transcriptome of *Arabidopsis* leaf senescence. Plant Cell and Environment 27, 521-549.

Hake, S.C. and Walbot, V. (1980). The genome of Zea mays, its organization and homology to related species. Chromosoma 79: 251-270.

Hurst, H.C. (1995). Leucine Zippers Transcription Factors. 72p. San Diego: ACADEMIC PRESS.

John,I. et al. (1997). Cloning and characterization of tomato leaf senescence-related cDNAs. Plant Molecular Biology 33, 641-651.

Kemper,E.L., Neto,G.C., Papes,F., Moraes,K.C., Leite,A., and Arruda,P. (1999) The role of opaque2 in the control of lysine-degrading activities in developing maize endosperm. Plant Cell 11:1981-1994.

Kikuchi,K. et al. (2000). Molecular analysis of the NAC gene family in rice. Molecular and General Genetics 262, 1047-1051.

Kornberg, R.D. (1999). Eukaryotic transcriptional control. Trend Cell Biol., 12:46-49.

Kuhlemeier, C. (1992). Transcriptional and post-transcriotional regulation of gene expression in plants. Plant Mol. Biol., 19:1-14.

Landschultz, W.H., Johnson, P.F. and Mcknight, S.L. (1998). The leucine zipper: A hypotetical struture common a new class of DNA binding protein. Science, 240:1759-1764.

Lazzeri, E.L. and Shewry, P.R. (1993) Biotechnology of cereals. Biotechnol. Genet. Eng. Rev. 11:79-146.

Lee, T.I. and Young, R.A. (2000). Transcription of eukaryotic protein-coding genes. Annual Rev. Genet., 34:77-137.

Leiva-Neto J.T., Grafi G., Sabelli P.A., Dante R.A., Woo Y.M., Maddock S., Gordon-Kamm W.J., Larkins B.A. (2004). A dominant negative mutant of cyclin-dependent kinase A reduces endoreduplication but not cell size or gene expression in maize endosperm. Plant Cell 16(7):1854-69.

Lopes, M.A. and Larkins, B.A. (1993) Endosperm origin, development and function. Plant Cell 5:1383-1389.

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.

Ludwig S R, Habera L F, Dellaporta S L and Wessler S R. (1989). Lc, a member of the maize R gene family responsible for tissue-specific anthocyanin production, encodes a protein similar to transcriptional activators and contains the myc-homology region. Proc. Natl. Acad. Sci. USA., 86: 7092-7096.

Luo,M. et al. (1999). Genes controlling fertilization-independent seed development in *Arabidopsis thaliana*. Proceedings of the National Academy of Sciences of the United States of America 96, 296-301.

Luscombe, N.M., Austin, S.E., Berman, H.M. and Thornton, J.M. (2000). An overview of the structures of protein-DNA complexes. Genome biology, 1:1-37.

Ma, H.; Schulze, S.; Lee, S.; Yang, M.; Mirkov, E.; Irvine, J.; Moore, P.; Paterson, A. (2003). An EST survey of the sugarcane transcriptome. Theor. Ap. Gen. 108:851-863.

Mertz, E. T., Bates, L. S., and Nelson, O. E. (1964) Mutant gene that changes protein composition and increases lysine content of maize endosperm. Science 145: 279-280.

Meshi, T. and Iwabuchi, M. (1995). Plant transcription factors. Plant Cell Physiol., 36(8):1405-1420.

MGDb - Maize Genetics/Genomics Database project homepage: http://www.maizegdb.org/

Millward DJ (1999). The nutritional value of plant-based diets in relation to human amino acid and protein requirements. Proceedings of the Nutrition Society, 58: 249-260

Motto, M., Maddaloni, M., Brembilla, M., Ponziani, G., Marotta, R., Di Fonzo, N., Soave, C., Thompson, R., and Salamini, F. (1988) Molecular cloning of the o2-m5 allele of Zea mays using transposon marking. Molecular Genetics and Genomics 212: 488-504.

Näär, A.M., Lemon, B.D., Tjian, R. (2001). Transcriptional coactivator complexes. Annu Rev Biochem. 70:475-501. Review.

Niggeweg R., Thurow C., Kegler C., Gatz C. (2000). Tobacco transcription factor TGA2.2 is the main component of as-1-binding factor ASF-1 and is involved in salicylic acid- and auxin-inducible expression of as-1-containing target promoters. J. Biol. Chem. 275:19897-19905.

Offler, C.E., McCurdy, D.W., Patrick, J.W. and Talbot, M.J. (2003). Transfer cells: cells specialized for a special purpose. Annu Rev Plant Biol. 54:431-54.

Olsen, O.A. (2001). Endosperm development: Cellularization and cell fate specification. Annual Review of Plant Physiology and Plant Molecular Biology 52: 233-267.

Olsen, O.A. (2004) Nuclear endosperm development in cereals and *Arabidopsis thaliana*. Plant Cell 16 Suppl:S214-27.

Onodera,Y., Suzuki,A., Wu,C.Y., Washida,H., and Takaiwa,F. (2001) A rice functional transcriptional activator, RISBZ1, responsible for endosperm-specific expression of storage protein genes through GCN4 motif. J.Biol.Chem. 276:14139-14152.

Opsahl-Ferstad, H.G., Le Deunff, E., Dumas, C. and Rogowsky, P.M. (1997). ZmEsr, a novel endosperm-specific gene expressed in a restricted region around the maize embryo. Plant J. 12(1):235-46.

Osterlund, M.T., Wei, N. and Deng, X.W. (2000). The roles of photorreceptor system and the COP-1 targeted distabilization of HY5 in the light control of Arabidopsis seedling development. Plant Physiol., 124: 1520-1524.

