

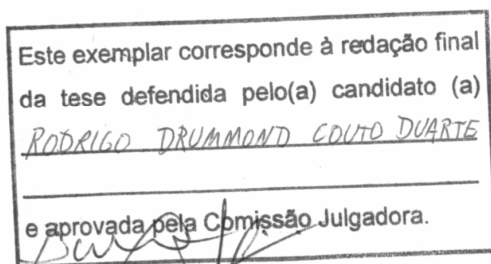
UNIVERSIDADE ESTADUAL DE CAMPINAS



Rodrigo Drummond Couto Duarte

**Métodos estatísticos para seleção de genes diferencialmente
expressos a partir de dados de arranjos de DNA:**

**Aplicação à análise da expressão gênica induzida em cana-de-açúcar
por deficiência de fosfato**



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

Orientador: Prof. Dr. Marcelo Menossi Teixeira

Co-Orientador: Prof. Dr. Aluísio de Souza Pinheiro

Campinas, 2006

**FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DO INSTITUTO DE BIOLOGIA – UNICAMP**

D85m

Duarte, Rodrigo Drummond Couto
Métodos estatísticos para seleção de genes
diferencialmente expressos a partir de dados de arranjos de
DNA: aplicação à análise da expressão gênica induzida em
cana-de-açúcar por deficiência de fosfato / Rodrigo
Drummond Couto Duarte . -- Campinas, SP: [s.n.], 2006.

Orientador: Marcelo Menossi Teixeira
Co-Orientador: Aluísio de Souza Pinheiro
Tese (doutorado) – Universidade Estadual de
Campinas, Instituto de Biologia.

1. Arranjos de DNA. 2. Análise de dados quantitativos.
3. Fosfato. 4. Cana-de-açúcar. I. Teixeira, Marcelo
Menossi. II. Pinheiro, Aluísio de Souza. III. Universidade
Estadual de Campinas. Instituto de Biologia. IV. Título.

Amr/ib

Título em inglês: Statistical methods for the selection of differentially expressed genes from DNA array data: application to the analysis of the gene expression induced in sugarcane by phosphate deficiency

Palavras-chave em inglês: DNA arrays, Data analysis, Phosphate deficiency, Sugarcane

Área de concentração: Bioinformática

Titulação: Doutor em Genética e Biologia Molecular.

Banca examinadora: Marcelo Menossi Teixeira, Ana Tereza Ribeiro de Vasconcelos, Gonçalo Amarante Guimarães Pereira, Sérgio Furtado dos Reis, Vicente Eugenio de Rosa Junior

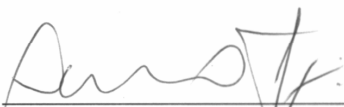
Data da defesa: 18/12/2006.

Programa de Pós-Graduação: Genética e Biologia Molecular.

Campinas, 18 de dezembro de 2006

BANCA EXAMINADORA

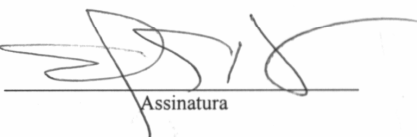
Prof. Dr . Marcelo Menossi Teixeira (Orientador)


Assinatura

Profa. Dra. Ana Tereza Ribeiro de Vasconcelos


Assinatura

Prof. Dr. Gonalo Amarante Guimarães Pereira


Assinatura

Prof. Dr. Srgio Furtado dos Reis


Assinatura

Dr. Vicente Eugenio de Rosa Junior


Assinatura

Prof. Dra. Glaucia Mendes Souza


Assinatura

Prof. Dr. Antonio Augusto Franco Garcia


Assinatura

Prof. Dr. Michel Georges Albert Vincentz


Assinatura

Dedico essa tese à memória de

meu pai, Sérgio Duarte,
minha avó, Isabel Drummond, e
do amigo Prof. Aquiles Eugênico
Piedrabuena.

Os três me apoiaram muito no início desse trabalho, e infelizmente não chegaram a vê-lo terminado.

Dedico também à minha mãe, Andréa Couto, e à minha esposa, Carla Vizeu. Com elas aprendi que o amor é o que mais importa.

Agradecimentos

Este trabalho foi financiado pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Agradeço a todos os colegas do laboratório pelas reuniões, discussões, idéias, enfim, pelo apoio e pela companhia ao longo do caminho. Não vou tentar citar todos os nomes, mas agradeço especialmente a Ju Felix, que foi a primeira a me chamar para trabalhar no CBMEG, e ao Renato Vicentini por toda a ajuda computacional. Também quero agradecer ao meu orientador, que sempre foi “Mano” e apostou em um aluno que não sabia ao certo se ainda era matemático ou se já tinha virado biólogo, e ao meu co-orientador, que me ajudou a recuperar minhas raízes matemáticas.

Agradeço aos grupos de pesquisa coordenados pelos professores Antônio Figueira (CENA-USP) e Gláucia Souza (IQ-USP), principalmente ao Raul Santin Almeida e à Flávia Rocha, pelas colaborações neste trabalho, sem as quais o mesmo não seria possível.

Novamente sem citar nomes, agradeço a todos os meus amigos e irmãos, que dividiram comigo os momentos de dúvidas, preocupações e, principalmente, de alegria. Meus agradecimentos também para minha mãe, pelo amor e apoio incondicionais, e para minha esposa, pela enorme paciência durante os momentos de estresse e por fazer parte da minha vida (sem ela tudo teria menos sentido).

Agradeço também ao meu pai espiritual Mário d'Ogum, que com seus conselhos e seu amor tornou mais fáceis as últimas etapas desse trabalho.

Por fim, gostaria de agradecer aos membros da banca por me darem a honra de participar comigo do último passo da longa caminhada que esse trabalho representou.

Resumo

Arranjos de DNA são uma poderosa técnica de monitoramento da expressão gênica em larga escala. No entanto, a grande quantidade de dados gerados com esse tipo de experimento requer um tratamento estatístico adequado às suas características. Uma aplicação importante dos arranjos de DNA é a identificação de genes diferencialmente expressos em diferentes amostras de RNA. Essa seleção demanda testes estatísticos apurados, capazes de distinguir, entre o grande número de genes usualmente presentes nos arranjos, aqueles cuja expressão é significativamente diferenciada. Neste trabalho nós desenvolvemos algoritmos para a análise estatística dos dados provenientes de arranjos de DNA, eficientes em lidar com os problemas usuais nesse tipo de dados, como o número limitado de réplicas. Aplicados a dados simulados, os algoritmos desenvolvidos mostraram-se competitivos com outros métodos de análises descritos na literatura, superando-os em algumas situações. A aplicação desses algoritmos foi também demonstrada em um experimento voltado à identificação de genes de cana-de-açúcar diferencialmente expressos em resposta a deficiência de fosfato. O fósforo é um macronutriente essencial, captado pelas plantas principalmente na forma de fosfato inorgânico (P_i). A deficiência de fosfato é freqüente na natureza, especialmente nos solos ácidos das regiões tropicais e subtropicais. Devido a grande importância econômica da cana-de-açúcar, a identificação de genes diferencialmente expressos em resposta à deficiência de fosfato nesta espécie é de grande interesse científico e agrônomo. Algoritmos de agrupamento foram também aplicados aos dados de expressão obtidos no experimento, identificando padrões de expressão gênica nos diferentes estágios da resposta da cana a esse estresse e proporcionando assim uma caracterização adicional da mesma. Futuramente, os resultados desse trabalho podem conduzir ao desenvolvimento de linhagens de cana-de-açúcar com melhor desempenho em solos pobres de fosfato, o que seria de extremo interesse agrônomo.

Abstract

DNA arrays are a powerful technique for monitoring gene expression in large scale. However, the great amount of data generated by this kind of experiment requires a statistical treatment adequate to its characteristics. An important application of DNA arrays is the identification of differentially expressed genes in different RNA samples. This selection demands refined statistical tests, able of distinguish among the great number of genes usually present in the arrays those which expression is significantly different. In this work, we have developed algorithms for the analysis of DNA array data, efficient in handling the usual problems in this kind of data, as the limited number of replicates. When applied to data simulations the developed algorithms showed to be competitive with other methods of analysis described in the literature and widely used, overperforming them in some situations. The application of these algorithms was also demonstrated in an experiment devoted to the identification of sugarcane genes differentially expressed in response to phosphate deficiency. Phosphorous is an essential macronutrient, absorbed by plants mostly in the form of inorganic phosphate (P_i). The phosphate deficiency is frequent in the nature, especially in the acid soils of tropical and subtropical areas. Because of the great economical importance of sugarcane, the identification of genes that are differentially expressed in response to phosphate deficiency in this species is of great scientific and agronomic interest. Clustering algorithms were also applied to the expression data obtained in the experiment, identifying patterns of gene expression in the different stages of the sugarcane response to the stress, thus providing an additional characterization of it. In the future, the results of this work can lead to the development of sugarcane cultivars that have better performance in phosphate deficient soils, what would be of great agronomic interest.

Índice

Introdução.....	1
Arranjos de DNA.....	2
Análise de dados de arranjos de DNA.....	3
A cana-de-açúcar.....	6
A deficiência de Fosfato.....	8
Estrutura do trabalho.....	11
Objetivos.....	15
Capítulo 1: Padronização do uso de macroarranjos em membranas de <i>nylon</i>	17
<i>Cross-species hybridization of nylon arrays containing sugarcane ESTs to identify Al-induced genes from maize....</i>	19
<i>Evaluation of the performance of different plastics used to seal nylon cDNA arrays.....</i>	49
Capítulo 2: Pré-processamento e normalização dos dados.....	65
2.1 - Subtração do <i>background</i>	65
2.2 – Identificação de problemas e filtragem dos dados.....	67
2.3 – Normalização dos dados.....	70
2.4 – Transformações dos dados.....	76
Capítulo 3: A aplicação de testes estatísticos clássicos à seleção de genes diferencialmente expressos.....	81
3.1 – Aplicações do teste t de Student.....	82
3.2 – Aplicações da análise de variância (ANOVA).....	84
3.3 – Correção de Bonferroni.....	86
<i>Identification of genes preferentially expressed in the pathogenic yeast phase of Paracoccidioides brasiliensis, using suppression subtraction hybridization and differential macroarray analysis.....</i>	89

<i>Transcriptionally active transposable elements in recent hybrid sugarcane.....</i>	101
<i>Isolation and characterization of Coffea genes induced during coffee leaf miner (Leucoptera coffeella) infestation.....</i>	113
Capítulo 4: Testes estatísticos específicos para dados de arranjos de DNA.....	123
4.1 – Seleção de genes baseada na distribuição global dos logaritmos das razões de expressão.....	123
4.2 – Estratégias de controle do número de falsos-positivos.....	125
4.3 – Significance Analysis of microarray (SAM).....	127
<i>Gene expression profiling of aluminum stress in maize roots.....</i>	131
Capítulo 5: O algoritmo ISER – seleção de genes diferencialmente expressos por transformações não lineares e ajustes locais.....	181
<i>ISER: selection of differentially expressed genes from DNA array data by non-linear transformations and local fitting.....</i>	183
<i>ISER: application to simulated data sets.....</i>	187
<i>ISER: analysis of a real macroarray data set.....</i>	205
Capítulo 6: O algoritmo ISER 2 – seleção de genes diferencialmente expressos em experimentos com réplicas.....	209
6.1 – Experimentos apenas com réplicas técnicas, sem réplicas biológicas.....	210
6.2 – Experimentos com réplicas técnicas e biológicas.....	214
6.3 - O controle do FDR no algoritmo ISER 2.....	221
6.3.1 – Mistura beta-uniforme.....	225
6.3.2 – Mistura beta-uniforme-beta.....	228
6.3.3 – FDR e q-valores.....	231

6.4 – Testes com dados simulados.....	232
6.4.1 – Modelo adotado nas simulações de dados.....	233
6.4.2 – Análise dos dados simulados.....	237
6.4.3 – Resultados das simulações.....	238
Capítulo 7: Efeito da deficiência de fosfato na expressão gênica de cana-de-açúcar.....	245
7.1 – Tratamento vegetal e extração de RNA.....	246
7.2 – Hibridações de microarranjos de DNA.....	247
7.3 – Análise preliminar dos dados.....	251
7.4 – Resultados da algoritmo ISER.....	253
7.5 – Resultados da algoritmo ISER2.....	255
7.6 – Confirmação da expressão diferencial por real-time PCR.....	260
7.7 – Discussão dos resultados.....	262
Capítulo 8: Agrupamentos dos genes de cana-de-açúcar cuja expressão responde à deficiência de fosfato.....	271
8.1 – Algoritmos hierárquicos de agrupamento na visualização de dados de expressão gênica.....	278
8.2 – Mapas auto organizáveis (<i>self-organizing maps</i>).....	281
Conclusão.....	295
Material e Métodos.....	297
Referências.....	303

Introdução

Os projetos genoma têm levado à identificação de um número cada vez maior de genes das mais diferentes espécies de organismos. Com isso, tornou-se necessário desenvolver estudos voltados à identificação da função desses genes, o que pode ser feito medindo sua expressão (as concentrações relativas de seus transcritos) em diferentes tecidos ou tratamentos. Recentes avanços tecnológicos tornaram possível o estudo simultâneo da expressão de milhares de genes através do uso de arranjos (*arrays*) de DNA (Schena *et al.*, 1998; Freeman *et al.*, 2000), abrindo novas possibilidades para a compreensão dos mecanismos de regulação gênica.

Os dados provenientes de arranjos de DNA possibilitam a identificação de genes diferencialmente expressos em diferentes indivíduos, tecidos ou condições fisiológicas. Para isso, porém, é necessária a aplicação de testes estatísticos apurados, capazes de distinguir, entre o grande número de genes usualmente presentes nos arranjos, aqueles cuja expressão é significativamente diferenciada. Desta forma, o desenvolvimento de testes estatísticos adequados às características desse tipo de dados é extremamente importante para maximizar o aproveitamento dos resultados de experimentos que empregam essa metodologia.

Neste trabalho foram desenvolvidos algoritmos para a análise estatística dos dados provenientes de arranjos de DNA, capazes de lidar eficientemente com os problemas usuais nesse tipo de dados, como o número limitado de réplicas. Em simulações de dados os algoritmos desenvolvidos mostraram-se competitivos com outros métodos de análises descritos na literatura e largamente utilizados,

superando-os em algumas situações. Além disso, a aplicação desses algoritmos foi demonstrada em um experimento voltado à identificação de genes de cana-de-açúcar diferencialmente expressos em resposta a deficiência de fosfato.

Arranjos de DNA

Um arranjo de DNA consiste em um suporte sólido (lâmina de vidro ou membrana de *nylon*) no qual são fixados, de forma ordenada, fragmentos de DNA correspondentes aos genes que serão estudados (em geral obtidos de uma biblioteca de cDNA). Assim em cada ponto do arranjo existe cDNA ou oligonucleotídeo correspondente a um gene. O arranjo é então submetido à hibridação com uma sonda de cDNA marcado com um isótopo radiativo ou um fluoróforo (Freeman *et al.*, 2000), obtido a partir do RNA total ou poli-A de uma amostra, de modo que as fitas de cDNA da sonda se pareiem com os fragmentos no suporte (Schena *et al.*, 1995, Zhao *et al.*, 1995, Desprez *et al.*, 1998, Schummer *et al.*, 1999). Desta forma, medindo a emissão radioativa ou fluorescência de cada ponto do suporte, obtemos para cada gene um valor proporcional à abundância de seus transcritos na amostra de RNA. Os arranjos de DNA podem ser divididos em micro e macroarranjos, de acordo com o suporte utilizado e com a densidade de pontos de fixação de DNA (Freeman *et al.*, 2000). Costuma-se chamar de macroarranjos (ou *nylon arrays*, *filter arrays* ou *high density membranes*) os arranjos em membranas de *nylon*, que apresentam uma densidade de *spots* menor e empregam sondas marcadas radiativamente. O termo microarranjo refere-se aos arranjos em lâminas de vidro, também chamados de DNA chips, que costumam empregar sondas marcadas com fluoróforos. Estes

últimos são muito mais utilizados que os arranjos em membranas de *nylon*, apesar de serem relativamente mais dispendiosos. Uma grande vantagem dos microarranjos de DNA é a possibilidade de hibridar simultaneamente ao arranjo duas sondas marcadas com fluoróforos diferentes, o que facilita muito a comparação entre os níveis de expressão gênica nas sondas. Por outro lado, os macroarranjos constituem uma opção mais acessível para a maioria dos laboratórios de biologia molecular, uma vez que empregam metodologias menos dispendiosas, robustas e de fácil implementação.

Através da hibridação de arranjos de DNA com sondas obtidas de diferentes amostras de RNA (diferentes tecidos ou tratamentos), pode-se identificar genes cuja expressão varia de uma amostra para outra. Com esta tecnologia é possível encontrar grupos de genes envolvidos em diferentes processos biológicos, a exemplo dos estudos sobre hipertrofia cardíaca em camundongos e resistência sistêmica adquirida em *Arabidopsis*, realizados respectivamente por Friddle *et al.* (2000) e Maleck *et al.* (2000). Outros desenhos experimentais também podem ser empregados em experimentos com arranjos de DNA, como por exemplo séries temporais (Yang e Speed, 2002).

Análise de dados de arranjos de DNA

A grande quantidade de dados gerados por experimentos utilizando arranjos de DNA requer uma metodologia de análise específica, devido às suas características peculiares. Uma dificuldade inerente à análise desses dados é a presença de diversas fontes de flutuações aleatórias, como variações na quantidade de DNA fixado em cada ponto dos arranjos, diferenças na eficiência da

marcação das sondas e em sua hibridação aos arranjos, diferenças na detecção e quantificação dos sinais obtidos a cada ponto, dentre outras (Freeman *et al.*, 2000). Além disso, outra dificuldade usual na análise dos dados de arranjos de DNA é a dimensão dos mesmos: um grande número de genes é analisado em paralelo (o que implica em múltiplos testes estatísticos), enquanto o número de réplicas experimentais é em geral reduzido (o que reduz o poder dos testes estatísticos aplicados).

Um aspecto delicado da análise de dados de arranjos de DNA é o pré-processamento dos mesmos, que procura torná-los mais adequados à aplicação de testes estatísticos. Com o objetivo de minimizar tendências sistemáticas, torna-se necessário normalizar os dados provenientes de arranjos de DNA, e diversas estratégias de normalização são largamente adotadas (Quackenbush, 2002, Yang *et al.*, 2002, Smith e Speed, 2003). Procurando ainda tornar esses dados mais adequados a aplicação de testes estatísticos, é usual aplicar transformações aos mesmos, com o objetivo de obter valores com distribuições conhecidas. A transformação logarítmica é a mais freqüentemente adotada (Quackenbush, 2002), uma vez que os logaritmos das razões de expressão entre tratamentos (expressão no tratamento 1 / expressão no tratamento 2) apresentam, em geral, distribuição normal (Friddle *et al.*, 2000).

Experimentos com arranjos de DNA freqüentemente têm como objetivo a identificação de genes diferencialmente expressos entre diversos tratamentos ou tecidos (Schena *et al.*, 1996). A seleção desses genes, porém, é muito dificultada pelo fato do número de réplicas nesses experimentos ser em geral reduzido, enquanto o número de genes analisado em paralelo chega a milhares (Lee *et al.*,

2000). Essa desproporção torna a maioria dos testes estatísticos tradicionais inadequados aos dados de arranjos, já que estes poderiam levar a um grande número de genes erroneamente classificados, e diversos estudos vêm sendo realizados procurando desenvolver métodos estatísticos mais adaptados às características desses dados (Kerr *et al.*, 2000, Tusher *et al.*, 2001, Efron *et al.*, 2001, Mutch *et al.*, 2002, Loguinov *et al.*, 2004, Newton *et al.*, 2004).

Outra estratégia de análise freqüentemente adotada em dados de arranjos de DNA é o agrupamento dos genes de acordo com seus padrões de expressão em diferentes tratamentos ou tecidos, procurando assim identificar grupos de genes expressos conjuntamente (Kerr e Churchill, 2001, Datta e Datta, 2003). Algoritmos hierárquicos (Chu *et al.*, 1998, Eisen *et al.*, 1998) e “*Self Organizing Maps*” (Tamayo *et al.*, 1999) são os mais freqüentemente utilizados nesse tipo de análise.

A análise de dados de arranjos de DNA constitui um campo de estudo em rápido desenvolvimento. Um grande número de métodos para a análise desses dados vem sendo desenvolvido, e apenas em poucos aspectos dessa análise existe algum consenso entre a comunidade científica (Allison *et al.*, 2005). A identificação de genes diferencialmente expressos a partir de dados de arranjos de DNA tem sido aplicada em diversos ramos da genética. Na área de genética vegetal, a identificação de genes envolvidos em respostas a estresses bióticos e abióticos ou relacionados a características interessantes do ponto de vista econômico pode ser de grande ajuda no melhoramento de espécies de interesse agrônômico. Como exemplo, podemos citar a identificação de genes de arroz diferencialmente expressos em resposta a deficiência de fosfato (Wasaki *et al.*,

2003) e a seleção de genes envolvidos na resposta a baixas temperaturas em cana-de-açúcar (Nogueira *et al.*, 2003).

A cana-de-açúcar

O cultivo da cana-de-açúcar é uma atividade agrícola muito importante mundialmente, sendo praticado em mais de 90 países, em áreas tropicais e subtropicais, e contribuindo com 60% do açúcar produzido no mundo (Grivet e Arruda, 2001). O Brasil é o maior produtor mundial de cana-de-açúcar, seguido pela Índia e Austrália. Na safra 2003-2004 foram cultivadas aproximadamente 350,3 milhões de toneladas de cana-de-açúcar no Brasil, dos quais 85% são provenientes da região Centro-Sul do país. A maior parte deste montante (55%) é utilizada na produção de etanol, enquanto 45% são destinados à produção de açúcar. No presente, cerca de 4,5 milhões de hectares do território brasileiro são ocupados pelo cultivo de cana-de-açúcar (Pessoa-Jr *et al.*, 2005). De acordo com o Registro Administrativo do Ministério do Trabalho e do Emprego, em 2002 os setores produtivos de cana-de-açúcar, açúcar e álcool geravam, conjuntamente, mais de 760 mil empregos formais (Moraes, 2005).

A cana-de-açúcar pertence à família Poacea, assim como o arroz, e à tribo Andropogoneae, como o milho e o sorgo. Suas cultivares modernas são frutos de cruzamentos entre *Saccharum officinarum*, uma espécie domesticada a partir da espécie selvagem *S. robustum* e que apresenta alto teor de açúcar, e *S. spontaneum*, uma espécie selvagem vigorosa. Devido a essa origem múltipla, o genoma da cana-de-açúcar é considerado um dos mais complexos dentre as espécies de importância agrícola, apresentando alto grau de ploidia. Clones de *S.*

officinarum usualmente apresentam $2n=80$ cromossomos, enquanto para *S. spontaneum* vários citotipos já foram descritos, com $2n$ variando de 40 a 128. As cultivares modernas de cana-de-açúcar apresentam entre $2n=100$ e $2n=130$ cromossomos, indicando que a aneuploidia é comum (Grivet e Arruda, 2001). Essa complexidade do genoma da cana-de-açúcar dificulta bastante a aplicação de técnicas convencionais de genética e melhoramento a essa espécie.

Dada a relevância econômica da cana-de-açúcar, torna-se importante a aplicação de novas tecnologias no estudo genético da espécie, visando com isso o melhoramento genético de suas cultivares. Projetos voltados para o estudo e seqüenciamento de genes dessa espécie se encontram em andamento na Austrália, África do Sul e Estados Unidos. No Brasil, uma iniciativa conjunta da Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp) e da Cooperativa dos Produtores de Cana, Açúcar e Alcool do Estado de São Paulo (Copersucar) deu origem ao SUCEST (**S**ugarcane **EST** Project), cujo objetivo final foi a identificação de cerca de 30.000 unigenes de cana-de-açúcar. As prováveis funções de muitos destes genes foram determinadas pelos grupos de *Data Mining* envolvidos no SUCEST, porém a função de muitos mais só poderá ser determinada através de estudos mais aprofundados, como por exemplo a quantificação de sua expressão em diferentes tratamentos.

Após a identificação de milhares de genes de cana-de-açúcar pelo projeto SUCEST, um consórcio de grupos de pesquisa iniciou o projeto SUCEST-FUN, cujo objetivo era a caracterização funcional de grande parte desses genes. Em uma primeira etapa, o transcrito da cana-de-açúcar vem sendo estudado através de arranjos de DNA e outros experimentos, procurando avaliar a

expressão dos genes estudados em diferentes tecidos e em plantas submetidas a diversas condições, como estresses bióticos e abióticos e tratamentos com hormônios. Os financiadores do projeto são a Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp), o Centro de Tecnologia Canavieira (CTC) e a Central de Álcool Lucélia.

A deficiência de fosfato

O fósforo é um macronutriente essencial a todos os organismos vivos e constitui cerca de 0,2% do peso seco das plantas (Schachtman *et al.*, 1998). Ele exerce várias funções biológicas básicas: é elemento estrutural de moléculas indispensáveis como os ácidos nucleicos, fosfolípidios e ATP, atua no metabolismo de energia e na ativação de intermediários metabólicos e colabora na transdução de sinais e na regulação de enzimas (Daram *et al.*, 1999).

A forma do fósforo mais facilmente captada pelas plantas é o fosfato inorgânico (Pi). Entretanto, devido a sua baixa mobilidade, existência em forma orgânica e a interações com outros constituintes do solo como Al, Fe e Ca, o fosfato é um dos elementos menos disponíveis na rizosfera. A concentração de Pi na solução do solo é tipicamente próxima a 2 μ M, enquanto as plantas contêm entre 5 e 20 mM de Pi (Bielecki, 1973). A deficiência de fosfato é freqüente na natureza, especialmente nos solos ácidos das regiões tropicais e sub-tropicais, onde se torna um fator limitante da produção agrícola (Raghotama, 2000). O cerrado brasileiro constitui 25% do território nacional (aproximadamente 205 milhões de hectares) e é um exemplo de região pobre em fosfato, onde a

agricultura depende do uso de fertilizantes enriquecidos com o nutriente e correção do pH do solo (Santana e Sans, 1999).

No decorrer da evolução as plantas desenvolveram vários mecanismos adaptativos à deficiência de fosfato. Alterações no crescimento e na arquitetura radicular, bem como o aumento na produção de fosfatases, RNases e transportadores de fosfato já foram comprovados como parte desses mecanismos (Raghotama, 1999). Diversas evidências têm indicado que muitas das alterações bioquímicas, fisiológicas e morfológicas que ocorrem como resposta à deficiência são associadas a mudanças na expressão gênica (Hammond *et al.*, 2004). Tais mudanças surgem como uma resposta direta e específica à deficiência de fosfato, e alguns dos genes induzidos estão diretamente envolvidos no aumento da disponibilidade de P_i e em promover sua maior absorção.

A avaliação dos dados disponíveis indica que a deficiência de fosfato deve levar à expressão coordenada de genes, similar ao *pho* regulon em *Saccharomyces cerevisiae*, no qual vários genes são induzidos ou reprimidos pela disponibilidade de P_i (Yoshida *et al.*, 1989). Dois genes para RNases associadas à senescência (*RNS1* e *RNS2*) são induzidos por esta deficiência em *Arabidopsis* (Bariola *et al.*, 1999). Na alga verde *Chlamydomonas reinhardtii* constatou-se o acúmulo de diversos mRNAs correspondentes a proteínas ribossomais e enzimas envolvidas no catabolismo da glicose durante a deficiência de fosfato (Dumont *et al.*, 1993).

Experimentos com culturas de células de tabaco (*Nicotiana tabacum*) privadas de fosfato levaram ao isolamento de quatro cDNAs induzidos, sendo dois deles homólogos aos genes *parA* e *parB* de *N. tabacum* (Ezaki *et al.*, 1995). Em

Brassica nigra foram constatadas, como respostas à deficiência de fosfato, a síntese *de novo* de fosfatases ácidas (Duff *et al.*, 1991), bem como a indução de uma fosfofrutoquinase dependente de P_i (Theodorou e Plaxton, 1994). Em tomate (*Lycopersicon esculentum* L.), observou-se o aumento da expressão de Ca²⁺-ATPase (Muchhal *et al.*, 1995) e maior atividade de fosfatase ácida, fitase e ribonuclease, enzimas mobilizadoras de fosfato (Bosse e Köck, 1998). Nessa mesma espécie detectou-se a rápida indução de um gene (*TPS11*), que provavelmente faz parte de uma resposta primária (*early*) à deficiência (Liu *et al.*, 1997). Um estudo envolvendo *Arabidopsis thaliana* e tomate encontrou evidências de que genes induzidos pela deficiência de fosfato estão submetidos a uma regulação negativa, permanecendo reprimidos na presença do nutriente (Mukatira *et al.*, 2001). Trabalhando com tremçoço branco (*Lupinus albus*), uma espécie conhecida por sua tolerância à baixa disponibilidade de P, Uhde-Stone *et al.* (2003) identificaram 35 genes induzidos durante a formação de raízes proteóides promovida como resposta a deficiência de P, incluindo genes envolvidos na aquisição e reciclagem de P_i (fosfatases ácidas) e na síntese de ácidos orgânicos (malato e citrato). Wu *et al.* (2003) e Hamond *et al.* (2003) realizaram estudos com *Arabidopsis thaliana*, utilizando microarranjos para avaliar alterações na expressão gênica, em escala genômica, causadas pela privação de fosfato. O primeiro grupo detectou 1.835 genes cuja expressão responde a essa deficiência nutricional, sendo 1398 em folhas e 730 em raízes, enquanto o segundo identificou a expressão diferencial de 119 genes em folhas (essa discrepância no número de genes pode ser devida a diferenças entre os métodos experimentais e de análise dos dados). Analisando o transcrito de raízes de arroz (*Oryza sativa*), Wasaki

et al. (2003) identificou ainda 208 genes diferencialmente expressos em resposta a essa deficiência nutricional. Esses resultados indicam a presença, em diferentes plantas, de uma resposta complexa à deficiência de fosfato, envolvendo a ativação de vários genes em diferentes estágios da deficiência (resposta primária, intermediária e tardia). Em cana-de-açúcar, até onde sabemos, o único estudo sobre a expressão gênica em resposta a essa deficiência foi realizado por Almeida (2002, tese de mestrado) que identificou e caracterizou transportadores de fosfato expressos preferencialmente no sistema radicular, sob deficiência de fósforo. Porém, estudos da coordenação da resposta da cana-de-açúcar à deficiência de fosfato ainda não foram publicados.

Estrutura do trabalho

A estrutura desse trabalho inclui artigos publicados por nosso grupo de pesquisa em revistas científicas ou submetidos para publicação nas mesmas. Os artigos encontram-se na forma em que foram publicados ou submetidos, redigidos em inglês.

O trabalho encontra-se dividido em oito capítulos. O Capítulo 1 é dedicado à padronização do uso de macroarranjos em membranas de nylon, uma técnica de análise da expressão gênica em larga escala robusta e acessível à maioria dos laboratórios de biologia molecular. Esse capítulo é constituído por dois artigos submetidos para publicação, sendo um deles voltado para a padronização da confecção dos arranjos e do uso de hibridação interespecífica na análise da expressão gênica e o outro voltado para a análise da interferência do plástico utilizado para selar as membranas nos resultados obtidos.

O Capítulo 2 é voltado ao pré-processamento aplicado aos dados de arranjos de DNA, antes de qualquer análise estatística. O capítulo contém uma revisão das principais estratégias usualmente aplicadas de subtração do background (emissão de fundo dos arranjos), filtragem dos dados, normalização e transformação dos mesmos.

O Capítulo 3 trata da aplicação de testes estatísticos clássicos (teste t de Student e análise de variância) aos dados de arranjos de DNA. Para exemplificar a aplicação desses testes na análise de dados provenientes de experimentos reais, três artigos publicados por nosso grupo de pesquisa foram anexados ao capítulo.

O Capítulo 4 revisa algumas estratégias de análise específicas para dados de arranjos de DNA já descritas na literatura. A aplicação de um algoritmo de análise largamente utilizado na literatura é demonstrada em um artigo anexado ao capítulo. Esse artigo foi voltado à identificação de genes de milho relacionados à tolerância a toxidez do alumínio e encontra-se submetido para publicação.

Os Capítulos 5 e 6 apresentam a parte principal deste trabalho: os algoritmos de análise de dados de arranjos de DNA desenvolvidos. O Capítulo 5 é constituído por um artigo publicado e por seus materiais suplementares, no qual é descrito um algoritmo voltado para a análise de dados de experimentos sem réplicas. O Capítulo 6 apresenta o segundo algoritmo de análise de dados desenvolvido nesse trabalho, voltado para experimentos que apresentam poucas réplicas biológicas (ainda que tenham diversas réplicas técnicas).

O Capítulo 7 descreve o experimento realizado para avaliar a resposta da cana-de-açúcar à deficiência de fósforo, bem como a aplicação dos algoritmos desenvolvidos aos dados de arranjos de DNA obtidos. Os genes identificados

como diferencialmente expressos em resposta a essa deficiência nutricional são discutidos, procurando assim caracterizar a resposta da cana-de-açúcar a esse estresse.

Por fim, o Capítulo 8 trata da aplicação de algoritmos de agrupamento aos dados de expressão dos genes selecionados como diferencialmente expressos no experimento descrito no Capítulo 7. O capítulo apresenta uma breve introdução aos algoritmos de agrupamento e os resultados obtidos com a aplicação de dois desses algoritmos aos dados do experimento de fosfato. Procurou-se com isso caracterizar melhor os diversos estágios da resposta da cana-de-açúcar à deficiência de fosfato.

Objetivos

O objetivo geral deste trabalho foi o desenvolvimento de algoritmos para a análise de dados de arranjos de DNA, que procuram identificar os genes cuja expressão diferencial é mais significativa. Além disso, tivemos também como objetivos:

- Aplicar os algoritmos desenvolvidos para identificar genes de cana-de-açúcar diferencialmente expressos em resposta a deficiência de fosfato, através da análise de dados resultantes da hibridação de microarranjos de DNA, contendo ESTs de cana-de-açúcar, com sondas provenientes de raízes de plantas desta espécie submetidas a diferentes regimes de fosfato.

- Empregar algoritmos de agrupamento na análise dos dados de expressão dos genes selecionados como diferencialmente expressos, buscando assim caracterizar os diferentes estágios da resposta da cana-de-açúcar a essa deficiência nutricional.

Capítulo 1: Padronização do uso de macroarranjos em membranas de nylon

Arranjos de DNA em lâminas de vidro (microarrays) são a tecnologia mais utilizada em estudos de expressão gênica em larga escala. Porém a implementação desta tecnologia requer investimentos vultosos e freqüentemente longos períodos de padronização. Uma alternativa mais acessível à maioria dos laboratórios de biologia molecular são os arranjos de DNA em membranas de náilon, também chamados de nylon arrays, filter arrays ou macroarrays. Apesar deste tipo de arranjos acomodar um número menor de genes, sua utilização requer metodologias rotineiras, robustas e de fácil implementação (Felix *et al.*, 2002), baseando-se na hibridação de sondas marcadas radiativamente a filtros de náilon, um procedimento usual na maioria dos laboratórios de biologia molecular.

Nosso grupo de pesquisa realizou diversos testes procurando padronizar e otimizar o uso de arranjos de DNA em membranas de náilon. Esses testes foram reunidos em dois artigos, anexados a seguir:

Cross-species hybridization of nylon arrays containing sugarcane ESTs to identify AI-induced genes from maize

Felix J.M., Drummond R.D., Maron L.G., Nogueira F.T.S., Jorge R.A., Arruda P. & Menossi M.

Artigo em vias de submissão à revista *Brazilian Journal of Medical and Biological Research*.

***Evaluation of the performance of different plastics used to seal nylon
cDNA arrays.***

Costa-Netto A.P., Drummond R.D., Felix J.M., Jorge R.A. & Menossi M.

Artigo submetido à revista *Ciência e Agrotecnologia*.

RUNNING TITLE: Cross-species hybridization of nylon arrays

TITLE: Cross-species hybridization of nylon arrays containing sugarcane ESTs to identify Al-induced genes from maize.

KEYWORDS: nylon filters; macroarray; aluminum; stress; heterologous hybridization

Juliana de Maria Felix ^{a, b, §}, Rodrigo Drummond ^{a, b, §}, Fábio Tebaldi Silveira Nogueira ^{a, b}, Renato A. Jorge ^c, Paulo Arruda ^{a, b}, Marcelo Menossi ^{a, b, *}.

^a*Centro de Biologia Molecular e Engenharia Genética, State University of Campinas (Unicamp), CP 6010, 13083-970, Campinas, SP, Brazil,*

^b*Departamento de Genética e Evolução, Instituto de Biologia, State University of Campinas (Unicamp), Campinas, SP, Brazil,*

^c*Departamento de Físico-Química, Instituto de Química, State University of Campinas (Unicamp) Campinas, SP, Brazil*

[§] Both authors contributed equally to this work

Abbreviations: EST, expressed sequence tag; GFP, green fluorescence protein; IP, Image Plate

*Address for Correspondence:

Centro de Biologia Molecular e Engenharia genética, State University of Campinas (Unicamp), CP 6010, 13083-970, Campinas, SP, Brazil

Telephone number: +55 (19) 35211143

Fax number: +55 (19) 35211089

E-mail menossi@unicamp.br

Abstract

The hybridization of DNA arrays is a powerful methodology to perform large-scale gene expression studies. The high degree of gene conservation observed between related species such as grasses led us to the hypothesis that arrays containing sugarcane genes could be used to probe cDNA samples from maize plants. However, the accuracy of gene expression profiling with DNA arrays is sensitive to internal and external sources of fluctuations that had to be analyzed before evaluating the performance of cross-species array hybridization. We found that the hybridization with an oligonucleotide target that recognizes all the sequences on the array is an efficient strategy to estimate the fluctuations on the amounts of probe spotted, and that transferring the DNA with two pin strikes reduced by half the variability of the DNA amount fixed on the membrane. Finally, we have identified and validated by RNA blot a gene encoding a major intrinsic protein that is induced by AI in the maize root apex, demonstrating that cross-species hybridization is a powerful strategy for gene expression profiling.

Introduction

The hybridization of DNA arrays has allowed the analysis of gene expression in a genomic scale, and the changes in the expression of thousands of genes under different conditions can be monitored in parallel. The general approach is to immobilize gene-specific sequences (probes) onto solid-state matrix (nylon filters, glass microscope slides, silicon/ceramic chips). These sequences are then queried with labeled copies of nucleic acids from biological samples (targets). The cDNA arrays that rely on denaturated double-stranded DNA fragments deposited on nylon filters are usually called macroarrays or high-density cDNA arrays. Macroarrays are unique among hybridization arrays in that probes can be manually spotted, and they use radioactive target labeling (microarrays are robotically spotted, and generally use fluorescent dye-labeling detection). After radioactively labeling the target, different samples are hybridized to individual replicate arrays. Imaging plates (or less frequently X-ray film) are then used to detect the bound target. In this sense, the macroarray technology is relatively cheap because it does not need high-cost equipments and the filters can be rehybridized up to ten times (Kuhn 2001). Moreover, it is a straightforward approach, since it is based in simple and very well established methodologies.

Experiments involving cDNA arrays are generally sensitive to external and internal fluctuations, such as target radioactivity used to hybridize with arrays, probe amount variation intra and inter replicates filters and random fluctuations in hybridization efficiency. Since the reliability of interaction patterns extracted from array data is essential for their interpretation, a reduction in these fluctuations by proper averaging and normalization procedures is of great practical interest.

The major limitation of a large scale gene expression profiling is the availability of cDNA sequences to produce the filters. Due to the high degree of conservation between gene sequences from close-related species such as maize and sugarcane, we decided to test if macroarrays containing sugarcane ESTs from the SUCEST project (<http://sucest.lad.dcc.unicamp.br>) could be used to identify expression profiles from the maize transcriptome. If this hypothesis hold true, more than 40,000 sugarcane genes identified by the project could be a valuable resource to study gene expression in many other monocots species with agronomical importance, such as rice, barley and wheat.

As a working model to test this hypothesis we used maize (*Zea mays* L.) plants that were exposed to Aluminum, the most abundant metal in the earth's crust, comprising about 7% of its mass (Delhaize and Ryan 1995). When solubilized in acid soils, Al is toxic to many crops and is the major limiting factor for plant production. Acid soils occupy 3.95 billion ha (30%) of the world's ice-free land area, and around 66% of land surface in South America (Baligar and Ahlrichs 1998), thus the elucidation of the genetic and physiologic basis of the response to Al-stress would be of great interest to agriculture around the world.

In the current study we evaluated several aspects that affect the results from macroarrays, such as the amount of target on the spots, the linearity between the detected signal and the amount of transcript and the sensitivity of the method. We have also validated the use of cross-species hybridization as a valuable tool to discover gene expression profiles, isolating the gene encoding a major intrinsic protein (MIP), which was induced by Al in maize. The putative role of this gene in Al tolerance is discussed.

Material and methods

Plant material

The maize inbred line Cat 100-6, (Al-tolerant) was obtained from the Germplasm Bank of the Universidade Estadual de Campinas, Campinas, Brazil. Plants were grown in the field and self-pollinated.

The commercial sugarcane variety used in the SUCEST project (SP80-3280) were field-grown at the Cane Technology Center (CTC), Piracicaba, São Paulo, Brazil.

Maize plant growth

Seeds were surface-sterilized with 70% ethanol for 1 min, 0.5% (v/v) sodium hypochlorite for 20 min, rinsed four times with sterile water and germinated between layers of moist filter paper at 28 °C for 60 h. After 3 days, seedlings were transferred to polystyrene holders that were floated on an aerated nutrient solution (Moon *et al.*, 1997). When required, Aluminum ($KAl(SO_4)_3 \cdot 12H_2O$) was added to the nutrient solution to the appropriate activity, estimated by the GEOCHEM-PC program (Parker, 1995). The pH of the solution was maintained at 4.0, and the plants were grown under a photoperiod of 16 h.

RNA isolation

Total RNA from maize root apices was extracted according to Logemann *et al.* (1987). For sugarcane roots, the total RNA was isolated using Trizol® according to the recommendations of the manufacturer (Invitrogen, Brazil). The RNA samples

were quantified in a spectrophotometer and loaded onto 1.0% agarose/formaldehyde gels for a quality inspection.

Preparation of cDNA macroarrays

The plasmids containing sugarcane ESTs were obtained from the SUCEST RT1 cDNA library. This cDNA library was generated from sugarcane roots (Vettore *et al.*, 2001). Plasmidial DNA was denatured with 0.2 M NaOH and spotted on nylon filters (Hybond-N, AP-Biotech, USA) with a hand-held 96-pin printhead tool which deposited typically 0.1 μ l (10 ng) of DNA solution (V&P-Scientific, USA), as described by Schummer *et al.* (1997). DNA was fixed to the filters by baking at 80 °C for two hours. Additionally, in the normalization experiments we also spotted a plasmid containing a cDNA encoding the green fluorescence protein (GFP), kindly provided by Dr. M. Schummer (University of Washington, Seattle, USA).