Pabo, C.O. and Sauer, R.T. (1992). Transcriptional factors:Strutural families and principles of DNA recognition. Ann. Rev. Biochem., 61:1053-1095.

Pontier, D., Miao, Z-H and Lam, E. (2001). Trans-dominant suppression of plant TGA factors reveals their negative and positive roles in plant defense responses. The Plant Journal 27(6): 529-538.

Radicella J P, Turks D and Chandler L V. (1991) Cloning and nucleotide sequence of a cDNA encoding B-Peru, a regulatory protein of anthocyanin pathway in maize. Plant Mol. Biol., 17: 127-130.

Russell, S.D. (1992) Double fertilization. Int. Rev. Cytol. 140:357-388.

Sablowski,R.W.M. and Meyerowitz,E.M. (1998). A homolog of NO APICAL MERISTEM is an immediate target of the floral homeotic genes APETALA3/PISTILLATA. Cell 92, 93-103.

Satoh R, Fujita Y, Nakashima K, Shinozaki K and Yamaguchi-Shinozaki K. (2004). A Novel Subgroup of bZIP Proteins Functions as Transcriptional Activators in Hypoosmolarity-Responsive Expression of the ProDH Gene in Arabidopsis. Plant Cell Physiol. 45(3):309-317.

Schuler, G. (1997). Pieces of the puzzle: expressed sequences tags and the catalog of human genes. J. Mol. Med. 75:694-698.

Schmidt,R.J., Burr,F.A., and Burr,B. (1987) Transposon tagging and molecular analysis of the maize regulatory locus opaque-2. Science 238:960-963.

Schmidt,R.J., Ketudat,M., Aukerman,M.J., and Hoschek,G. (1992) Opaque-2 is a transcriptional activator that recognizes a specific target site in 22-kD zein genes. Plant Cell 4:689-700.

Souer, E., van Houwelingen, A., Kloos, D., Mol, J. and Koes, R. (1996). The no apical meristem gene of petunia is required for pattern formation in embryos and flowers and is expressed at meristem and primordia boundaries. Cell 85, 159-170.

Telles, G.P.; Braga, M.D.V; Dias, Z.; Lin, T-L; Quitzau, J.A.A.; Da Silva, F.R.; Meidanis, J. (2001). Bioinformatics of the sugarcane EST project. Gen. Mol. Biol. 24:9-15.

Thompson, R.D., Hueros, G., Becker, H. and Maitz, M. (2001). Development and functions of seed transfer cells. Plant Sci. 160(5):775-783.

Tran, L.S.P. et al. (2004). Isolation and functional analysis of *Arabidopsis* stress-inducible NAC transcription factors that bind to a drought-responsive cis-element in the early responsive to dehydration stress 1 promoter. Plant Cell 16, 2481-2498.

Uno, Y., Furihata, T., Abe, H., Yoshida, R., Shinozaki, K. and Yamagushi-Shinozaki, K. (2000). Arabidopsis basic leucine zipper transcription factors involved in as abscisic acid-dependent signal transduction pathway under drought and high salinity conditions. Proceedings of the National Academy of Sciences of the United States of America, 97: 11632-11637.

Ulm R, Baumann A, Oravecz A, Mate Z, Adam E, Oakeley EJ, Schafer E and Nagy F. (2004). Genome-wide analysis of gene expression reveals function of the bZIP transcription factor HY5 in the UV-B response of Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America 101(5):1397-1402.

Vettore, A.L.; Da Silva, F.R.; Kemper, E.L.; Arruda, P. (2001). The libraries that made SUCEST. Gen. Mol. Biol. 24:1-7.

Vettore, A. L., Yunes, A. J., Cord Neto, G, da Silva, M. J., Arruda, P., and Leite, A. (1998) The molecular and functional characterization of an Opaque 2 homologue gene from Coix and a new classification of plant bZIP proteins. Plant Molecular Biology 36: 249-263.

Walsh, J.W. and Feeling, M. (1997). The maize gene liguleles 2 encodes a basic leucine zipper involved in the establishment of the leaf blade-sheath boundary. Genes and development, 11: 208-218.

White, J.A.; Todd, J.; Newman, T.; Focks, N.; Girke, T.; De Ilarduya, O. M.; Jaworski, J.G.; Ohlrogge, J. B.; Benning, C. (2000). A new set of *Arabidopsis thaliana* expressed sequence tags from developing seeds. The metabolic pathway from carbohydrates to seed oil. Plant Physiol. 124:1582–1594.

Wu, J. et al. (2002). A comprehensive rice transcript map containing 6591 expressed sequence tag sites. Plant Cell 14:525–535.

Xie,Q., Frugis,G., Colgan,D. & Chua,N.H. (2000). *Arabidopsis* NAC1 transduces auxin signal downstream of TIR1 to promote lateral root development. Genes and Development 14, 3024-3036.

Xie,Q., Sanz-Burgos,A.P., Guo,H.S., Garcia,J.A. and Gutierrez,C. (1999). GRAB proteins, novel members of the NAC domain family, isolated by their interaction with a geminivirus protein. Plant Molecular Biology 39, 647-656.

Yunes,J.A., Vettore,A.L., da Silva,M.J., Leite,A., and Arruda,P. (1998) Cooperative DNA binding and sequence discrimination by the Opaque2 bZIP factor. Plant Cell 10:1941-1955.

Zhang, Y., Fan, W., Kinkeman, M., Li, X. and Dong, X. (1999). Interaction of NPR1 with basic leucine zipper protein transcription fator that bind sequences required for salicylic acid induction of the PR-1 gene. Proceedings of the National Academy of Sciences of the United States of America, 96: 6523-6528.