Target labeling and hybridization protocols

The amount of probe on the spots was estimated with an overgo probe that recognizes the sequence of the *Amp^r* gene of the vector used in the SUCEST cDNA libraries (pSPORT1, Invitrogen, USA). The target was synthesized using the primers 5'-GTGGTCCTGCAACTTTATCCGC-3' and 5'-TAGACTGGATGGAGGCGGATAA-3' in the presence of α -³³P-dCTP and hybridized to the nylon arrays for 16 h at 58°C, according to the protocol from J.D. McPherson (<http://www.tree.caltech.edu/protocols/overgo.html>). After the detection

of the radioactivity (see below) the filters were stripped twice with boiling 0.5% SDS and stored.

The relationship between the signal intensity and the transcript abundance in a complex cDNA target preparation was performed with a polyadenylated RNA from the *GFP* gene. A PCR was done with the *mGFP4* gene (Haseloff *et al* 1997) using 10 μ M of the primers 5'-ACAATGAAAGGAGAAGAAGCTT-3' and 5' T₂₂ATTTGTATAGTTCATCCATGC-3'. The PCR product was cloned in the pGemT-Easy (Clontech, USA). The polyadenylated GFP transcript was synthesized from 300 ng/ μ L of the linearized DNA using the SP6 RNA polymerase (Langdale 1994). The RNA was quantified using a spectrophotometer and mixed at different masses (0.0028, 0.014, 0.028, 0.056, 0.28, 1.4, 2.8 and 8.4 ng) with 30 μ g of total RNA from sugarcane roots to simulate a regular labeling reaction. The synthesis of the α -³³P-dCTP labeled cDNA target, the hybridization and washing of the filters were performed as described by Schummer *et al* (1999), except that our targets were purified with G-50-columns (AP-Biotech, USA) before the hybridization.

The heterologous hybridization was done with cDNA target synthesized from total RNA from maize roots of seedlings exposed to several AI activities (0, 5, 15, 50 and 85 x 10⁻⁶). The target synthesis and hybridization were done as described above for the sugarcane and GFP RNA mixes. The targets were independently hybridized with replicate filters containing 768 random SUCEST ESTs in duplicate, totalizing 1536 spots/filter.

Signal intensity detection

After the hybridization, the arrays were exposed to image plates from the FLA3000-G screen system (Fuji Photo Film, Japan). The radioactive intensity of each spot from the digitalized images was converted to relative intensity value and quantified using the *Arrayvision*TM software (Imaging Research, Canada). The local background was subtracted from relative intensity of each spot and the intensity data were rearranged into MS Excel (Microsoft, USA) files to further analysis.

Northern Blot Analysis

Ten micrograms of total RNA from maize root apex was electrophoresed in formaldehyde-containing agarose gel and transferred to nylon filter (Hybond-N⁺, AP-Biotech, USA). A SUCEST ESTs and the maize cDNAs were labeled with α -³²P-dCTP and hybridized to the filters according to Sambrook (1989) using high stringency conditions. Digitized images of RNA blot hybridization signals were detected with the FLA3000-G screen system (Fuji Photo Film, Japan) and quantified with the *Image Gauge* software (Fuji Photo Film, Japan).

Results and discussion

Random fluctuations in the target volume

The intensity of the signal obtained from the radioactive emission of a spot depends not only on the amount of the transcript present in the target sample, but also on the amount of probe spotted on the array (Perret, 1998). In this sense, fluctuations on the amount of DNA on the spots can lead to errors in the analyses of gene expression. Perret (1998) suggested that these variations can be estimated by hybridizing the filters with a target that binds to the plasmid used to amplify the genes, common to all sequences on the arrays.

To evaluate the efficiency of this strategy in our homemade arrays we spotted serial dilutions (100-3.125 ng/ μ l DNA) of a plasmidial DNA containing the GFP gene (Haseloff *et al* 1997). The arrays were hybridized with an overgo probe, which recognizes the *Amp^r* gene of the GFP construct and also of the pSPORT vector, used in the libraries construction of the SUCEST project. Due to the fact that different spots compete for the same target, the amount of target bound in each spot is supposed to be proportional to the amount of probe cDNA in the spot. The observed signal intensity showed linearity with the amount of DNA spotted on the arrays (Fig. 1), indicating that the overgo probe can estimate fluctuations on the amounts of cDNA fixed on our nylon arrays. The same experiment was repeated with other four filters, giving similar results (data not shown).

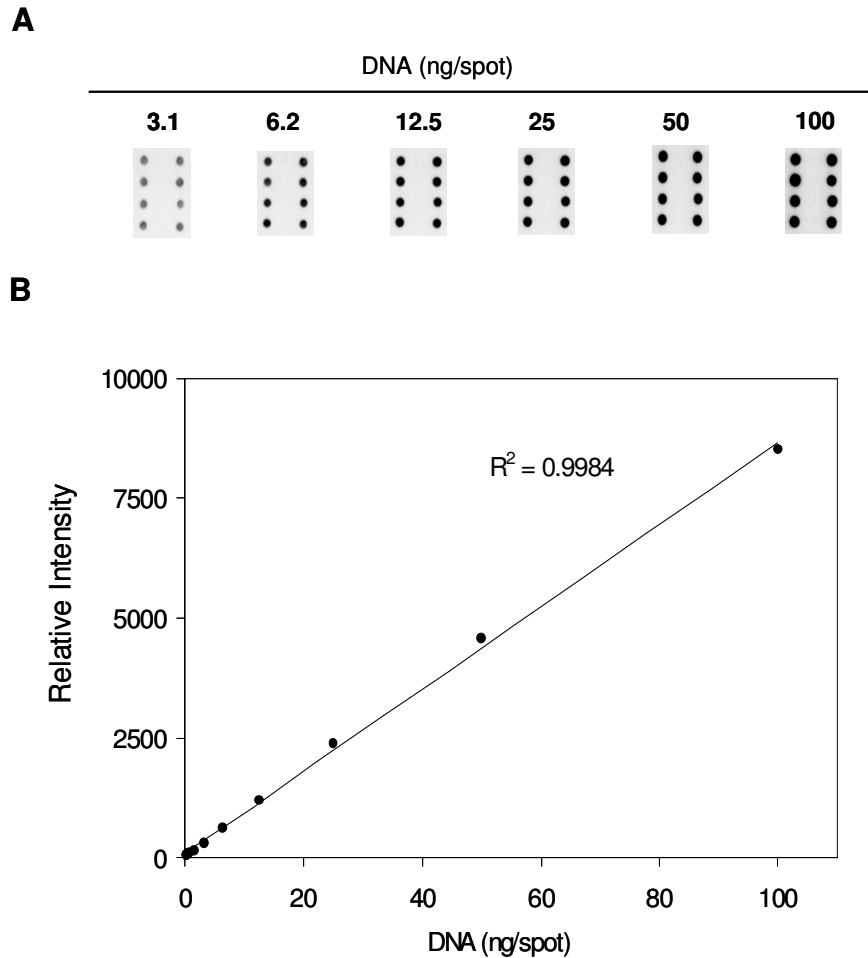


Figure 1: Linearity between the hybridization signal from the overgo probe and the amount of DNA on the filters. Six serial dilutions starting from 100 ng/ μ l of a plasmidial DNA containing a *Amp^r* gene was spotted on nylon arrays (eight spots/dilution) with a handheld device that takes 0.1 μ l of the DNA sample. The arrays were hybridized with ^{33} P-labeled overgo probe. **(A)** Close-ups of the filter regions containing the spots after a 24 h exposition to imaging plates. The amount of DNA on the spots is indicated over the pictures. **(B)** The signals from the images shown in (A) were quantified with a phosphorimager and their means were plotted versus the amount of DNA.

A possible way to reduce these fluctuations is to use multiple pin strikes to transfer the DNA to the filters (L. Reis, Ludwig Institute, personal communication). A pin strike is defined as a single transfer of DNA from the source plate to a spot

on the filter surface (English *et al* 1999). When the DNA is transferred by a single pin strike, any fluctuation on the amount of DNA transferred will be reflected on the final amount of DNA fixed on the spot, while the use of multiple strikes makes it less probable that the same imperfection occurs more than once on the same spot.

To verify the efficiency of this strategy twelve different ESTs were fixed on an array. Twelve spots containing 10 ng of cDNA probe represented each EST. On four of these spots the cDNA were transferred by a single pin strike (from a 100 ng/ μ l DNA stock); on other four spots the cDNA were transferred by two pin strikes (from a 1:1 DNA dilution) and on the remaining four spots, three strikes were used (from a 1:2 dilution). The filter was then hybridized with the overgo probe and the signal was quantified as above. The coefficient of variation (CV, the percentage of the mean signal between replicate spots represented by the standard deviation of these signals) for the three groups of spots for each EST was calculated, and its mean for each spotting strategy is shown in Figure 2. The mean CV using one pin strike was reduced from 31% to 15% when using two pin strikes and the use of three pin strikes caused only a minor further reduction (CV = 13%). Based on these results we decided to transfer DNA with two pin strikes. We evaluated five replica filters containing 768 ESTs in duplicate and we found that in 98% of the spots the ratios of the overgo signal between replica spots lies in the range 0.5-2.0, i.e., most ESTs did not show a variation of the DNA amount between replicate spots above the threshold of two-fold (data not shown). This result is in agreement with Schummer *et al.* (1997).

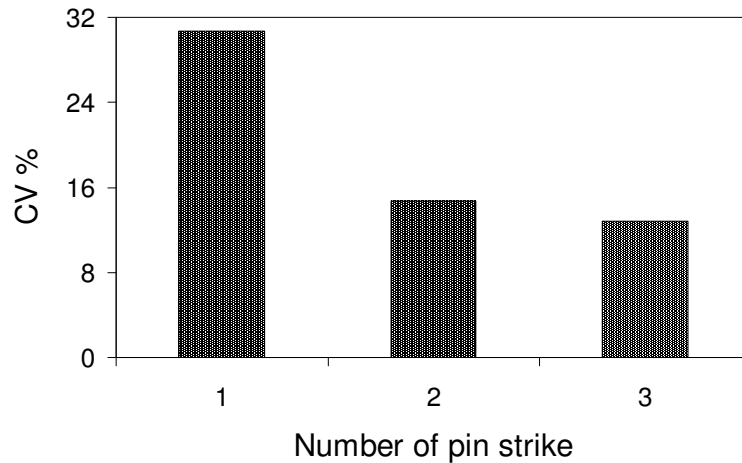


Figure 2: Effect of multiple pin strikes on the variation in the amount of DNA transferred to the filter. Nylon arrays with spots containing 10 ng of plasmidial DNA from twelve sugarcane ESTs transferred using one, two and three pin strikes were hybridized with ^{33}P -labeled overgo probe. The mean of the coefficients of variation of the spots for each number of pin strikes for the twelve genes is shown.

Normalization of fluctuations of the target signal

The efficiency of the hybridization reactions is influenced by a number of experimental parameters, notably temperature, time, buffering conditions and the overall amount of target molecules used for hybridization (Schuchhardt *et al.*, 2000). In order to compare hybridization signatures of two identical filters that have been hybridized with different targets it is needed to normalize the signals to a standard. When working with targets that came from different treatments it is usually expected that most of the genes will not change their expression, so it is expected that the median of the signals from all genes in each filter will be conserved between filters, unless there is some fluctuation on the efficiency of hybridization (Quackenbush, 2002). These considerations lead to the normalization strategy of dividing all the values in each filter by its median.

To check the efficiency of this strategy in our arrays five replicate filters containing 12 ESTs (8 spots/EST) were hybridized individually with five complex cDNA targets obtained from the same RNA sample extracted from sugarcane roots. The signals from three representative ESTs before and after the normalization with the median are shown in the Figure 3A and 3B, respectively. The CV between the five hybridizations is around 40%, mainly due to the fourth hybridization, that gave a weak signal (Fig. 3C). After the normalization with the median, the CV dropped in all cases to around 15%, confirming the usefulness of this strategy in the macroarrays.

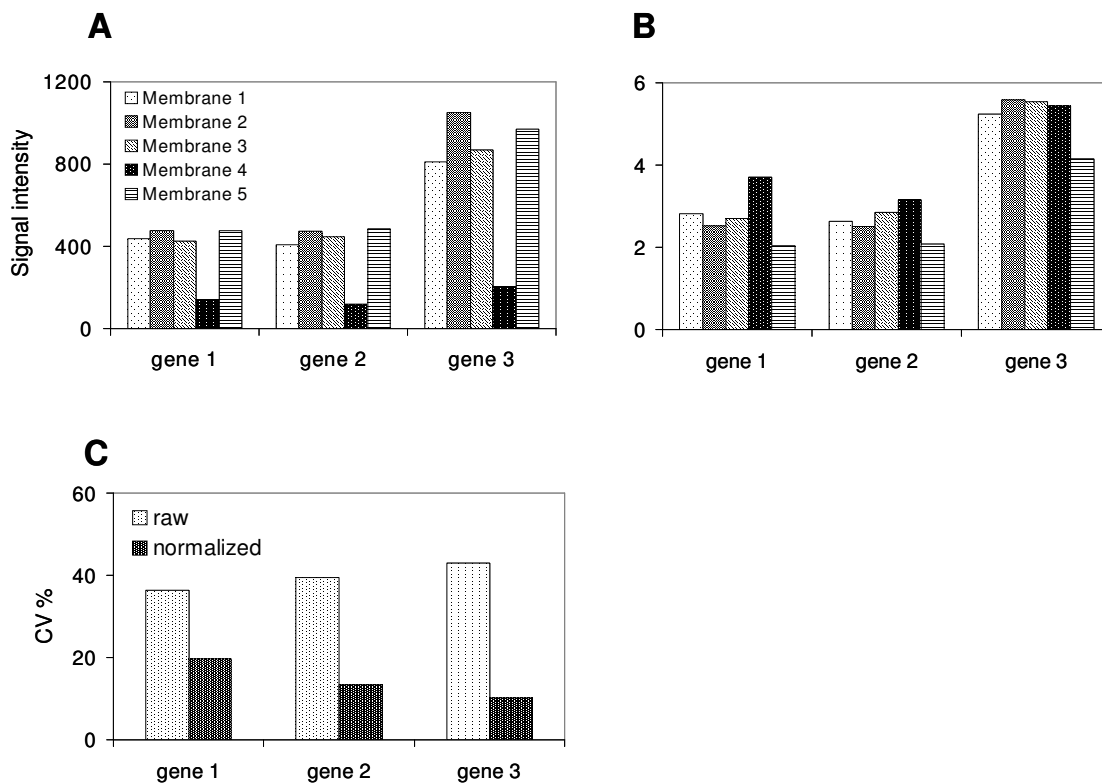


Figure 3: Normalization using the median of genes to correct for variation between replicates experiments. **(A)** Raw signal obtained for three genes in five replicate filters; **(B)** The signals in (A) were normalized using the median for all genes on each filter; **(C)** Coefficients of variation of the signals from the three genes obtained from different filters before (shown in A) and after the normalization (shown in B).

Sensitivity and relationship between signal intensity and the abundance of RNA transcripts

Other relevant aspects in array analysis are the detection threshold and the relationship between the detected signal and the abundance of the related transcript. To evaluate these parameters we hybridized arrays with targets obtained from 30 µg of total RNA from sugarcane mixed with several masses of a polyadenilated RNA of the GFP gene (0.0028, 0.014, 0.028, 0.056, 0.28, 1.4, 2.8 and 8.4 ng) simulating rare (~10 copies/cell), intermediate (~300 copies/cell) and abundant (~12000 copies/cell) transcripts, according to the estimations from Huang *et al.* (1999).

Since there is no homologue to GFP gene in plants, the signals obtained on this experiment are due to the amount of GFP transcripts that were added to each aliquot of total sugarcane RNA before the synthesis of cDNA. The targets generated from these samples were hybridized with the filters containing 10 ng/spot of the GFP plasmid. As can be seen in Figure 4, it was possible to detect a signal above background even when the target contains 0.0028 ng of GFP transcript, corresponding to rare genes, indicating that the method has a good detection limit, similar to the observed by other authors working with nylon arrays (Pietu *et al* 1996; Schummer *et al* 1999).

We also evaluated the linearity and the sensitivity on spots containing lower amounts of the GFP plasmid and it is worth to emphasize that this level of sensitivity hold true only with spots with at least 2.5 ng of DNA (results not shown).

Based on these results we decided to adopt 10 ng of DNA as the standard amount to be transferred onto each spot.

Our results also showed that the relation between the detected signal and the amount of GFP transcript had a good linearity ($R^2=0.95$) when both values are expressed in a logarithmic scale (Fig. 4), indicating the log values are the best choice to compare the expression between experiments.

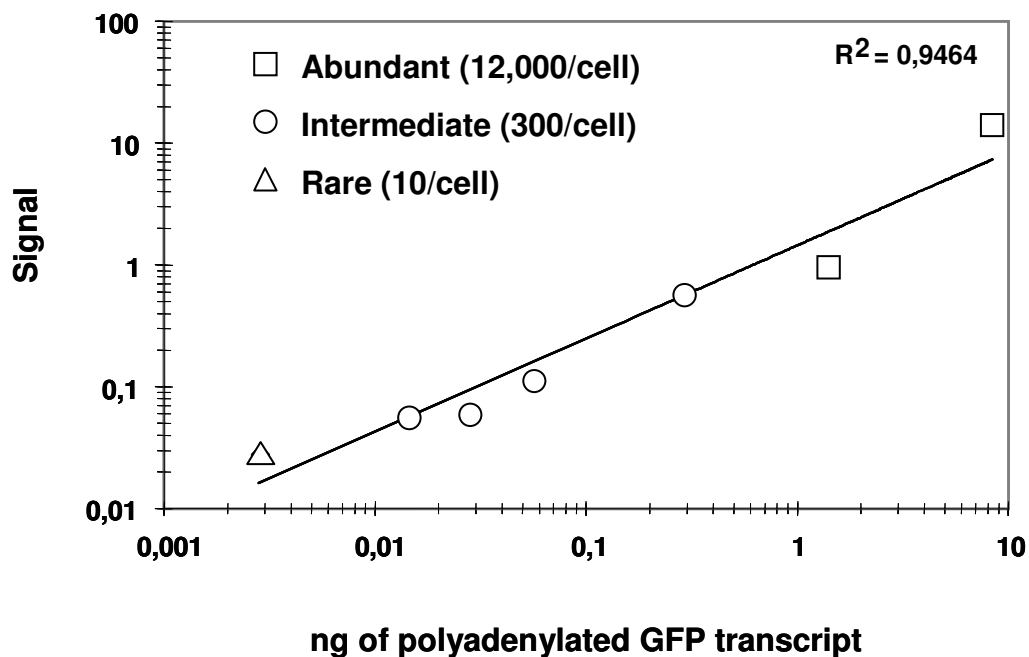


Figure 4: Evaluation of the detection limit and the relationship between the signal and mRNA abundance in the RNA pool. Different amounts (0.0028 to 8 ng) of a synthetic polyadenylated RNA of the GFP gene were mixed with 30 μ g of total sugarcane RNA, reverse transcribed in the presence of $\alpha^{33}\text{P}$ -dCTP, and hybridized to nylon arrays containing 10 ng of GFP DNA in each spot. The signal was quantified after a 96 h exposition (data represent the mean of the signal intensity). Symbols indicated GFP transcript amounts equivalent to rare (Δ), intermediate ($\circ$) and abundant ($\square$) genes.

Evaluation of fluctuations in the amount of DNA fixed on the spots.

According to Perret *et al.* (1998) the signal from any particular spot hybridized with a complex target is proportional to the signal detected from a labeled oligonucleotide that recognizes the vector sequence, which reflects the amount of DNA fixed on the spot. These authors found that if the signal detected from complex cDNA targets is divided by the signal from an oligonucleotide in every spot, the fluctuations due to different amounts of cDNA fixed on the filters can be corrected. We tested this assumption using the overgo probe in our arrays and we found different results.

To this end the replicate arrays containing spots with several masses of the GFP plasmid, previously hybridized with the mixture of polyadenilated GFP transcript and 30 µg of sugarcane RNA (see above), were hybridized with the overgo probe. The signal was detected and plotted against the signal from the GFP target. Although the signal obtained with the overgo probe depends linearly on the amount of fixed cDNA (Fig. 1), the same did not hold true for the signal obtained with the GFP target. When the GFP transcripts were in lower concentrations, the lack of linearity was higher (results not shown). In fact, the relation between the signals obtained for the targets with lower concentrations of GFP transcripts and the ones obtained with the overgo probe is better described by a polynomial function of third degree than by a linear function. The simple division of the complex target signal by the overgo would lead to a wrong estimation of the gene expression level. In our hands the overgo signal has been used only to "flag" the

spots that have variations of more than two-fold in the amount of DNA fixed on the filter, thus excluding these spots of further analysis.

Cross-species hybridization of macroarrays

The high level of similarity of gene sequence among grasses provides a powerful tool for understanding and manipulating grass biology (Bennetzen and Freeling, 1997; Guimaraes *et al.*, 1997; Draye *et al.*, 2001). For instance, comparative maps among different grass species demonstrated that the information from one species can be extrapolated to another, for breeding, ecology, evolution and molecular biology (Guimarães *et al.*, 1997). The close relationship between maize and sugarcane (Bennetzen and Freeling, 1997; Draye *et al.*, 2001) prompted us to test whether the gene expression profile of maize roots under Al stress could be evaluated using macroarrays containing sugarcane ESTs. Moreover, Girke *et al* (2000), using microarrays containing *Arabidopsis* ESTs, was found that the expression profile of seeds and leaves from oilseed rape were similar to those observed from the same tissues from *Arabidopsis*.

Aluminum is a major factor limiting plant growth in large areas because it is toxic at micromolar concentration to most crop plants, such as maize. The main site of the Al toxicity is the root apex (Ryan and Kochian 1993). The cloning of genes induced in response to Al is a key step to clarify the mechanisms of Al tolerance and toxicity in plants.

One putative Al-induced maize gene was detected when the macroarrays were hybridized with cDNA targets from maize root tips exposed to Al. The Figure 5 shows a representative macroarray hybridized with cDNA from root apices from

plantlets grown for 24h in the absence of Al or in the presence of 15×10^{-6} Al activity. To validate the expression pattern observed in the macroarrays we used the putative Al-induced EST in northern blots containing RNA from a replicated experiment.

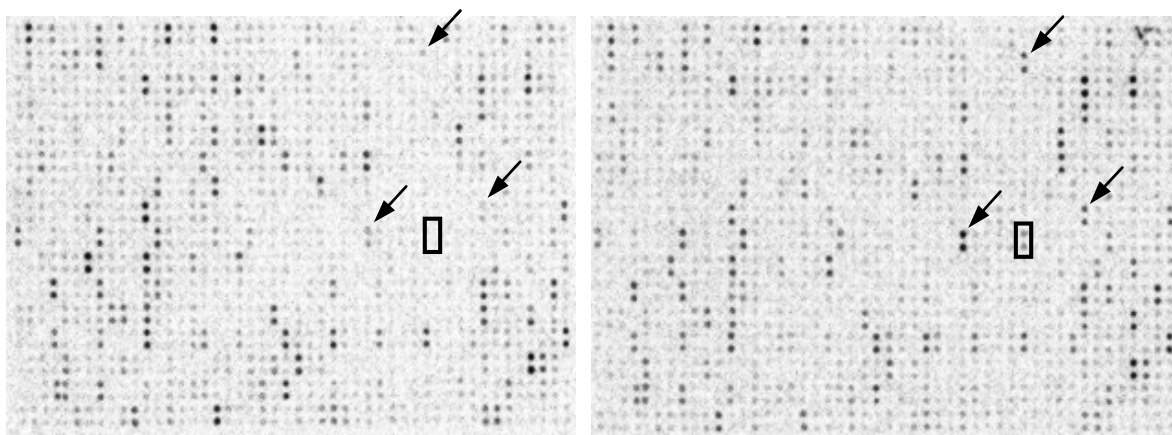


Figure 5: Identification of putative Al-induced genes from maize using arrays containing sugarcane ESTs. Replica macroarrays containing 768 sugarcane ESTs were hybridized with $\alpha^{33}\text{P}$ -labeled targets from total RNA extracted from the root apex of the maize line Cat 100-6 grown on solutions without Al (left) and containing 15×10^{-6} Al activity (right). Arrows indicate putative Al-induced genes. Boxes represent a sugarcane EST chosen for further analysis (see text).

The corresponding sugarcane cDNA clone was sequenced again, confirming its identity. The sugarcane ESTs are clusterized according to the cap3 sequence assembly program (Huang *et al.*, 1999). One of the sugarcane reads (*SCAGLR2033E03.g*) that was differentially expressed in Al-treated maize root showed 93% identity (results not shown) with a transmembrane protein from maize (CAA57955.1, Chevalier *et al.*, 1995). The expression profile of RNA blots hybridized with the sugarcane EST and the zmTRAPRO (kindly supplied by Dr.

Chevalier C.) were very similar (Fig. 6A). The signals from the northern blots and the macroarray were quantified with a phosphorimager and although the absolute values in the RNA blots were not identical to those from the macroarrays, a similar trend was observed (Fig. 6B). Based on these results, we concluded that sugarcane EST macroarrays are a very useful tool to analyze gene expression profiles in related grasses species, such as maize. In fact, we also have been using these macroarrays to identify genes induced by abiotic stress in rice (unpublished results).

Expression profile in response to A1-stress in maize

In higher plants numerous physiological processes might involve the modulation of transmembrane water transport which is in part controlled by transmembrane or major intrinsic proteins (MIP). This protein family exhibits essentially two distinct types of channel properties: (1) specific water transport by the aquaporins, and (2) small neutral solutes transport, such as glycerol by the glycerol facilitators (Maurel, 1997). Aquaporins and their homologs have been found in the plasma membrane and in the tonoplast (vacuolar membrane), thus being called respectively PIPs (Plasma membrane Intrinsic Proteins) and TIPs (Tonoplast Intrinsic Proteins) (Johansson *et al.*, 2000). The presence of aquaporins in cellular membrane is not only a prerequisite for rapid transmembrane water transport, but also enables the organism to regulate the flux of water between cells and within each cell. Some aquaporins are constitutively expressed while the expression of others is regulated by environmental factors such as drought, salinity, hormones and blue light. Gibberellins increase the expression of the gene

corresponding to γ -TIP in an *Arabidopsis* mutant with low levels of endogenous gibberellins. The *Arabidopsis pip1b* gene is activated by blue light, abscisic acid (ABA) and gibberellins. The expression of the gene encoding the root-specific tobacco tonoplast aquaporin, *TobRB7*, was induced by infection by root-knot nematodes. These parasites induce the formation of a feeding site within the plant root, altering the expression of plant genes at this site (Johansson *et al.*, 2000).

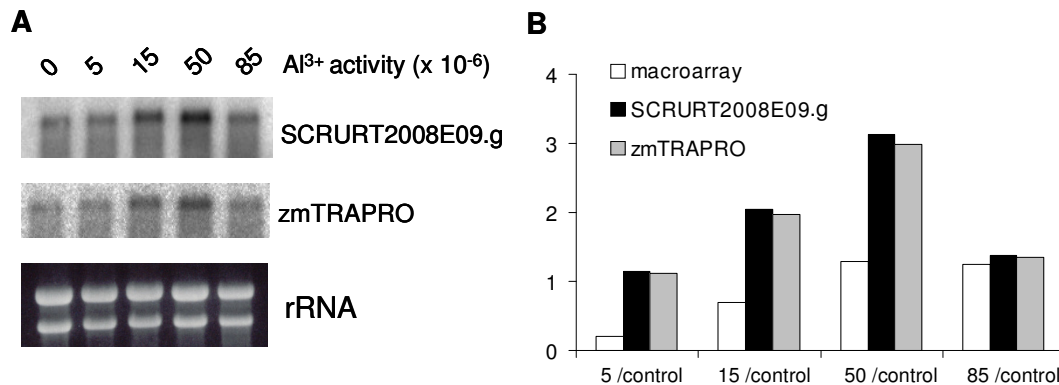


Figure 6: Confirmation of the Al induction by RNA blot analysis. **(A)** Individual targets synthesized from the sugarcane EST coding for a transmembrane protein (SCRURT2008E09.g) and its maize ortholog clone (zmTRAPRO) were hybridized to membranes containing total RNA from the root apex of the maize line, Cat 100-6 (Al-tolerant) exposed to increasing Al concentrations. The total RNA stained with ethidium bromide before being transferred to the membrane is also shown. **(B)** Correlation between the signal from macroarrays and RNA blot. The blots and macroarray data were quantified by using specific software. The macroarray data were normalized with the median of all genes in each filter (see text).

The hydraulic conductivity (L_p) of roots varies in response to a variety of environmental factors including diurnal cycling, drought and salinity, low temperature, oxygen starvation and variation in the supply of nutrients such as N,

P, and S (Maurel, 1997). Studies with cotton plants, grown in a controlled environment, revealed that both N- and P-deficient conditions restricted leaf expansion, reduced transpiration and decreased the hydraulic conductivity of the plants (Radin and Ackerson, 1981; Radin and Eidenbock, 1984). With P-deficient plants it was established that the root was the site of diminished hydraulic conductance, in cortical cells of cotton roots the L_p of the plasma membrane was smaller than in nutrient-sufficient controls, by approximately 60% and 85% in P- and N-stressed plants, respectively (Radin and Matthews, 1989). The L_p of excised roots of *Lotus japonicus* was found to vary over a 5-fold range during a day/night cycle. When mRNAs from *L. japonicus* roots were probed with cDNA from the *Arabidopsis* aquaporin *AthPIP1a* gene, an abundant transcript was found to vary in abundance diurnally under high-stringency conditions, the pattern of the L_p fluctuations of excised roots of *L. japonicus* resembled closely the diurnal pattern of variation in root L_p , suggesting that the diurnal variations in hydraulic conductivity can be correlated with the expression of putative aquaporins in the roots of *L. japonicus* (Henzler *et al.*, 1999).

The first and major symptom of Al toxicity in plants is the inhibition of root growth, which causes a decrease in water and nutrient uptake and a subsequent inhibition of plant growth. Al ions alter the membrane electrical properties, causing destabilization of the cellular membranes (Deleers *et al.*, 1986; Ahn *et al.*, 2001). Moreover, the Al ions cause phosphate precipitation by formation of insoluble composites (Novais and Smyth, 1999). As a result, Al ions interfere in P uptake, causing a P deficiency in plants grown on acid soils or in nutrient solutions (Foy *et al.*, 1978; Jan and Pettersson, 1989). In turn, this metal, reducing the water and

ions uptake, will probably also affect the root L_p . We observed that the highest up-regulation of putative aquaporin genes was on 50×10^{-6} of Al activity (Fig. 6 A and B). These conditions are very important to keep the enzymatic metabolism, which can magnify the Al-stress tolerance in maize root apex.

Conclusions

In the last few years, the hybridization of DNA arrays has proved to be a powerful methodology to perform large-scale gene expression studies. However, some work is still to be done on the standardization of the method and on the analysis of its output. Working on this direction, we also attempted to validate the use of heterologous target to perform large-scale gene expression profiling.

The strategy of hybridizing the filter arrays with the overgo labeled target has shown to be efficient on estimating the amount of DNA fixed on each spot. Using this method, we also validated the use of multiple pin strikes to reduce the fluctuations on the fixed DNA amounts. However, we also concluded that the signals obtained with the overgo probe are not indicated to normalize directly the fluctuations on an array experiment output caused by variations on the amounts of fixed DNA, because the output signals of an array hybridization do not vary linearly with this amounts of DNA. We only use the signals obtained with the overgo probe as a spot quality control. The strategy of normalizing an array output signals by their median has also proved to be efficient to reduce fluctuations on the signals caused by fluctuations on the hybridization efficiency or target activity.

Finally, the confirmation of our array results by RNA-blot validated the use of heterologous targets on macroarrays experiments, showing that this strategy can be a powerful tool to perform fast and large-scale gene expression profiling between closely related species.

Acknowledgments

We would like to thank R. V. dos Santos and the Sucest/Mazest technicians' team for their excellent assistance. We are grateful to Dr. M. Schummer for supplying the GFP clone and for his help in the initial phase of this work and to Dr. C. Chevalier for supplying the zmTRAPRO clone. This work was supported by grants from PADCT/CNPq (62.0472/98-7) to MM. JMF and FTSN were recipients of fellowships from FAPESP. RD was recipient from a fellowship from CAPES. MM received a Research Fellowship from CNPq.

References

- Ahn, S.J., Sivaguru, M., Osawa, H., Chung, G.C. and Matsumoto, H. (2001). Aluminum inhibits the H⁺-ATPase activity by permanently altering the plasma membrane surface potentials in squash roots. *Plant Physiol.* 126: 1381-1390.
- Azaizah, H. and Steudle, E. (1991). Effects of salinity on water transport of excised maize (*Zea mays*) roots. *Plant Physiol.* 97: 1136-1145.
- Baligar, V.C. and Ahlrichs, J.L. (1998). Nature and distribution of acid soils in the world. In: *Proceedings of a workshop to develop a strategy for collaborative research and dissemination of technology in sustainable crop production in acid savannas and other problem soils of the world.* (Schaffert, R. E.), Purdue University, West Lafayette, IN 1-11.
- Bennetzen, J.L. and Freeling, M. (1997). The unified grass genome: synergy in synteny. *Genome Res.* 7: 301-306.
- Chaumont, F., Barrieu, F., Jung, R. and Chrispeels, M.J. (2000). Plasma membrane intrinsic proteins from maize cluster in two sequence subgroups with differential aquaporin activity. *Plant Physiol.* 122: 1025-1034.
- Chevalier, C., Bourgeois, E., Pradet, A. and Raymond, P. (1995). Molecular cloning and characterization of six cDNAs expressed during glucose starvation in excised maize (*Zea mays* L.) root tips. *Plant Mol. Biol.* 28: 473-485.
- Delahaba, G. and Cantoni, G.L. (1959). Enzymatic synthesis of S-Adenosyl-L-Homocysteine from adenosine and homocysteine. *J. Biol. Chem.* 234: 603-608.
- Deleers, M., Servais, J.P. and Wulfert, E. (1986). Neurotoxic cations induce membrane rigidification and membrane fusion at micromolar concentrations. *Biochim. Biophys. Acta* 855: 271-276.
- Delhaize, E. and Ryan, P.R. (1995). Aluminum toxicity and tolerance in plants. *Plant Physiol.* 107: 315-321.

- Draye, X., Lin, Y.R., Qian, X.Y., Bowers, J.E., Burow, G.B., Morrell, P.L., Peterson, D.G., Presting, G.G., Ren, S.X., Wing, R.A. and Paterson, A.H. (2001). Toward integration of comparative genetic, physical, diversity, and cytomolecular maps for grasses and grains, using the sorghum genome as a foundation. *Plant Physiol.* 125: 1325-1341.
- English, J., Pearson, G., Wilsbacher, J., Swantek, J., Karandikar, M., Xu, S.C. and Cobb, M.H. (1999). New insights into the control of map kinase pathways. *Exp. Cell Res.* 253: 255-270.
- Espartero, J., Pintortoro, J.A. and Pardo, J.M. (1994). Differential accumulation of S-Adenosylmethionine Synthetase transcripts in response to salt stress. *Plant Mol. Biol.* 25: 217-227.
- Foy, C.D., Chaney, R.L. and White, M.C. (1978). Physiology of metal toxicity in plants. *Annu. Rev. Plant Physiol.* 29: 511-566.
- Girke, T., Todd, J., Ruuska, S., White, J., Benning, C. and Ohlrogge, J. (2000). Microarray analysis of developing Arabidopsis seeds. *Plant Physiol.* 124: 1570-1581.
- Gowri, G., Bugos, R.C., Campbell, W.H., Maxwell, C.A. and Dixon, R.A. (1991). Stress responses in alfalfa (*Medicago sativa*) .10. molecular- cloning and expression of S-Adenosyl-l-methionine-caffeic acid 3-o-methyltransferase, a key enzyme of lignin biosynthesis. *Plant Physiol.* 97: 7-14.
- Guimaraes, C.T., Sills, G.R. and Sobral, B.W.S. (1997). Comparative mapping of andropogoneae: *Saccharum* L. (sugarcane) and its relation to sorghum and maize. *Proc. Natl. Acad. Sci. USA.* 94: 14261-14266.
- Haseloff, J., Siemering, K.R., Prasher, D.C. and Hodge, S. (1997). Removal of a cryptic intron and subcellular localization of green fluorescent protein are required to mark transgenic arabidopsis plants brightly. *Proc. Natl. Acad. Sci. USA.* 94: 2122-2127.
- Henzler, T., Waterhouse, R.N., Smyth, A.J., Carvajal, M., Cooke, D.T., Schaffner, A.R., Steudle, E. and Clarkson, D.T. (1999). Diurnal variations in hydraulic conductivity and root pressure can be correlated with the expression of putative aquaporins in the roots of *Lotus japonicus*. *Planta* 210: 50-60.

- Huang, G.M., Ng, W.L., Farkas, J., He, L., Liang, H.A., Gordon, D., Yu, J. and Hood, L. (1999). Prostate cancer expression profiling by cDNA sequencing analysis. *Genomics* 59: 178-186.
- Jan, F. and Pettersson, S.. (1989). Varietal Diversity of upland rice in sensitivity to aluminum. *J. Plant Nutr.* 12: 973-993.
- Johansson, I., Karlsson, M., Johanson, U., Larsson, C. and Kjellbom, P. (2000). The role of aquaporins in cellular and whole plant water balance. *Biochim. Biophys. Acta - Biomembranes* 1465: 324-342.
- Kawalleck, P., Plesch, G., Hahlbrock, K. and Somssich, I.E. (1992). Induction by fungal elicitor of S-Adenosyl-L-Methionine Synthetase and S-Adenosyl-L-Homocysteine Hydrolase messenger- RNAs in cultured-cells and leaves of *Petroselinum-crispum*. *Proc. Natl. Acad. Sci. USA.* 89: 4713-4717.
- Kim, C.S., Kwak, J.M., Nam, H.G., Kim, K.C. and Cho, B.H. (1994). Isolation and characterization of 2 cDNA clones that are rapidly induced during the wound response of *Arabidopsis thaliana*. *Plant Cell Rep.* 13: 340-343.
- Kuhn, E. (2001) From library screening to microarray technology: strategies to determine gene expression profiles and to identify differentially regulated genes in plants. *Ann. Bot.* 87: 139-155.
- Logemann, J., Schell, J. and Willmitzer, L. (1987). Improved method for the isolation of RNA from plant-tissues. *Anal. Biochem.* 163: 16-20.
- Maurel, C. (1997). Aquaporins and water permeability of plant membranes. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 48: 399-429.
- Mayne, M.B., Coleman, J.R. and Blumwald, E. (1996). Differential expression during drought conditioning of a root- specific S-adenosylmethionine synthetase from jack pine (*Pinus banksiana*) seedlings. *Plant Cell Environ.* 19: 958-966.
- Parker, D.R. (1995). Root-growth analysis - an underutilized approach to understanding aluminum rhizotoxicity. *Plant Soil* 171: 151-157.
- Peleman, J., Boerjan, W., Engler, G., Seurinck, J., Botterman, J., Alliotte, T., Vanmontagu, M. and Inze, D. (1989). Strong cellular preference in the

- expression of a housekeeping gene of *Arabidopsis thaliana* encoding S-Adenosylmethionine Synthetase. *Plant Cell* 1: 81-93.
- Perret, E., Ferrán, E.A., Marinx, O., Liauzun, P., Dumont, X., Fournier, J., Kaghad, M., Ferrara, P. and Caput, D. (1998). Improved differential screening approach to analyze transcription variations in organized cDNA libraries. *Gene* 208: 103-115.
- Pietu, G., Alibert, O., Guichard, V., Lamy, B., Bois, F., Leroy, E., Mariagesamson, R., Houlgatte, R., Soularue, P. and Auffray, C. (1996). Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. *Genome Res.* 6: 492-503.
- Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature* 32: 496-501
- Radin, J.W. and Ackerson, R.C. (1981). Water relations of cotton plants under nitrogen deficiency .3. stomatal conductance, photosynthesis, and abscisic-acid accumulation during drought. *Plant Physiol.* 67: 115-119.
- Radin, J.W. and Eidenbock, M.P. (1984). Hydraulic conductance as a factor limiting leaf expansion of phosphorus-deficient cotton plants. *Plant Physiol.* 75: 372-377.
- Radin, J.W. and Matthews, M.A. (1989). Water transport-properties of cortical-cells in roots of nitrogen-deficient and phosphorus-deficient cotton seedlings. *Plant Physiol.* 89: 264-268.
- Ryan, P.R. and Kochian, L.V. (1993). Interaction between aluminum toxicity and calcium-uptake at the root apex in near-isogenic lines of wheat (*Triticum aestivum*) differing in aluminum tolerance. *Plant Physiol.* 102: 975-982.
- Sambrook, J., Fritsch, E.F. and Maniats, T. (1989). *Molecular cloning: a laboratory manual*. Cold Spring Harbor, New York.
- Sasaki, M., Yamamoto, Y. and Matsumoto, H. (1996). Lignin deposition induced by aluminum in wheat (*Triticum aestivum*) roots. *Physiol. Plant.* 96: 193-198.

- Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H. and Herzel, H. (2000). Normalization strategies for cDNA microarrays. *Nucleic Acids Res.* 28: E47.
- Schummer, M., Ng, W.V., Baumgarner, R.E., Nelson, P.S., Schummer, B., Bednarski, D.W., Hassell, L., Baldwin, R.L., Karlan, B.Y. and Hood, L. (1999). Comparative hybridization of an array of 21,500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. *Gene* 238: 375-385.
- Schummer, M., Ng, W., Nelson, P.S., Bumgarner, R.E. and Hood, L. (1997). Inexpensive handheld device for the construction of high- density nucleic acid arrays. *Biotechniques* 23: 1087-1092.
- Shimizu, M., Ito, T., Terashima, S., Hayashi, T., Arisawa, M., Morita, N., Kurokawa, S., Ito, K. and Hashimoto, Y. (1984). Inhibition of lens aldose reductase by flavonoids. *Phytochemistry* 23: 1885-1888.
- Tabor, C.W. and Tabor, H. (1984). Methionine adenosyltransferase (S-Adenosylmethionine Synthetase) and S-Adenosylmethionine Decarboxylase. *Adv. Enzymol. Relat. Areas Mol. Biol.* 56: 251-282.
- Vanbreusegem, F., Dekeyser, R., Gielen, J., Vanmontagu, M. and Caplan, A. (1994). Characterization of a S-Adenosylmethionine Synthetase gene in rice. *Plant Physiol.* 105: 1463-1464.
- Woodson, W.R., Park, K.Y., Drory, A., Larsen, P.B. and Wang, H. (1992). Expression of ethylene biosynthetic-pathway transcripts in senescing carnation flowers. *Plant Physiol.* 99: 526-532.
- Yang, S.F. and Hoffman, N.E. (1984). Ethylene biosynthesis and its regulation in higher-plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 35: 155-189.

RUNNING TITLE: Effects of sealing plastics on nylon cDNA arrays

TITLE: Evaluation of the performance of different plastics used to seal nylon cDNA arrays.

KEYWORDS: signal detection, nylon arrays, sensitivity

Antonio Paulino da Costa Netto^{1,2a}, Rodrigo Duarte Drummond^{1,3a}, Juliana de Maria Felix^{1,3}, Renato Atílio Jorge⁴, Marcelo Menossi^{1,2,3*}

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

²*Departamento de Bioquímica, Instituto de Biologia, Universidade Estadual de Campinas (UNICAMP), Campinas, SP, Brazil.*

³*Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas (UNICAMP), Campinas, SP, Brazil.*

⁴*Departamento de Físico-Química, Instituto de Química, Universidade Estadual de Campinas (UNICAMP), Campinas, SP, Brazil.*

^aBoth authors contributed equally to this work.

*Address for Correspondence:

Centro de Biologia Molecular e Engenharia Genética, State University of Campinas (Unicamp), CP 6010, 13083-970, Campinas, SP, Brazil

Telephone number: +55 (19) 35211143

Fax number: +55 (19) 35211089

E-mail menossi@unicamp.br

Abstract

cDNA arrays are a powerful tool for discovering gene expression patterns. Nylon arrays have the advantage that they can be re-used several times. A key issue in high throughput gene expression analysis is sensitivity. In the case of nylon arrays, signal detection can be affected by the plastic bags used to keep membranes humid. In this study, we evaluated the effect of five types of plastics on the radioactive transmittance, number of genes with a signal above the background, and data variability. A polyethylene plastic bag 69 μm thick had a strong shielding effect that blocked 68.7% of the radioactive signal. The shielding effect on transmittance decreased the number of detected genes and increased the data variability. Other plastics which were thinner gave better results. Although plastics made from polyvinylidene chloride, polyvinyl chloride (both 13 μm thick) and polyethylene (29 and 7 μm thick) showed different levels of transmittance, they all gave similarly good performances. Polyvinylidene chloride and polyethylene 29 μm thick were the plastics of choice because of their easy handling. For other types of plastics, it is advisable to run a simple check on their performance in order to obtain the maximum information from nylon cDNA arrays.

Introduction

Around one-third of the genes found in every genome project have an unknown function, and studies aimed at determining the physiological roles of gene products are one of the principal tasks in genomics. cDNA array technology is the main strategy for assessing gene expression profiles on a large scale. cDNA clones corresponding to the genes under study are arrayed on a solid support, such as glass or nylon. The arrays are then hybridized with labeled probes obtained from the samples under study. The amount of labeled probe bound at the cDNAs on the arrays is proportional to the number of mRNA molecules present in the initial population of cells. Because each array contains thousands of genes, the expression profile of a relevant part of the cell transcriptome can be easily assessed.

cDNA arrays based on glass slides (microarrays) are by far the most common way to study gene expression on a large scale, whereas cDNA arrays on nylon filters (macroarrays) are an alternative to glass slides. Although the latter accommodate a lower number of genes, they are cheaper and more robust since they are based on the hybridization of radiolabeled probes to filters, a routine procedure in most molecular biology laboratories. Using macroarrays, we have successfully identified several genes that are induced by cold stress in sugarcane (Nogueira *et al.*, 2003), and several recent studies on gene expression profiling have also used macroarrays, including the identification of angiogenesis factors in lung cancer tissues (Ohira *et al.*, 2002), the response to sulfur starvation in *Arabidopsis thaliana* plants (Hirai *et al.*, 2002), and the identification of genes involved in ceramide-dependent neuronal apoptosis (Decraene *et al.*, 2002).

One of the key issues in cDNA array analysis is the detection of genes that are expressed at low levels, particularly since these correspond to the largest part of the transcriptome in most cells (Jongeneel *et al.*, 2003; Bishop *et al.*, 1974). Thus, finding ways to increase the signal detection is of great importance, since the higher the signals the greater the reliability of the measurements.

During our studies using macroarrays we noted that little attention has been given to the plastic bags used to seal the nylon filters (the latter can be re-used after stripping, but keeping the membranes wet is essential for subsequent rounds of hybridization). We found that a significant fraction of the radioactivity is blocked by the most commonly used plastic material. We have therefore compared several plastics to determine which provides the best transmittance and evaluated their effect on the detection of gene expression and other aspects of cDNA array analysis.

Material and Methods

Plant material

The maize inbred line Cat 100-6 (Al-tolerant) was obtained from the Germplasm Bank of the Universidade Estadual de Campinas, Campinas, Brazil. The plants were grown in the field and self-pollinated.

Maize plant growth

Seeds were surface-sterilized with 70% ethanol for 1 min and with 0.5% (v/v) sodium hypochloride for 20 min then rinsed four times with sterile water and germinated between layers of moist filter paper at 28 °C for 60 h. After 3 days, the seedlings were transferred to polystyrene holders that were floated on an aerated nutrient solution (Moon *et al.*, 1997).

Nylon array hybridization

Plasmids containing sugarcane ESTs were obtained from the SUCEST RT1 library made from sugarcane root RNA (Vettore *et al.*, 2001). Plasmidial DNA obtained by alkaline lysis was denatured with 0.2 M NaOH. DNA was spotted on nylon filters (Hybond-N, AP-Biotech, USA) with a hand-held 96-pin printhead tool which typically deposited 0.1 µl (around 10 ng) of DNA solution (V&P-Scientific, USA), as described by Schummer *et al.* (1997). The DNA was fixed to the filters by baking at 80 °C for 2 h. The filters contained 768 ESTs in duplicate, to give 1536 spots per filter.

³³P-Labelled cDNA was produced from 30 µg of total RNA and hybridized to nylon arrays essentially as described by Nogueira *et al.* (2003). Membranes were washed with 1× SSC and 0.1% SDS for 15 min at 65 °C, and twice with 0.1× SSC and 0.1% SDS at 65 °C (15 min each).

Detection of gene expression

Five different plastics sold in supermarkets were compared: polyvinyl chloride (PVC, 13 µm thick, MajiPack™, Inproco, Brazil), polyvinylidene chloride (PVDC, 13 µm thick, Saran Wrap™, Johnson & Son, USA) and three polyethylene plastics 69, 29 and 7 µm thick (PE-69, PE-29 and PE-7, Garoa, Brazil). Four nylon filters were sealed with each plastic, and four additional filters without sealing were used as controls. After sealing, the filters were exposed for 96 h to imaging plates in an FLA3000-G screen system (Fuji Photo Film, Japan). The intensity of each spot on the digitalized images was quantified using Arrayvision™ software (Imaging Research, Canada). The local background was subtracted from the intensity of each spot and the data were rearranged into MS Excel (Microsoft, USA) files for further analysis.

Statistical analysis

Three parameters were used to evaluate the interference of each type of plastic on the gene expression signal detected: the total radiation on each membrane (i.e., the sum of the intensities of all spots), the number of spots that showed a signal at least 10% greater than the local background emission (around

the spot) and the number of genes for which the corresponding pair of replicate spots showed signals that varied in intensity by less than two fold. For each parameter, the data from each type of plastic and from naked nylon arrays were compared in pairs. To avoid assumptions regarding the distribution of the data, statistical comparisons were done using the Wilcoxon-Mann-Whitney rank sum test (Woolson, 1987), a nonparametric test for two independent samples.

Results and Discussion

During a study of gene expression in maize plants exposed to aluminum stress, we decided to check the effect of commonly used plastic sealing bags on signal detection. Nylon cDNA arrays have the advantage they can be reused several times, in the same way as Southern and northern blots. We have reused nylon arrays up to seven times without noticeable loss of signal (unpublished results). However, the membranes need to be sealed to avoid drying, and this can lead to covalent linking between the immobilized probes on the array and the radioactive target probes. The inconvenience of sealing with plastics is their blocking effect on the radioactive signal (Figure 1).

Initially, the effects of three different plastics (PE-69, PVC and PVDC) were compared to naked membranes (NM group). Nylon membranes that had previously been exposed without plastic were wrapped in the three types of plastic and the signals were then measured again. The set of naked membranes was exposed twice without wrapping in plastic in order to control for the loss of signal resulting from radioisotope decay. The PE-69 plastic, routinely used by our group, had a strong blocking effect, with only 31.3% of the original emission from naked membranes being detected (Figure 2). PVC and PVDC showed less blockage, with 70.9% and 75.5% of the emission being detected. The latter values were significantly higher ($p < 0.05$) than those observed with PE-69, but were not significantly different from naked membranes.

One of the main problems with the decreased signal caused by the interference of plastics is that genes with low levels of expression are not detected. Since several regulatory genes are expressed at low levels, this loss must be

minimized. The number of detected spots with a signal at least 10% higher than the local background was used as a parameter to assess this effect. This percentage is usually used as a cut-off value in cDNA array experiments, with only signals above this level being considered in gene expression analysis. Compared to the naked membranes, a very similar number of spots was detected when using PVC and PVDC (97.5% and 98.9%, respectively; Figure 2). The lowest detection was observed with PE-69 (92.3%). In a typical assay with 3000 genes, the difference between PE-69 and PVDC would represent a loss of 198 genes, which is high. The largest difference in the number of detected spots (between PE-69 and NM) was 7.7%, much lower than the difference between the total emission (68.7%; Figure 2).

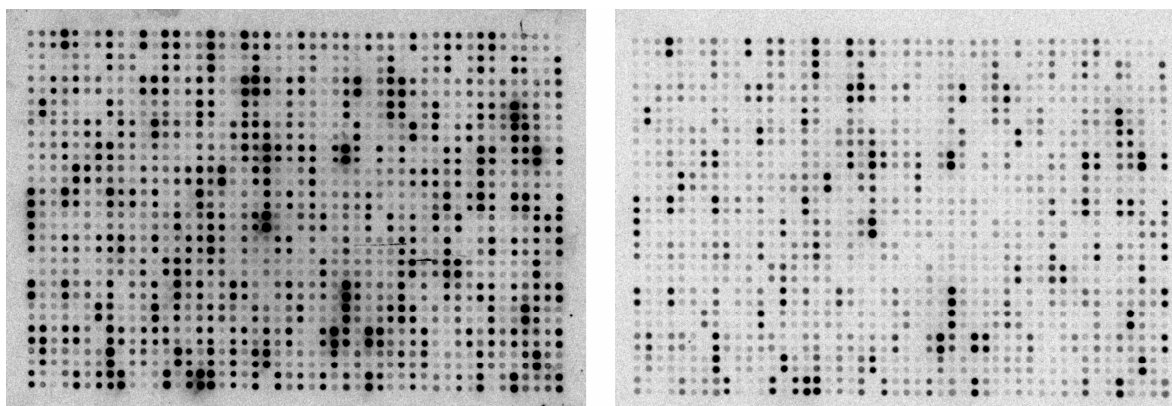


Figure 1: Example of the effect of plastics on signal detection. Filters containing 768 ESTs fixed in duplicate, giving a total of 1536 spots per filter, were exposed with different sealing plastics. Left: image of radioactive emission of a naked membrane. Right: image of radioactive emission of a replicate membrane sealed with PE-69.

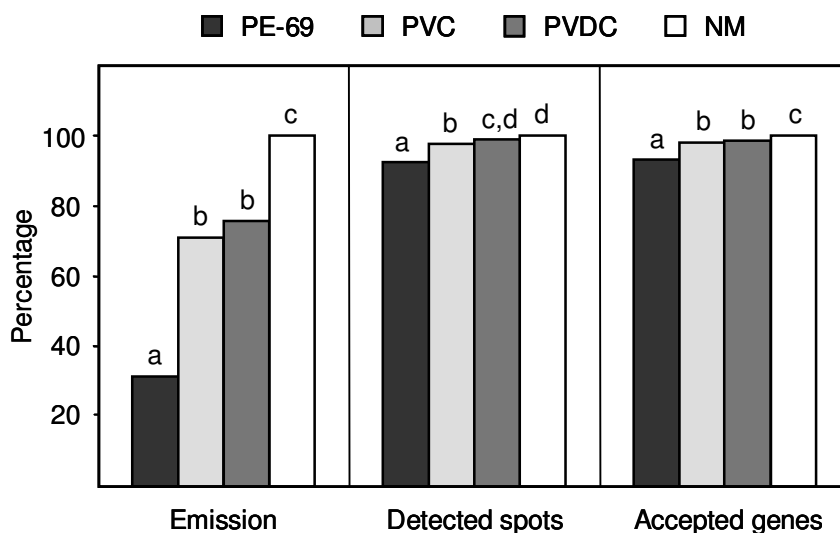


Figure 2: Effect of sealing plastics on the geometric means of the ratios of total emission, detected spots and accepted genes. The values obtained for each sealed filter were divided by those previously detected for the same filter without sealing. In the NM group, the values were determined twice without sealing in order to control for radioisotope decay. The geometric mean of four ratios is shown for each group. Letters at the top of each column indicate significantly different groups ($p < 0.05$).

In cDNA array experiments, each gene is usually spotted twice in the same array. Random fluctuations during spotting can affect the amount of DNA in each spot. Although in glass arrays these fluctuations are not a problem (a control target is usually also hybridized with the test target), in nylon arrays, which are hybridized with only one target at a time, these fluctuations may result in false positives, i.e. genes that are not up-regulated, but give higher signals in one array because of different amounts of DNA (the same is true for false negatives). To avoid this type of fluctuation in nylon arrays experiments, it is usual to eliminate from subsequent analysis all genes whose signals from different spots on the same membrane vary more than two-fold (Larkin *et al.*, 2002). Since the shielding effect of sealing

plastics decreases the signal intensities, it also increases data variability, causing more spots to be discarded by this criterion. To evaluate this effect, we analyzed the differences in the number of genes discarded by this criterion in nylon filters wrapped in different sealing plastics. The percentage of accepted genes in nylon arrays sealed with PVC and PVDC were only 2% lower than the percentage accepted in naked membranes (Figure 2). In contrast, on arrays sealed with PE-69, 7% of the spots were rejected.

These results showed that PVC and PVDC were the best sealing plastics for arrays compared to PE-69. However, working with 13 μm thick PVC can be troublesome because the plastic easily wrinkles and the hybridization solution sometimes leaks from the terminally sealed edges. Since PVDC is also not easily available in countries other than the USA, we tested additional polyethylene-based plastics which are used to produce plastic bags worldwide. These plastics were thinner than PE-69 (7 μm and 29 μm) and were compared with PVC (Figure 3). The radioactive emission was the only parameter that showed significant variation between the three types of plastics. Whereas PE-7 had a similar transmittance than PVC, PE-29 clearly had a higher shielding effect, transmitting only 82% of the emission detected with PVC. The differences in the number of detected spots and accepted genes between the three plastics were very small and not significant. Based on these results, and because PE-29 is easier to handle than PE-7 and PVC, we concluded that sealing arrays with PE-29 plastic was a good alternative for increasing the signal detection in nylon arrays.

The differences we found may be explained by the two factors that most influence the transmittance of any material, namely, chemical composition and thickness. Thus, PVDC and PVC had the same thickness and very similar performances. Among the polyethylene plastics, the lower performance of PE-69 probably reflected its greater thickness. However, although the difference in thickness between PE-29 and PE-7 affected the transmittance, no significant differences were found in the number of spots detected or accepted. This may indicate that transmittance becomes a problem only below a certain threshold level.

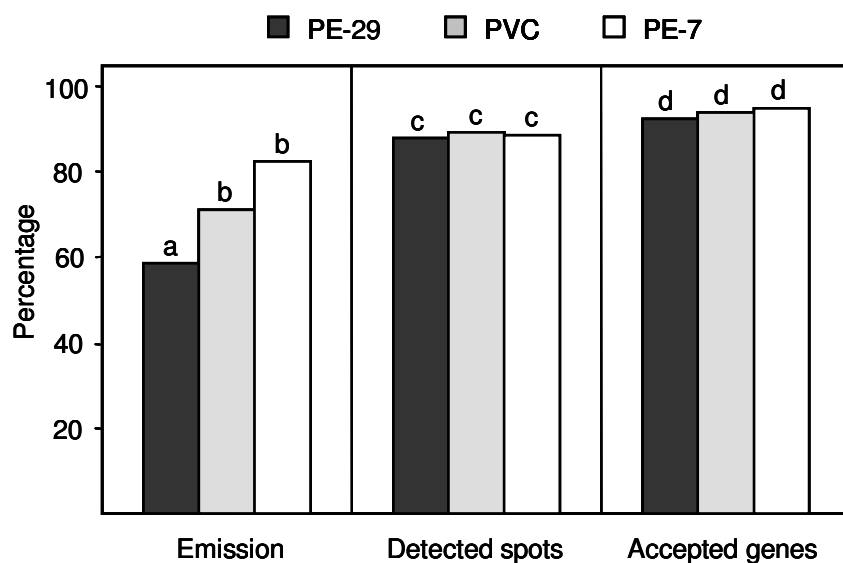


Figure 3: Performance of polyethylene plastics compared to PVC. The ratios were determined as in figure 2. Letters at the top of each column indicate significantly different groups ($p < 0.05$).

In conclusion, our findings indicate that the choice of the sealing plastic is very important for obtaining the maximum information from nylon arrays. Among the plastics tested, PVDC showed the best performance, although in countries where this is not available, PE-29 proves a good alternative. Based on these findings, we have used PE-29 in various experiments, including the search for cold-induced genes from sugarcane (Nogueira *et al.*, 2003), and the results from the hybridizations used in this work will be published elsewhere. We recommend that researchers using other plastics should check the shielding effect in order to obtain the most information from their nylon arrays.

Acknowledgements

This work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (00/07433-0), the European Commission INCO-DC (ICA4-CT2000-30017) and PADCT/CNPq (62.0472/98.7). APCN, RDD and JMF were recipients of fellowships from CNPq, CAPES and FAPESP, respectively. MM received a Research Fellowship from CNPq.

References

- Bishop, J.O., Morton, J.G., Rosbash, M. and Richardson, M. (1974). Three abundance classes in HeLa cell messenger RNA. *Nature* 250: 199–204.
- Decraene, C., Brugg, B., Ruberg, M., Eveno, E., Matingou, C., Tahi, F., Mariani, J., Auffray, C. and Pietu, G. (2002). Identification of genes involved in ceramide-dependent neuronal apoptosis using cDNA arrays. *Genome Biol.* 3: 1-22.
- Hirai, M.Y., Fujiwara, T., Awazuhara, M., Kimura, T., Noji, M. and Saito, K. (2003). Global expression profiling on sulfur-starved Arabidopsis by DNA macroarray reveals the role of O-acetyl-L-serine as a general regulator of gene expression in response to sulfur nutrition. *Plant J.* 33: 651-663.
- Jongeneel, C.V., Iseli, C., Stevenson, B.J., Riggins, G.J., Lal, A., Mackay, A., Harris, R.A., O'Hare, M.J., Neville, A.M., Simpson, A.J.G. and Strausberg, R.L. (2003). Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. *Proc. Natl. Acad. Sci. USA.* 100: 4702-4705.
- Larkin, P., Sabo-Attwood, T., Kelso, J. and Denslow, N.D. (2002). Gene expression analysis of largemouth bass exposed to estradiol, nonylphenol, and p,p'-DDE. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* 133: 543-557.
- 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.
- Ohira, T., Akutagawa, S., Usuda, J., Nakamura, T., Hirano, T., Tsuboi, M., Nishio, K., Taguchi, F., Ikeda, N., Nakamura, H., Konaka, C., Saijo, N. and Kato, H. (2002). Up-regulated expression of angiogenesis factors in post-chemotherapeutic lung cancer tissues determined by cDNA macroarray. *Oncol. Rep.* 9: 723-728.
- Schummer, M., Ng, W.V., Bumgarner, R.E., Nelson, P.S., Schummer, B., Bednarski, D.W., Hassell, L., Baldwin, R.L., Karlan, B.Y. and Hood, L. (1999). Comparative hybridization of an array of 21,500 ovarian cDNAs for

the discovery of genes overexpressed in ovarian carcinomas. *Gene* 238: 375-385.

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.

Woolson, R.F. (1987). *Statistical Methods for the Analysis of Biomedical Data*. Wiley, New York, 513pp.

Capítulo 2: Pré-processamento e normalização dos dados

Experimentos com arranjos de DNA baseiam-se na hipótese de que os sinais detectados para cada gene refletem os seus níveis de expressão nas diferentes amostras de RNA analisadas (Schena et al., 1995). Porém, os dados provenientes desse tipo de experimento são em geral afetados por diversas fontes de ruído (Schuchhardt *et al.*, 2000). Assim, antes de se aplicar qualquer análise estatística é necessário realizar um pré-processamento dos mesmos. Procura-se com isso eliminar tendências devidas a diferenças na marcação das amostras, na eficiência da hibridação das sondas aos arranjos e diferenças na emissão de fundo dos arranjos (*background*), entre outros fatores. Além disso, genes que apresentam outros problemas nos seus dados, como sinais muito fracos, são identificados para evitar que venham a motivar falsas conclusões.

2.1 - Subtração do background

O primeiro passo na análise de dados de arranjo de DNA consiste em subtrair-se, do sinal obtido para cada ponto do arranjo (*spot*), um valor, usualmente chamado de *background* ou emissão de fundo, que corresponde à porção do sinal devida à hibridação não específica de moléculas de DNA marcado à superfície do suporte do arranjo. Esse valor é calculado diretamente pelo programa utilizado para a quantificação da imagem, e pode ser obtido de diversas formas.

Pontos específicos do arranjo, onde nenhum DNA havia sido depositado, podem ser selecionados e a média dos sinais obtidos nesses pontos pode ser considerada como o background de todo o arranjo. A desvantagem dessa estratégia é que um único valor de background é definido para todo o arranjo, assim desprezando possíveis variações na emissão de fundo em diferentes regiões do mesmo.

Quando o arranjo apresenta uma estrutura de *spots* divididos em subconjuntos (*subgrids*, como por exemplo os pontos cujo DNA foi depositado por uma mesma agulha do equipamento que montou o arranjo), regiões ao redor de cada um desses subconjuntos podem ser selecionadas para quantificar o *background*. A média dos pontos em cada uma dessas regiões pode ser então considerada como *background* para o respectivo *subgrid*.

Uma região circular ao redor de cada ponto de fixação do DNA no arranjo (*spot*) pode ser utilizada para quantificar o *background*. Nesse caso a média (ou mediana) dos sinais obtido nos pontos (*pixels*, ou seja, pontos da imagem) dessa região é considerada como o valor de *background* para o respectivo *spot*. A grande vantagem dessa estratégia é que o cálculo local do *background* compensa qualquer irregularidade na emissão de fundo ao longo do arranjo. Uma desvantagem, porém, é que caso um *spot* tenha sinal excessivamente forte, este pode contaminar a região ao redor do mesmo, aumentando assim o *background* do *spot* e com isso, após a subtração do mesmo, o respectivo sinal tende a ser subestimado (isto ocorre principalmente em arranjos hibridados com cDNA marcado radiativamente). Ainda assim, essa estratégia foi considerada por nosso grupo de pesquisas como uma das mais eficientes para a quantificação do

background, e foi utilizada na quantificação da maioria das hibridações de arranjos realizadas neste trabalho, incluindo as hibridações do experimento de deficiência de fosfato em cana-de-açúcar (descrito no Capítulo 7).

O algoritmo *lowess* (*locally weighted linear regression*, Cleveland, 1979) pode ser aplicado na determinação do *background*. Como dito acima, o *background* calculado ao redor de cada ponto de fixação pode sofrer influência do sinal do próprio ponto, ou outras flutuações indesejáveis. Uma forma de diminuir esses efeitos é aplicar o algoritmo *lowess* para alisar os valores do *background* em função das posições dos pontos no arranjo. Com isso, valores aberrantes do *background* são evitados, devido a sua substituição por uma média dos mesmos com os valores de *background* dos pontos de fixação vizinhos, ponderada em função da distância de cada ponto vizinho ao ponto em questão (Colantuoni *et al.*, 2002). Essa estratégia é principalmente indicada para arranjos hibridados com cDNA marcado radiativamente, onde os valores do *background* costumam ser elevados em relação aos sinais detectados nos *spots*.

2.2 - Identificação de problemas e filtragem dos dados

O processo de confecção dos arranjos de DNA, assim como sua hibridação e quantificação da imagem obtida, é passível de erros e irregularidades. Tais artefatos podem levar a conclusões equivocadas. Por isso, é usual filtrar os dados provenientes de arranjos de DNA para eliminar das análises futuras aqueles provenientes de *spots* com problemas. Diversos critérios podem ser utilizados nessa filtragem.

Pode-se eliminar genes cujos sinais dos respectivos *spots* réplicas apresentam problemas numa hibridação teste com sonda *overgo*. Esse critério aplica-se apenas aos macroarranjos de cDNA em membranas de *nylon*, que podem ser hibridados diversas vezes (as sondas hibridadas podem ser removidas e o arranjo pode ser reutilizado). Nesse caso, para verificar a qualidade do arranjo e a quantidade de cDNA fixado em cada *spot*, pode-se hibridá-lo com uma sonda complementar a uma região do vetor de clonagem usado para amplificar os cDNAs fixados, denominada sonda *overgo*. Essa sonda, marcada radiativamente, é adicionada em grande quantidade na hibridação, de maneira que a grande maioria das moléculas de cDNA fixadas no arranjo são hibridadas com moléculas de sonda. Após a quantificação da imagem obtida dessa hibridação, pode-se eliminar os *spots* que apresentam sinal fraco, pois isso indica que neles há pouco cDNA fixado. Os genes que apresentam uma grande variação entre os sinais de seus *spots* réplicas também podem ser eliminados, pois isso indica que há diferenças nas quantidades de cDNA para esse gene fixadas nos diferentes *spots*. Essas diferenças poderiam levar a falsas conclusão quanto ao padrão de expressão desse gene nas amostras de RNA que vierem a ser analisadas com o arranjo em questão. Nos arranjos de cDNA em *nylon* confeccionados pelo nosso grupo, descartamos os *spots* que apresentavam, na hibridação com sonda *overgo*, sinal abaixo de 10% da média dos sinais de todos os *spots* na membrana nessa mesma hibridação, bem como os genes que apresentavam uma variação de mais de duas vezes entre os sinais de seus *spots* réplicas.

É possível eliminar os *spots* com sinais iguais a zero em uma proporção (predefinida) das hibridações. Isso pode indicar que a deposição ou a fixação do

DNA nesses *spots* foi ineficiente. Nesse caso, se o respectivo gene estiver fixado em outros *spots* do arranjo onde o problema não se repete, apenas o *spot* em questão é eliminado e o gene pode continuar nas análises.

Pode-se eliminar os *spots* com sinais mais fracos ou com razões sinal/ruído muito pequenas. Isso também pode indicar problemas na fixação do DNA no *spot*. Nesse caso, para cada hibridação, estabelece-se um valor mínimo abaixo do qual os sinais são considerados como muito fracos e, conseqüentemente, descarta-se os respectivos *spots*. Esse valor mínimo pode ser tomado, por exemplo, como o desvio padrão dos valores de *background* na hibridação multiplicado por uma constante pré-definida. Procura-se com isso eliminar os *spots* cujos respectivos sinais podem ser apenas resultado de hibridação não específica que, por uma flutuação aleatória, foi maior na área do *spot* que na sua respectiva região de *background*. Outra opção é a eliminação de *spots* cuja razão sinal/ruído fica abaixo de um limite pré-estabelecido. Essa razão é calculada automaticamente por alguns programas de quantificação dos dados (por exemplo o *ArrayVison*, APBiotech, Brasil), e é a razão entre o sinal (subtraído do *background*) obtido no *spot* e o desvio dos sinais obtidos em cada ponto (*pixel*) da respectiva região de *background* (esse desvio é uma medida do ruído aleatório presente nessa região do arranjo).

Alternativamente, em lugar de eliminar os genes que apresentam problemas pode-se apenas etiquetá-los como tais. Com isso tais genes continuam a participar das análises futuras, porém de forma que não interfiram nos resultados dos demais genes. Além disso, caso algum desses genes problemáticos seja selecionado como diferencialmente expresso, a sua etiqueta

nos alerta para a pouca confiabilidade desse resultado, assim só consideraremos o gene como induzido ou reprimido se isso se confirmar em outros experimentos (como *northern blots* ou *real-time PCR*) .

2.3 - Normalização dos dados

Dados provenientes de arranjos de DNA são afetados por diversas fontes de flutuações aleatórias (Schuchhardt *et al.*, 2000), a saber:

- A extração do mRNA pode apresentar variações no rendimento e na qualidade do RNA obtido, principalmente quando o estudo envolve amostras de diferentes tecidos, devido a diferentes taxas de degradação do RNA.

- A reação de transcrição reversa pode apresentar variações nos tamanhos dos cDNAs obtidos.

- A marcação das sondas pode variar aleatória ou sistematicamente, dependendo da frequência de nucleotídeos na composição dos cDNAs.

- A amplificação dos clones por PCR, empregada em alguns casos, pode apresentar grandes variações no rendimento.

- Os pinos empregados para depositar o DNA nos arranjos, componentes do equipamento utilizado na confecção dos mesmos, podem apresentar pequenas diferenças na sua estrutura. Com isso, pode surgir uma variação sistemática nas quantidades de DNA depositadas.

- A quantidade de DNA depositada pode apresentar também flutuações aleatórias, mesmo quando o mesmo pino depositante é empregado.

- A fração do DNA depositado que é efetivamente fixada nos arranjos, ligando-se quimicamente à superfície do suporte do mesmo, é desconhecida.
- A eficiência da hibridação dos arranjos pode variar em função de diferentes parâmetros, como temperatura, tempo, etc.
- As sondas podem ser distribuídas de forma não homogênea sobre os arranjos durante a hibridação. Com isso, esta pode ocorrer de maneira desigual em diferentes partes dos mesmos.
- A hibridação não específica é uma típica fonte de erro que não pode ser excluída.
- Os sinais de pontos de fixação (*spots*) adjacentes podem influenciar no sinal obtido no *spot*, principalmente se são muito fortes.
- A quantificação da imagem obtida pode ser imprecisa devido a saturação de pontos do arranjo ou a variações na forma e na posição dos pontos de fixação do DNA.

Todos esses fatores diminuem a precisão e a confiabilidade dos dados gerados por arranjos de DNA. Assim, para que possamos extrair conclusões válidas desses dados, após a subtração do *background* é preciso adotar procedimentos que eliminem ou minimizem esses efeitos, ou seja, é preciso normalizá-los. Diversas estratégias de normalização para este tipo de dados já foram propostas (Quackenbush, 2002), sendo as mais freqüentemente adotadas descritas a seguir.

A normalização pela média de todos os sinais no arranjo foi a primeira estratégia de normalização proposta para dados de arranjos de DNA. Baseia-se na suposição de que a maioria dos genes presentes nos arranjos apresenta níveis

de expressão constantes nas amostras estudadas (o que só é verdade se esses genes forem selecionados ao acaso, e não genes sabidamente regulados no experimento em estudo). Com isso, espera-se que a média de seus sinais se mantenha constante nas hibridações. Dividem-se então todos os sinais obtidos no arranjo pela média dos mesmos, impondo-se assim que a média dos sinais normalizados seja igual a 1 em cada hibridação. Desta forma, procura-se eliminar tendências globais causadas por diferenças na eficiência da marcação das sondas ou da hibridação das mesmas aos arranjos.

A normalização pode ser feita também pela mediana de todos os sinais no arranjo. O procedimento é equivalente à normalização pela média, mas substituindo-a pela mediana de todos os sinais obtidos em cada hibridação. A mediana é uma medida mais robusta da tendência central dos dados, sendo menos alterada pelos valores extremos dos mesmos, isto é, menos sensível às alterações nos sinais provenientes da porção dos genes em estudo que efetivamente são diferencialmente expressos.

Em alguns experimentos, considera-se que um subconjunto dos genes fixados nos arranjos apresenta níveis de expressão constantes nas amostras analisadas (os chamados genes *house-keeping*). Com isso, pode-se normalizar os sinais de todos os genes dividindo-os, a cada hibridação, pela média ou pela mediana dos sinais desse subconjunto de genes. A desvantagem dessa estratégia é que se corre o risco de a expressão dos genes considerados *house-keeping* não ser tão estável quanto se supunha, e pequenas variações na mesma levarem a uma tendência indesejável nos dados normalizados.

Uma opção adotada em alguns trabalhos é a adição de pequenas quantidades de RNAs exógenos às amostras estudadas, antes da marcação das sondas e da hibridação dos arranjos. Esses RNAs não podem ter homólogos nas amostras, e seus cDNAs também são fixados nos arranjos. Dessa forma, a concentração desses RNAs nas amostras é controlada (como se eles representassem genes de expressão constante) e os seus sinais, de maneira análoga aos genes *house-keeping*, podem ser usados para normalizar os sinais dos genes em estudo. A desvantagem dessa estratégia é que a concentração desses RNAs, bem como a intensidade dos sinais detectados para os mesmos, podem apresentar flutuações aleatórias inesperadas e assim causar tendências indesejáveis nos dados normalizados.

Uma alternativa menos utilizada é a normalização pela emissão de fundo (*background*). Em experimentos onde os genes fixados nos arranjos foram previamente selecionados, por haver evidências de que seus níveis de expressão são diferenciados entre as amostras em estudo, não se pode supor que a média (ou mediana) dos sinais desses genes deveria manter-se constante nas hibridações. Nesses casos, o ideal seria fixar nos arranjos, juntamente com a seleção de genes que se pretende estudar, um conjunto de genes *house-keeping* cujos sinais poderiam ser utilizados na normalização dos dados. Porém, nem sempre isso é feito e, assim, não restam muitas opções de normalização aplicáveis. Uma alternativa quando isso ocorre é utilizar a emissão de fundo (*background*) para normalizar os dados, procurando com isso ao menos eliminar tendências nos dados devido a diferenças na intensidade das imagens captadas das hibridações e em sua quantificação. Seleciona-se uma região do arranjo, a

parte dos pontos de fixação do DNA, onde o *background* global da membrana é quantificado em cada hibridação. Os sinais de todos os genes na respectiva hibridação são então divididos por esse valor.

As estratégias de normalização descritas acima têm em comum a determinação de um fator de normalização por hibridação, que será aplicado a todos os sinais provenientes da mesma. Porém, já foi descrito que em dados de arranjos podem ocorrer tendências sistemáticas dependentes da intensidade dos sinais dos genes (Yang *et al.*, 2002), isto é, os fatores de normalização aplicados precisam variar de um gene para o outro, de acordo com a intensidade dos sinais obtidos. Para determinar esse fator de normalização dependente da intensidade, Yang *et al.* (2002) propôs a utilização do algoritmo conhecido como *lowess* (locally weighted linear regression, Cleveland, 1979). A melhor maneira de visualizar essas tendências dependentes da intensidade é através de um gráfico dos logaritmos das razões de expressão - $\log(\text{sinal da amostra1} / \text{sinal da amostra2})$ - versus os logaritmos das intensidades - $\log(\text{sinal da amostra1} \times \text{sinal da amostra2})/2$ (Figura 2.1, a transformação logarítmica dos dados é discutida abaixo). A aplicação do algoritmo *lowess* na normalização de dados de arranjos consiste em detectar, nesse gráfico, a tendência central dos logaritmos das razões de expressão em função dos logaritmos das intensidades. De acordo com a suposição de que a maioria dos genes não varia entre as amostras em estudo (hipótese necessária à aplicação desta estratégia), seria esperado que essa tendência central fosse constante e igual a 0 ($\log(1)=0$). Assim, a normalização consiste em subtrair de todos os logaritmos das razões de expressão o valor da tendência central, dependente do logaritmo da intensidade, calculado pelo *lowess*.

Obtêm-se com isso valores normalizados dos logaritmos das razões de expressão, que serão utilizados nas análises subsequentes.

Todas as estratégias de normalização descritas acima são aplicadas conjuntamente a todos os genes presentes nos arranjos. Porém, em algumas situações, é conveniente normalizar separadamente os dados provenientes de subconjuntos dos genes depositados (Yang et al., 2002). Por exemplo, quando o equipamento utilizado na confecção dos arranjos emprega diversos pinos para depositar o DNA, pequenas irregularidades nos pinos podem levar a deposição de diferentes quantidades de DNA, e assim é recomendável aplicar as estratégias de normalização separadamente aos dados provenientes dos pontos de fixação depositados por cada pino. Nesses casos, as estratégias de normalização pela média, mediana ou por *lowess* podem ser aplicadas independentemente aos dados provenientes de cada subconjunto de pontos do arranjo. O mesmo vale para a estratégia de normalização por genes *house-keeping*, desde que esses genes estejam presentes em todos os subconjuntos de pontos do arranjo (o que raramente é feito). A estratégia de normalização pelo *background* também poderia ser aplicada independentemente a subconjuntos dos pontos dos arranjos, mas nesse caso seria necessário que estes subconjuntos estivessem separados espacialmente na superfície do arranjo, de maneira que fosse possível determinar um valor de *background* para cada um deles.

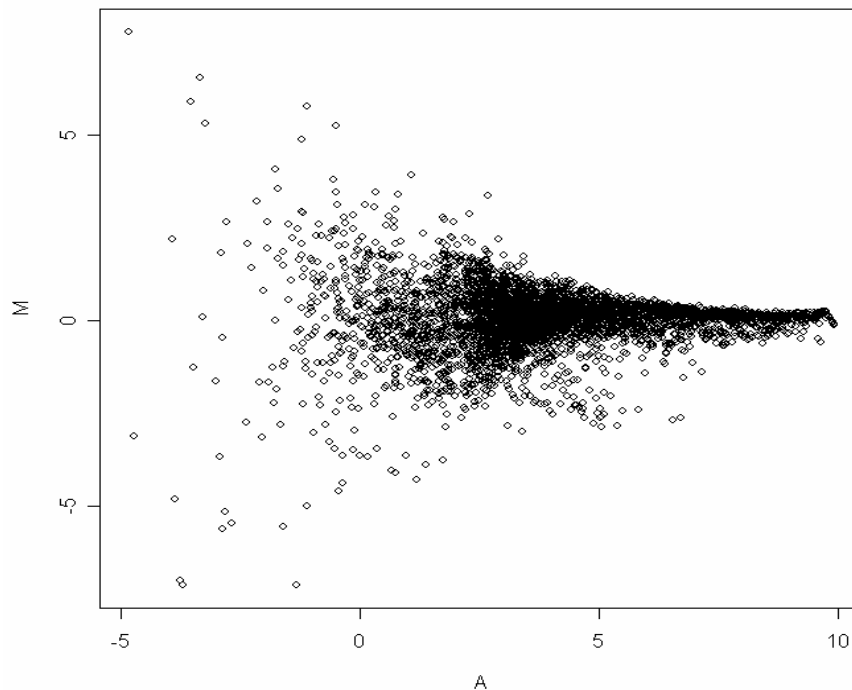


Figura 2.1: dispersão dos logaritmos das razões (M) em função das intensidades médias dos genes (A, também em escala logarítmica).

2.4 - Transformações dos dados

As razões de expressão (sinal na amostra 2/sinal na amostra 1) fornecem uma medida bastante intuitiva do quanto cada gene tem sua expressão alterada entre as diferentes amostras analisadas. No entanto, os valores destas razões não são adequados para as análises estatísticas empregadas na seleção de genes diferencialmente expressos, devido a sua distribuição ser assimétrica e muito influenciada pela intensidade do sinal dos genes. Assim, torna-se necessário aplicar transformações nos dados provenientes de arranjos de DNA, de maneira a obter valores mais indicados para as análises subsequentes.

A grande maioria dos trabalhos com arranjos de DNA emprega a transformação logarítmica nos dados obtidos. Isto se deve a algumas características bastante interessantes desta transformação, a saber:

- A normalização dos valores dos logaritmos, bem como o cálculo de razões, torna-se aditiva e, portanto, mais simples.

- Os logaritmos das razões de expressão são simétricos para genes induzidos e reprimidos. Por exemplo, um gene cuja expressão é duas vezes maior numa amostra proveniente de um tratamento que no seu respectivo controle apresenta um logaritmo da razão de expressão (em base 2) igual a 1, enquanto um gene que é duas vezes menos expresso apresenta esse logaritmo igual a -1 (o que é mais simétrico que os valores brutos das razões de expressão, que seriam respectivamente 2 e 0,5). Além disso, genes com expressão constante nas amostras analisadas apresentam um logaritmo da razão de expressão igual a 0 ($\log(1)=0$), valor mais intuitivo para representar a ausência de resposta do gene ao tratamento em estudo.

- A distribuição dos logaritmos das razões de expressão é aproximadamente normal (Friddle *et al.*, 2000; Hoyle *et al.*, 2002), o que é extremamente interessante para a aplicação de testes estatísticos utilizados na seleção de genes diferencialmente expressos.

A transformação logarítmica pode ser aplicada com qualquer base (vale lembrar que logaritmos em diferentes bases diferem apenas pela multiplicação por uma constante), e é usual a utilização de logaritmos em base 2, por gerarem resultados considerados mais intuitivos.

Alternativas à transformação logarítmica são fornecidas pela família Box-Cox de transformações (Box e Cox, 1964), definidas pela fórmula:

$$BC(x) = \begin{cases} \log(x) & \text{se } \lambda = 0 \\ \frac{x^\lambda - 1}{\lambda} & \text{se } \lambda \neq 0, \end{cases}$$

onde o parâmetro λ é escolhido de maneira a obter valores transformados cuja distribuição seja o mais semelhante possível à distribuição normal (de acordo com a função de verossimilhança). Vale observar que a transformação logarítmica, em base neperiana, pertence à família Box-Cox. Apesar de as demais transformações da família (quando $\lambda \neq 0$) não apresentarem todas as vantagens da transformação logarítmica descritas acima, a seleção de uma dessas transformações, feita automaticamente a cada experimento de acordo com as características dos dados obtidos, é extremamente interessante. Este procedimento permite obter valores transformados com distribuição aproximadamente normal, o que auxilia muito a seleção de genes diferencialmente expressos.

Outras opções de transformações de dados são as transformações estabilizadoras da variância, propostas simultaneamente por dois grupos de pesquisadores (Rocke & Durbin, 2003; Huber *et al.*, 2002). Apesar de a variação dos logaritmos das razões costumar ser menos dependente da intensidade dos sinais dos genes que a variação das próprias razões, ainda assim se observa nos logaritmos uma tendência a apresentarem maior variação para genes de menor intensidade (Mutch *et al.*, 2002). As transformações estabilizadoras da variância procuram evitar essa tendência, tornando a variância dos valores transformados independente da intensidade dos sinais dos genes. Essas transformações são

bem mais complexas e dependem de parâmetros calculados diretamente dos dados, porém têm a grande vantagem da homogeneidade da variância, que facilita bastante a aplicação de testes estatísticos para a seleção de genes diferencialmente expressos.

Capítulo 3: A aplicação de testes estatísticos clássicos à seleção de genes diferencialmente expressos

Os primeiros trabalhos realizados com arranjos de DNA (Schena et al., 1995, Zhao *et al.*, 1995) utilizavam um critério simples para a seleção de genes diferencialmente expressos: aqueles que apresentassem uma razão de expressão acima de um valor pré-fixado arbitrariamente (2 ou 3, na maioria das vezes) eram considerados induzidos, assim com os que apresentassem essa razão abaixo do inverso do limite estabelecido ($1/2$ ou $1/3$, respectivamente) eram selecionados como reprimidos. Com o aumento do número de experimentos realizados utilizando essa metodologia, tornou-se claro que tal critério de seleção era inadequado, uma vez que o valor limite era arbitrário e poderia levar à classificação errônea de muitos genes (dependendo da distribuição das razões de expressão, esse critério poderia ser pouco conservativo ou conservativo demais). Ficou evidente à comunidade científica que era necessário aplicar testes estatísticos aos dados provenientes de arranjos de DNA e, conseqüentemente, que era necessário realizar esse tipo de experimento com réplicas (Herwig et al., 2001), o que não era a regra até então. Os primeiros testes estatísticos adotados foram aqueles já consagrados pelo uso em todos os campos da ciência, como o teste t de Student e derivados da análise de variância.

3.1 - Aplicações do teste t de Student

A análise de dados de arranjos de DNA à procura de genes diferencialmente expressos em geral consiste em comparar, a cada gene, os valores dos sinais obtidos nas hibridações de amostras de RNA provenientes de dois tratamentos (ou tecidos, indivíduos, etc). O teste t é exatamente indicado para a comparação de dois grupos de valores, pareados ou não (diz-se que os dados são pareados se para cada valor em um grupo corresponde um valor no outro grupo), quando se supõe que os valores de cada grupo apresentam distribuições normais com a mesma variância. Vários trabalhos com arranjos de DNA têm usado o teste t, como por exemplo o estudo realizado por Yu et al. (2005), que identificou genes diferencialmente expressos durante o desenvolvimento do gametófito feminino de *Arabidopsis thaliana*, e o trabalho de Salzman et al. (2005), que identificou genes de sorgo regulados por diferentes hormônios.

Na aplicação do teste t aos dados de arranjos de DNA, para cada gene as médias empíricas dos sinais de expressão ou, mais freqüentemente, de seus logaritmos, são calculadas a cada tratamento (por exemplo, chamemos essas médias de m_t e m_c , respectivamente para amostras provenientes de um tratamento e seu controle). As variâncias desses valores também são calculadas (S_t^2 e S_c^2 , respectivamente) e utilizadas para computar o valor do teste, uma distância normalizada entre os dois conjuntos de valores, na forma:

$$t = \frac{(m_t - m_c)}{\sqrt{\frac{S_t^2}{n_t} + \frac{S_c^2}{n_c}}},$$

Em que n_t e n_c são, respectivamente, os números de observações (sinais de diferentes pontos de fixação dos arranjos) em cada grupo. Quando esse valor ultrapassa um determinado limite, de acordo com uma margem de erro pré-fixada (em geral 5%), o gene é considerado como diferencialmente expresso. Vale observar que, como o teste t é aplicado gene a gene, seus resultados não são afetados pelo fato de as variâncias dos sinais detectados serem heterogêneas entre os genes. Por outro lado, o poder desse teste (sua probabilidade de selecionar corretamente um gene diferencialmente expresso) fica reduzido quando os tamanhos amostrais (n_t e n_c na fórmula acima) são pequenos.

Nosso grupo de pesquisa colaborou em dois trabalhos nos quais o teste t foi empregado na análise dos dados de arranjos de DNA. No primeiro (Marques *et al.*, 2004), foram identificados genes preferencialmente expressos na fase patogênica da levedura *Paracoccidioides brasiliensis*. Após a eliminação dos genes com níveis de expressão próximos do *background*, os sinais dos genes restantes foram normalizados pela mediana de todos os sinais na membrana e seus logaritmos (em base 2) foram calculados. O teste t foi aplicado aos valores dos logaritmos, e os genes que apresentaram um valor acima de 3.00 no teste (margem de erro de 1%) foram selecionados como diferencialmente expressos (artigo em anexo).

No segundo trabalho com que cooperamos (Araujo *et al.*, 2005), o teste t foi empregado com um objetivo diferente da seleção de genes diferencialmente expressos: estudar a transcrição de elementos transponíveis em cana-de-açúcar. Para isso, amostras de RNA de calos e de três tecidos diferentes (meristema apical, folhas e flores) foram extraídas e utilizadas na confecção de sondas

radiativas, que foram então hibridadas a macroarranjos em membranas de *nylon* contendo cDNAs para 68 elementos transponíveis que haviam sido identificados em bibliotecas do projeto SUCEST (Vettore et al., 2003). A análise dos dados provenientes dos arranjos procurava comprovar se os elementos transponíveis apresentavam expressão nas amostras de RNA estudadas, assim o teste t para dados pareados foi empregado para verificar se, para cada elemento transponível, a emissão quantificada nos seus pontos de fixação nos arranjos era significativamente maior que a emissão de fundo (*background*) em volta dos mesmos pontos. Os elementos transponíveis que apresentaram um valor acima de 2.92 no teste t (confiança de 95%) foram considerados como expressos na amostra de RNA em questão (artigo em anexo).

3.2 - Aplicações da análise de variância (ANOVA)

Assim como o teste t é freqüentemente utilizado na comparação de dois conjuntos de dados, a procura de diferenças significativas entre as médias dos mesmos, a análise de variância costuma ser aplicada quando mais de dois grupos estão presentes nos dados analisados (desde que se possa supor que cada conjunto tem distribuição normal com a mesma variância). Na verdade, quando restrita a dois grupos a análise de variância é equivalente ao teste t, assim pode-se encará-la como uma versão mais complexa do teste t, que procura identificar diferenças significativas nas médias dos grupos de dados analisados. Kerr et al. (2000) propuseram a aplicação da análise de variância aos dados de arranjos de DNA, apresentando um modelo onde todas as possíveis fontes de variação num

experimento com microarranjos são computadas. Diversos estudos baseados em arranjos de DNA empregaram a análise de variância para identificar genes cuja expressão fora regulada nas amostras de RNA analisadas, como por exemplo o trabalho realizado por Sagi *et al.* (2004) estudando a regulação de genes de tomate (*Lycopersicon esculentum*) em resposta ao silenciamento de genes responsáveis pela "respiratory burst oxidase" (Rboh).

A aplicação da análise de variância aos dados de arranjos de DNA, após a normalização e transformação logarítmica dos mesmos, em linhas gerais consiste em estimar, a cada gene, a variabilidade presente nos subconjuntos de dados provenientes de cada amostra (ou grupo de amostras réplicas) e a variabilidade presente entre as médias desses subconjuntos. Testa-se então se a variabilidade entre as médias é significativamente maior que a variabilidade interna nos subgrupos, de acordo com uma margem de erro pré-fixada (em geral 5%), e se esse for o caso seleciona-se o gene como sendo diferencialmente expresso entre as amostras (ou grupos de amostras) analisadas.

Nosso grupo de pesquisa empregou a análise de variância na análise de dados de arranjos de DNA de um experimento voltado para a identificação de genes de café regulados durante a resposta a infestação por bicho mineiro (*Leucoptera coffeella*). Arranjos em membranas de *nylon* foram confeccionados durante o projeto, contendo genes selecionados através de bibliotecas de subtração por responderem a deposição dos ovos e eclosão das larvas de bicho mineiro nas folhas do cafeeiro (Mondego *et al.*, 2005). Os arranjos foram hibridados com sondas radiativas obtidas a partir de amostras de RNA de duas linhagens de café, uma resistente e outra suscetível ao bicho mineiro, extraídas de

folhas após a oviposição e após a eclosão das larvas do bicho mineiro, bem como de folhas controle não infestadas. Como os genes presentes no arranjo eram pré-selecionados, a única estratégia de normalização que pode ser adotada foi a normalização pelo *background* (descrita no capítulo 2). Após o cálculo dos logaritmos dos sinais normalizados, a análise de variância foi aplicada, a cada gene, para identificar aqueles cuja expressão variava significativamente em resposta a infestação pelo parasita. (artigo em anexo)

3.3 - Correção de Bonferroni

A aplicação de testes estatísticos aos dados de arranjos de DNA gera um problema devido à dimensão dos dados analisados. Um número muito grande de testes é feito em paralelo, já que o número de genes estudado chega a centenas ou milhares, e assim a margem de erro adotada, ainda que pequena, pode levar a um grande número de falsos positivos (por exemplo, se estivermos analisando 2000 genes e adotarmos uma margem de erro de 5% a cada teste, poderemos chegar até a 100 genes erroneamente selecionados como diferencialmente expressos). Uma maneira de evitar que tantos falsos positivos sejam encontrados é diminuir o valor da margem de erro adotada. No entanto, esse procedimento diminui também a capacidade de selecionar genes que verdadeiramente são diferencialmente expressos. A correção de Bonferroni é uma estratégia para controlar a probabilidade global de selecionar incorretamente ao menos um gene, conhecida como *family-wise error rate* (FWER). Ela consiste em adotar como margem de erro para cada teste o valor obtido dividindo-se a margem de erro

global desejada (probabilidade de ocorrer ao menos um erro) pelo número de genes analisado. Esta estratégia pode ser aplicada tanto ao teste t como a análise de variância, e também foi adotada no trabalho que nosso grupo de pesquisa realizou identificando genes de café induzidos durante a infestação por bicho mineiro (artigo em anexo).

E. R. Marques · M. E. S. Ferreira · R. D. Drummond
J. M. Felix · M. Menossi · M. Savoldi · L. R. Travassos
R. Puccia · W. L. Batista · K. C. Carvalho
M. H. S. Goldman · G. H. Goldman

Identification of genes preferentially expressed in the pathogenic yeast phase of *Paracoccidioides brasiliensis*, using suppression subtraction hybridization and differential macroarray analysis

Received: 6 January 2004 / Accepted: 14 April 2004 / Published online: 8 May 2004
© Springer-Verlag 2004

Abstract *Paracoccidioides brasiliensis*, a thermomorphing fungus, is the causative agent of paracoccidioidomycosis (PCM), the most prevalent systemic mycosis in Latin America. Pathogenicity appears to be intimately related to the dimorphic transition from the hyphal to the yeast form, which is induced by a shift from environmental temperature to the temperature of the mammalian host. Little information is available on the *P. brasiliensis* genes that are necessary during the pathogenic phase. We have therefore undertaken Suppression Subtraction Hybridization (SSH) and macroarray analyses with the aim of identifying genes that are preferentially expressed in the yeast phase. Genes identified by both procedures as being more highly expressed in the yeast phase are involved in basic metabolism, signal transduction, growth and morphogenesis, and sulfur metabolism. In order to test whether the observed changes in gene expression reflect the differences between the growth conditions used to

obtain the two morphological forms rather than differences intrinsic to the cell types, we performed real-time RT-PCR experiments using RNAs derived from both yeast cells and mycelia that had been cultured at 37°C and 26°C in either complete medium (YPD or Sabouraud) or minimal medium. Twenty genes, including *AGSI*(α -1,3-glucan synthase) and *TSAl*(thiol-specific antioxidant), were shown to be more highly expressed in the yeast cells than in the hyphae. Although their levels of expression could be different in rich and minimal media, there was a general tendency for these genes to be more highly expressed in the yeast cells.

Keywords *Paracoccidioides brasiliensis* · Dimorphic fungi · Suppression subtraction hybridization · Macroarray analysis · Real-time RT-PCR

Communicated by C P. Hollenberg

E. R. Marques · M. E. S. Ferreira · M. Savoldi
G. H. Goldman (✉)
Departamento de Ciências Farmacêuticas, Faculdade de Ciências
Farmacêuticas de Ribeirão Preto, Universidade de São Paulo,
Av. do Café S/N, CEP 14040-903 Ribeirão Preto, São Paulo, Brazil
E-mail: ggoldman@usp.br
Tel.: +55-16-6024280/81
Fax: +55-16-6331092

R. D. Drummond · J. M. Felix · M. Menossi
Departamento de Genética e Evolução, Instituto de Biologia,
Universidade Estadual de Campinas (UNICAMP),
CEP 13084-971 Campinas, São Paulo, Brazil

L. R. Travassos · R. Puccia · W. L. Batista · K. C. Carvalho
Departamento de Microbiologia, Imunologia e Parasitologia,
Universidade Federal de São Paulo,
CEP-04023-062 São Paulo, Brazil

M. H. S. Goldman
Departamento de Biologia, Faculdade de Filosofia,
Ciências e Letras de Ribeirão Preto, Universidade de São Paulo,
CEP 14040-901 Ribeirão Preto, Brazil

Introduction

The term 'fungal dimorphism' refers to the ability of some fungal pathogens to reversibly switch between yeast (Y) and mycelial (M) forms. *Paracoccidioides brasiliensis*, a thermomorphing fungus, is the causative agent of paracoccidioidomycosis (PCM), the prevalent systemic mycosis in Latin America. Epidemiological data indicate a broad geographic distribution in Central and South America, from Mexico to Argentina (San-Blas and Niño-Veja 2001). It is estimated that as many as ten million individuals may be infected with *P. brasiliensis*. Infection of the human host generally occurs by inhalation of airborne microconidia, which reach the pulmonary alveolar epithelium where they transform into the parasitic yeast form (Restrepo et al. 2001). The dimorphic transition in *P. brasiliensis* occurs simultaneously with changes in cell wall composition involving structural alterations in the carbohydrate polymers, and reorganization of membrane lipids, especially glycos-

phingolipids (GLS) (Leverly et al. 1998, Toledo et al. 1999; Vigh et al. 1998). As the fungus adopts the yeast form, an increase in the chitin content is observed in the cell wall, followed by a change in the anomeric structure of the glucan from a β -1,3-linked polymer to an α -1,3-linked chain (San-Blas and Niño-Veja 2001). The surface α -glucan may act as a protective layer against the host defense mechanisms, in light of the incapacity of host phagocytic cells to digest α -1,3-glucan (San-Blas 1982). The morphological transitions (Y $\rightarrow$ M and vice versa) can be induced in vitro by temperature shifts (see below).

Recently, we reported a survey of the genes expressed in the pathogenic yeast phase of *P. brasiliensis* (Goldman et al. 2003). We obtained 13,490 ESTs from both 5' and 3'-ends of mRNAs, and clustering analysis yielded the partial sequences of 4692 expressed genes, which were functionally classified by similarity to known genes. We have identified several homologues of known *Candida albicans* virulence/pathogenicity factors in *P. brasiliensis*. The transcription of three heat-shock genes (*HSP-70*, *-82* and *-104*), of polyubiquitin (*UBI1*), and of a Δ -9-desaturase homologue (*OLE1*) peaks during the mycelium to yeast (M $\rightarrow$ Y) transition. These results suggested that stabilization of proteins and changes in membrane organization are important factors in the morphological transition to the yeast form at higher temperature.

Pathogenicity appears to be intimately linked to the dimorphic transition (Bialek et al. 2000; San-Blas and Niño-Veja 2001; Borges-Walmsley et al. 2002; Rooney and Klein 2002). In a search for proteins that are differentially expressed in the pathogenic yeast phase of *P. brasiliensis*, da Fonseca et al. (2001) identified and characterized new antigens homologous to catalase, fructose-1,6-biphosphate aldolase, glyceraldehyde-3-phosphate dehydrogenase, malate dehydrogenase, and triosephosphate isomerase. Yeast cells preferentially synthesized these antigens. By using differential display, Venancio et al. (2002) were able to isolate two cDNA fragments that were upregulated during the M $\rightarrow$ Y transition. Their predicted products showed no significant similarity to any of the previously described antigens.

Given the present lack of information regarding *P. brasiliensis* genes expressed in the pathogenic yeast phase, we have used Suppression Subtraction Hybridization (SSH) and differential macroarray analyses to identify genes that are preferentially expressed in the yeast cells. Here we report the characterization of several *P. brasiliensis* genes that are more highly expressed in the pathogenic yeast phase than in the mycelial phase.

Materials and methods

RNA extraction

For macroarray and suppression subtraction hybridization, total RNA was extracted from yeast and mycelium of *P. brasiliensis* isolate 18 grown in agitated liquid cultures to near the late log phase. Yeast cells were

cultivated at 36°C in modified YPD (mYPD, 0.5% yeast extract, 0.5% casein peptone, 1.5% glucose) and mycelium was cultured at 26°C in rich (Sabouraud) medium. Yeast cells were initially recovered from the spleens of intraperitoneally infected mice (the same as those used in Goldman et al. 2003), subcultured from mYPD slants in 100 ml of the same medium for 5 days, then transferred to 1 l of fresh medium and grown for another 7 days. Mycelia were obtained by transferring yeast cells subcultured from mYPD slants to liquid Sabouraud (100 ml) for 20 days at 26°C, then transferred to 1 l of the same medium and grown for 12 days. For RT-PCR validation experiments, total RNA was extracted from yeast and mycelium that had been obtained as just described, but were then grown in parallel in the same culture medium (100 ml). We compared the expression of genes from both phases of *P. brasiliensis* cultured in mYPD, Sabouraud and also in modified minimal medium (MM; Restrepo and Jiménez 1980; L-cysteine, and L-methionine were added at 100 mg/l each). RNA extraction and treatment with RNase-free DNase were carried out as described by Goldman et al. (2003). The absence of DNA contamination after treatment with DNase was verified by PCR amplification of the *PbGP43* gene (Cisalpino et al. 1996). Plasmid DNA extraction and sequencing were performed as described by Goldman et al. (2003).

Suppression Subtraction Hybridization (SSH)

All the SSH experiments were done with the *P. brasiliensis* isolate 18. SSH was carried out using double-stranded cDNA derived from the mycelium phase as the driver, and yeast cDNA as the tester. A PCR Select kit (Clontech) was used as recommended by the manufacturer. Briefly, the tester and driver cDNAs were first digested with *Rsa* I, a restriction enzyme with a 4-bp recognition site that yields cDNA fragments with blunt ends. The tester cDNA fragments were divided into two equal aliquots and each one was ligated to a different adapter, resulting in two distinct populations of tester cDNAs. A small amount of each cDNA tester population (approximately 300 ng) was mixed with an excess of driver (2 μ g), and the mixture was heat-denatured and allowed to anneal for 8 h at 68°C. The two samples from the first hybridization were then combined and annealed with an additional, freshly denatured, sample of driver cDNA at 68°C overnight. An initial PCR using adapter primers was performed to amplify those cDNAs that represent differentially expressed genes. A second PCR amplification was carried out using nested adapter primers to minimize non-specific amplification. The cDNA fragments were fractionated by electrophoresis in a 1% agarose gel, excised from the gel and purified by using the Qiagen Plasmid Purification System. The isolated PCR products were then cloned into the cloning vector pCR2.1 TOPO-(Invitrogen) according to the manufacturer's instructions.

Data handling and analysis

A pipeline was designed to analyze and assemble the *P. brasiliensis* SSH sequences. Sequences were automatically edited for each EST using the programs Phred (Ewing et al. 1998; Ewing and Green 1998), CAP3 (Huang and Madan 1999), Consed (Gordon et al. 1998) and Crossmatch from Phrap (P. Green 1996; available at <http://bozeman.mbt.washington.edu/phrap.html>). Sequences originating from the pUC18 vector were filtered out, and only SSH sequences with a quality value of at least 20 were subjected to further analysis. Edited sequences were clustered using the CAP3 program (Huang and Madan 1999). To identify potential homologues to *P. brasiliensis* genes, the BLASTx and BLASTn algorithms (Altschul et al. 1997) were used to compare the predicted protein sequences with those in the NCBI nonredundant and *P. brasiliensis* EST project databases (<http://143.107.203.68/est/default.html>), respectively. The values presented for putative homologues derive from results obtained with sequences of known function as well as those of unknown function. The database sequence matches (putative homologues) were divided into highly significant (e values $\leq 10^{-19}$), moderately significant ($1e-10$ to $1e-19$), weakly significant ($1e-03$ to $1e-10$), and non-significant ($> 1e-03$) classes (data not shown). When e values higher than 10^{-3} were obtained, they were considered not to be statistically significant (no significant match).

Preparation and analysis of cDNA macroarrays

cDNA inserts representing 1397 *P. brasiliensis* genes (Goldman et al. 2003) were amplified by PCR using the M13 reverse and forward primers. The inserts were denatured with 0.2 M NaOH for 15 min at 37°C and then spotted onto Hybond-N filters (AP-Biotech) with a hand-held 96-pin print head tool (V&P Scientific). This tool typically deposited 0.1 μ l of DNA solution, which corresponded to approximately 5 ng of DNA. The DNA was fixed to the filters by baking at 80°C for 2 h, and the filters were then stored at room temperature until used for hybridizations. Each filter carried 96 clones in quadruplicate, giving a total of 384 spots per filter.

cDNA targets were obtained from total RNA isolated from mycelium and yeast phases, as described by Schummer et al. (1999), using [α - ^{32} P]dCTP. Filters were hybridized, washed and exposed to imaging plates for 96 h, and then scanned in a phosphorimager FLA3000-G (Fujifilm). The radioactive signals were completely removed from the membranes (Sambrook and Russell 2001), and those that had been hybridized with radioactive yeast cDNA were probed with radioactive mycelium cDNA, and vice-versa. The signals were evaluated again, and compared to those observed in the previous hybridization experiment. Signals were quantified using Array vision software (Imaging Research). The local background was subtracted automatically from each

spot. The intensities of the four replicate spots were averaged and used to calculate the expression ratios between yeast and mycelium. To reduce the fluctuations among replica filters due to differences in the experimental conditions, on each membrane, the signal from every spot was normalized using the median of all signals on that membrane; i.e. the median was set to 1 (Schummer et al. 1999). Genes with expression levels close to the background value were discarded and 536 genes (out of 1397) were used for further analysis. Differentially expressed genes were identified with Student's t -test using the $\log_2$ of normalized values from two replicated experiments. A cutoff value of 3.00 for the t -test was chosen ($P < 0.01$).

PCR and real-time RT-PCR

All PCRs and RT-PCRs were performed using an ABI Prism 7700 Sequence Detection System (Perkin-Elmer Applied Biosystems). Taq-Man EZ RT-PCR kits (Applied Biosystems) were used for RT-PCR reactions. The thermocycling conditions comprised an initial step at 50°C for 2 min, followed by reverse transcription for 30 min at 60°C and denaturation at 95°C for 5 min, and 40 cycles at 94°C for 20 s and 60°C for 1 min. Taq-Man PCR Reagent kits were used for PCRs. The reaction parameters were as follows: after initial incubation at 50°C for 2 min, and 10 min at 95°C, 40 cycles of 95°C for 15 s and 60°C for 1 min were performed. Since there is no ideal control for gene expression, we first compared the α -tubulin and β -tubulin probes as normalizers for the expression of these genes. We saw no difference between these two normalizers. Accordingly, the calibrator gene used for the expression experiments was the α -tubulin gene (data not shown). The reactions and calculations were performed according to Semighini et al. (2002). Primer and probe sequences are listed in Table 1.

Results

Use of macroarrays to identify differentially expressed genes

Differential macroarray analysis was used as a first approach to isolate genes that are preferentially expressed in the yeast phase of *P. brasiliensis*. PCR fragments of representative clones from 1397 contigs clustered from 8369 ESTs (5' and 3' ends) previously described by Goldman et al. (2003) were spotted on nylon membranes (two copies of each membrane and four repetitions for each clone). Yeast and mycelium cDNA probes were labeled, and equal amounts of radioactively labeled samples were separately hybridized to duplicate nylon membranes. The signals were converted to relative intensity values, quantified and statistically analyzed. Here we will describe only the genes that were more highly expressed in the yeast phase. Figure 1 shows the

Table 1 List of primers and fluorescent probes used in this work

Primers and probes ^a	Sequence (5' to 3')	<i>P. brasiliensis</i> Cluster ID	Locus name
Alpha	CCAGAACCAGGCAGTCCAAA	SSHCS001-E11	<i>AGS1</i>
P-alpha ^b	FAM-CACCTGCCTAACAAGATTGGACCAGG5 G	SSHCS001-E02	
H82f	ACCGCTGGTGCCGATATT	PbCC001-B11	<i>HSP82</i>
H82r	GGCCACGAGGTAGGCAGAG		
PH82	TET-TGATTGGTCAGTTTGGTGTGCGCTTCT-TAMRA		
APS1F	ACTAATCCCGGTAAACTCCTTCACA	RC2-Pb5NEW1-154G11	<i>APS1</i>
APS1R	CGGAGCAGAGAGACCCAAAG		
P-APS1	FAM-TCGCCTTCTTATACAGCCC-TAMRA-		
c101-F	AAGCCTTCTCCCCACCAT	PbCC002-A05	<i>C101</i>
c101-R	GCCAAAACATAGCCCACGA		
Pc101	FAM-TGGCTTGTTCGAAAGA-TAMRA		
c126F	CGTTGGTAATCCTGGCTTGG	PbCC002-C06	<i>GSP1</i>
c126R	CAACTGCGGTCTCAGGTGG		
Pc126	FAM-ATTCGTTGCCGATATTGCCCTCGC-TAMRA		
C1363F	ATTTGAGGGAGGCGGAATCC	Contig 1363 RST	<i>MEP1</i>
c1363R	TTTGGGAACGTCCAGTCTGTT		
Pc1363	FAM-CTGCCCTGTACTGACCTC-TAMRA		
c1412F	TGTCTGCAGGGCATCGATTA	PbCC015-F08	<i>TSA1</i>
c1412R	AGCTACCCTGCATCCACTGG		
Pc1412	FAM-TCGCAGCACCTCAGCAGCATTCC-TAMRA		
c162F	CTCGACAACGTCCAGGAGAAG	PbCC002-F06	<i>RHO1</i>
c162R	GGGATGACCCTGGCAGAAATG		
c162	FAM-AAGGACTTCAGAAATCCA-TAMRA		
C165F	TCCCCTATCATCCTGCAAGTGT	PbCC002-F09	<i>FBA1</i>
c165R	CCATTGGGAACGCCCTTC		
Pc165	FAM-CAAAGAAAGCAGCCCC-TAMRA		
C261F	CCTACGCGACATGCAGAAATATATC	PbCC003-F09	<i>LUC7</i>
c261R	CGCTGCGCCGAGTCTATC		
Pc261	FAM-CTCCGATTACAGTCATC-TAMRA		
c32F	GGCAAAGCCGAAGAAAAGC	PbCC001-C08	<i>ILV5</i>
c32R	TGTACAGGTAGCCGCTGCC		
Pc32	FAM CGGCATCGCCATC-TAMRA		
C354F	CCTACAATATTTCTGGCTCGTTTCCT	PbCC004-F06	<i>PAL1</i>
c354R	GGATTGCTCAACAACCTCCTCAT		
Pc354	FAM-ACTCCGGTAATTTTCG-TAMRA		
c5F	ACGACCACGGAAACTCATACC	PbCC001-A04	<i>ERG25</i>
c5R	GGATCATGCTGAGGCTTTTCCA		
Pc5	FAM-CTGTGCGCATCAATG-TAMRA		
c71F	ACCGTTGCCAGCACTT	PbCC001-F11	<i>CIT1</i>
c71R	AGCCCTCCCAGACCATAGACTT		
Pc71	FAM-AGGCCTCTCATTCTC-TAMRA		
c77F	GGAGCCGGAAGTCGAATAGG	PbCC001-G05	<i>C77</i>
c77R	ACCTACCATTGGAATGTGTGCATA		
Pc77	FAM-CCAAGACTCCAATGGT-TAMRA		
C836F	AACGGTGTTCAGACGATGGT	PbCC009-F08	<i>PAL2</i>
c836R	AACCACCACGACGACCAG		
Pc836	FAM-ACCTCGGCAACATC-TAMRA		
CHS1F	GAATGGGCTTCCGAGATAAATGC	VP1-Pb50001-178A02	<i>CHS1</i>
CHS1R	ACCGACACCAACACATCTCAA		
P-CHS1	FAM-CCGACATCCAAAGCCT-TAMRA		
COT1F	CCTGCTTCATATCCTCATCATGGA-	PbCC003-A05	<i>CTR3</i>
COT1R	TGACGCCTATACATGACCCTACAA-		
P-COT1	FAM-CACGTCCAGAGGAATGT-TAMRA-		
FLBA-F	CGGCTAGCAAGACTGTTTCG-	PbCC002-F06 162	<i>FLB1</i>
FLBA-R	TGCTTTTCTTGCTCCTGGATCTC-		
P-FLBA	FAM-CCTGCGAGCTGAACAT-TAMRA-		
GCL1F	GCATCAGCATCAGCAGCAA	RC2-Pb30001-119B07	<i>GCL1</i>
GCL1R	CTGCTTGTGCTGTTGGTATTGG		
P-GCL1	FAM-CAGCACGTCAACCCCA-TAMRA		
Tub2F	AGCCTTGCGTCGGAACATAG	PbCC014-D01	<i>TUB2</i>
Tub2R	ACCTCCATCCAGGAACCTTCA		
Ptub2	FAM-AGCGTGTAGGTGATCAG-TAMRA		
CDI1F	TGGCATGGTCGTGGATAGC	PbCC004-C05	<i>CDI1</i>
CDI1R	AGGAAACGGCAAAAGCAATCTG		
PCDI1	FAM-AGCGTGTAGGTGATCAG		
Tub1	ACCCAGCTTGGAACAGTGCT	PbCC001-E05	<i>TUB1</i>
Ptub1**	CACGGACCATCAGCCTTCAAACC5G		

Table 1 (Contd.)

Primers and probes ^a	Sequence (5' to 3')	<i>P. brasiliensis</i> Cluster ID	Locus name
C159F	GCCCCTGGAGGATAAGATAAAGAG	PbCC002-F03	<i>SEP!</i>
C159R	GAGCGAACGAATCGACTTTACTGTA		
PC159	FAM-CTAACGAAGAGAGAATCCCT-TAMRA		

^aP' primers were fluorescently labeled with the indicated markers (TET, 6-carboxy-4,7,2',7'-tetrachlorofluorescein; 6-FAM, 6-carboxy-fluorescein; TAMRA, 6-carboxy-*N,N,N',N'*-tetramethylrhodamine)

^bProbe Lux (Invitrogen)

scatter correlation graph for 536 genes (out of 1397) that were expressed at levels above the background value. We were able to identify 38 clusters whose expression levels were 1.5–6.6 times higher on the membrane hybridized with the yeast cDNA probe than on that hybridized with the mycelium cDNA probe (Table 2). Homology searches revealed that 26 of these clusters yielded highly significant *e* values ($\leq 10^{-19}$), while seven showed moderately (1e-10 to 1e-19), and five showed weakly (1e-03 to 1e-10) significant *e* values with sequences listed in databases. It is noteworthy that most members of the latter two classes show similarity to hypothetical proteins from *Neurospora crassa* and *Aspergillus nidulans*.

Use of SSH to select preferentially expressed genes

We also performed an SSH analysis using mycelium cDNAs as driver and yeast cDNAs as tester with the aim of isolating genes preferentially expressed in the *P. brasiliensis* yeast phase. When the original cDNA pools were used as probes on Northern blots loaded with yeast and mycelium mRNAs, similar hybridizing signals were detected, while the subtracted cDNAs used as a probe yielded stronger and diffuse hybridizing signals with the yeast mRNA (data not shown). The SSH procedure was also validated using real-time PCR to quantify the expression of the *HSP82* gene, which was previously shown to be overexpressed in the yeast phase (Goldman et al. 2003). *HSP82* RNA was enriched at least 150 fold in the subtracted cDNAs relative to the original cDNA pools (data not shown).

Since our subtracted cDNA population was enriched for cDNA fragments representing mRNAs from the yeast phase, we constructed a library from the SSH subtracted cDNAs and sequenced 750 ESTs whose inserts ranged from 200–700 bp in size. The sequences yielded 240 clusters following analysis with CAP3 (112 contigs and 128 singlets). These sequences were compared with nonredundant NCBI proteins using BLASTx, and with our previous EST collection from *P. brasiliensis* (Goldman et al. 2003) using BLASTn. Our analysis revealed 163 novel clusters (50 with known and 113 with unknown function), while 79 (46 with known and 33 with unknown function) were already represented in our database (Goldman et al. 2003). The most highly expressed sequences were screened by spotting one representative PCR-amplified insert from each of the 240 SSH clusters on nylon membranes and

hybridizing them with radioactively labeled subtracted cDNA. The inserts that corresponded to the genes showing the most intense signals, as well as those highly represented in the library (multiple clones bearing fragments of the same gene), were selected for further expression analysis by real-time RT-PCR (see Table 3). Some of these, such as *ERG25*, *CIT1*, *FBA1*, *ILV5*, *CTR3* (related to cell metabolism) and *CDI1* (related to sulfur metabolism) were also detected by differential macroarray analysis (Table 2). Apart from the clusters listed in Table 3, the SSH library contained homologues of the *HSP10*, -30, -70, -82, and -88 genes, which are important in the adaptation of microorganisms to temperature changes. The *HSP70* heat shock gene (da Silva et al. 1999) was also selected as being more expressed in the yeast phase by the macroarray technique (Table 2). Other potentially important gene homologues found in the SSH library, which were not present in our *P. brasiliensis* database, were: *MNN4* (related to mann-6-phosphorylation), *PKC1* (coding for a cAMP-dependent protein kinase) and *GST1* (coding for a glutathione-S-transferase). All these gene homologues showed highly significant *e* values ($\leq 10^{-19}$), with the exception of *AGS1* (*e*-value = 3e-16), which showed a moderately significant *e* value (1e-10 to 1e-19). However, we have isolated the same gene homologue in a cluster derived from a random sequencing tag collection

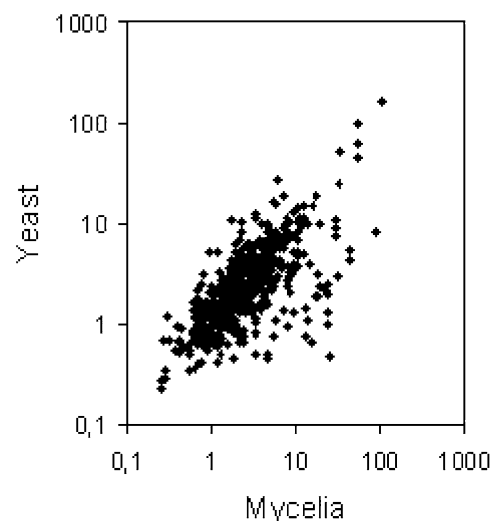


Fig. 1 Scatter plot of gene expression in yeast and mycelium. Points represent the average normalized signal intensity (arbitrary units) for 536 *P. brasiliensis* genes

Table 2 Macroarray analysis of contigs more highly expressed in the pathogenic yeast phase than in the mycelial phase of *P. brasiliensis*

<i>P. brasiliensis</i> Cluster ID	Homologue (function, source) and NCBI protein accession number	BLAST e-value (% identity/similarity)	Induction factor
PbCC001-A04	<i>ERG25</i> (C-4 methyl sterol oxidase, <i>S. cerevisiae</i>) NP_011574.1	8e-53 (60/78)	6.6
PbCC001-C03	<i>CIT1</i> (citrate synthase, <i>A. nidulans</i>) AAC49728.3	9e-61 (89/94)	6.0
PbCC002-A05	<i>FBA1</i> (aldolase, <i>S. cerevisiae</i>) NP_012863.1	e-136 (65/76)	4.7
PbCC005-A06	<i>ILV5</i> (Branched-chain amino acid biosynthesis, <i>S. cerevisiae</i>) AAB67753.1	2e-82 (86/91)	4.4
PbCC014-C11	Predicted protein (<i>N. crassa</i>) NCU06390	6e-05 (43/61)	3.6
PbCC003-F04	Hypothetical protein (<i>N. crassa</i>) NCU05891.1	6e-07 (36/49)	3.3
PbCC003-B12	Hypothetical protein (<i>N. crassa</i>) NCU04080	2e-10 (34/48)	3.0
PbCC003-E12	<i>ENO1</i> (enolase, <i>A. oryzae</i>) Q12560	0.0 (86/91)	2.9
PbCC001-F06	<i>HIS4</i> (histone H4, <i>A. nidulans</i>) P23751	2e-35 (100/100)	2.8
PbCC001-D02	<i>EEF3</i> (translation elongation factor eEF3, <i>C. albicans</i>) S25363	e-103 (74/89)	2.7
PbCC004-A08	<i>RL20</i> (60S ribosomal protein L20, <i>S. cerevisiae</i>) NP_014957.1	8e-67 (71/83)	2.5
PbCC004-H02	<i>A. nidulans</i> EST c4c08a1.r1	4e-04 (86/92)	2.5
PbCC005-C08	Hypothetical protein (<i>N. crassa</i>) AL513466	2e-18 (30/44)	2.4
PbCC001-A10	<i>FCF1</i> (zinc cluster transcription factor Fcrlp, <i>C. albicans</i>) AF057038	1e-15 (64/76)	2.3
PbCC001-C10	Unknown		2.2
PbCC004-A04	Probable glutaredoxin (<i>N. crassa</i>) AL353819	6e-17 (52/67)	2.1
PbCC004-A09	Unknown		2.0
PbCC004-A10	<i>GCV3</i> (glycine decarboxylase, <i>A. thaliana</i>) AC004667	8e-30 (53/69)	2.0
PbCC004-C10	Aminopeptidase B (<i>A. niger</i>) AJ418036	2e-52 (69/80)	2.0
PbCC003-G12	Inositol-1-phosphate synthase (<i>P. pastoris</i>) AF078915	1e-73 (48/63)	2.0
PbCC007-D02	Predicted protein (<i>N. crassa</i>) NCU02545.1	6e-11 (32/52)	1.9
PbCC001-H03	Unknown		1.9
PbCC003-E01	Unknown		1.8
PbCC003-C05	Unknown		1.8
PbCC002-C05	Putative ribosomal protein (<i>A. fumigatus</i>) AL807577	9e-56 (84/94)	1.8
PbCC008-C11	60S ribosomal protein L27A (L29) Y11394	7e-70 (79/91)	1.8
PbCC001-A12	Unknown		1.8
PbCC003-G06	Conserved hypothetical protein with HesB-like domain (<i>S. pombe</i>) NC_003423	9e-18 (48/64)	1.7
PbCC001-D03	Hydroxyproline-rich glycoprotein precursor <i>Z. mays</i> S45164	6e-08 (41/44)	1.7
PbCC005-D04	Unknown		1.7
PbCC001-A06	Probable snRNP protein (<i>N. crassa</i>) AL513466	6e-36 (71/81)	1.7
PbCC006-A01	Unknown		1.7
PbCC004-C05	<i>CD11</i> (cysteine dioxygenase, <i>C. elegans</i>) U51993	2e-36 (48/65)	1.7
PbCC007-B07	<i>LUC7</i> (RNA splicing factor Pad-1, <i>N. crassa</i>) CAD21082.1	3e-66 (73/77)	1.6
PbCC006-E05	Predicted protein (<i>N. crassa</i>) NCU07465.1	1e-06 (39/53)	1.6
PbCC003-A05	<i>CTR3</i> (high affinity copper transporter protein, <i>S. cerevisiae</i>) NC_003421	6e-20 (41/57)	1.6
PbCC006-A03	<i>HSP70</i> (heat shock protein 70) X98931	2e-23 (91/99)	1.5
PbCC004-D04	Hypothetical protein (<i>Mesorhizobium loti</i>) NC_002678	5e-10 (35/52)	1.5

(Nobrega and Goldman, manuscript in preparation). Partial sequencing of this gene confirmed that it indeed encodes a putative homologue of the α -glucan synthase gene from *A. fumigatus* (e-value = 5e-25; 70% identity with α -glucan synthase from *A. fumigatus*).

Use of real-time RT-PCR to validate the expression of the genes identified by macroarray and SSH analyses

For real-time RT-PCR analysis, we designed Taq-Man and Lux probes (Tables 1, 2 and 3): 13 for gene homologues selected from the SSH library (Table 2) and 4 from the macroarray analysis, i.e., *C77*, *C101*, *FBA1*, and *LUC7* (Table 3). The selection was based on the degree of expression in the yeast phase in the preliminary selection, but we also included four probes for *P. brasiliensis* gene homologues involved in sulfur metabolism (*MEP1*, *APS1*, *CHS1* and *GCL1*). Although these last genes were not selected in our primary screen, they have been shown to influence the morphological state of

Histoplasma capsulatum and other dimorphic fungi (Maresca and Kobayashi 2000; Hwang et al. 2003).

The total RNA preparations used in the initial screening for both SSH and macroarrays had been isolated from yeast cells grown in mYPD medium and from mycelia grown in Sabouraud medium. To check whether changes in the expression levels of the genes might reflect differences in the growth conditions used to maintain the morphological forms rather than functions required in one or other of the two distinct cell types, validation by RT-PCR experiments was carried out using a second batch of RNAs derived from yeast and mycelium that had been grown in the same medium at 37°C and 26°C, respectively. We compared the expression of these genes at the two temperatures in mYPD, Sabouraud and minimal media.

The results were expressed as the ratio of the transcript number in yeast to that in mycelium, with the transcript copy number in mycelium being set to 1 (Table 3). Our results suggest that many of the genes listed in Table 4 may be involved in the maintenance of

Table 3 *P. brasiliensis* genes chosen for the real-time RT-PCR experiments

<i>P. brasiliensis</i> Cluster ID(EST cluster)	Homologue (NCBI protein accession number)	BLAST e-value (% identity similarity)	Criteria for choosing the gene for RT-PCR analysis
Metabolism			
SSHCS001-G09 (PbCC010-B11)	<i>ILV5</i> (branched-chain aminoacid biosynthesis, <i>S. cerevisiae</i>) AAB67753.1	3e-20 (73/84)	4.4-fold induction in the macroarray; > 3 SSH fragments > 1.6-fold induction in the macroarray > 3 SSH fragments; > 6.0-fold induction in the macroarray Strong signal in the SSH hybridization
PbCC007-B11	<i>LUC7</i> (RNA splicing factor Pad-1, <i>N. crassa</i>) CAD21082.1	3e-66 (73/77)	
SSHCS001-A02 (PbCC012-B10)	<i>CIT1</i> (citrate synthase, <i>A. nidulans</i>) AAC49728.3	9e-61 (89/94)	
SSHCS001-B04 (PbCC003-A05)	<i>CTR3</i> (high-affinity copper transporter, <i>S. cerevisiae</i>) NC_003421	6e-20 (41/57)	
SSHCS001-E11 (PbCC002-C06)	<i>GSP1</i> (small G-protein Gsp1p, <i>C. albicans</i>) AF190700	2e-94 (79/82)	> 6.6-fold induction in the macroarray; Strong signal in the SSH hybridization Strong signal in the SSH hybridization Strong signal in the SSH hybridization > 4.7-fold induction in the macroarray > 1.8-fold induction in the macroarray > 2.2-fold induction in the macroarray
SSHCC001-A06 (PbCC001-A04)	<i>ERG25</i> (C-4 methyl sterol oxidase, <i>S. cerevisiae</i>) NP_011574.1	8e-53 (60/78)	
SSHCS001-C08 (PbCC004-F06)	<i>PAL1</i> (alkaline serine protease Penc2, <i>Penicillium citrinum</i>) AF098517_1	e-157/64/77	
SSHCC001-A12 (PbCC009-F08)	<i>PAL2</i> (aspartyl aminopeptidase, <i>Coccidioides immitis</i>) AF420303	7e-256 (75/89)	
SSHCC001-D05 (not present) ^a	<i>AGS1</i> (putative α -1,3-glucan synthase, <i>A. fumigatus</i>) AAL18964.1	3e-16 (67/78)	> 4.7-fold induction in the macroarray > 1.8-fold induction in the macroarray > 2.2-fold induction in the macroarray
PbCC002 A05 (<i>FBA1</i>)	<i>FBA1</i> (aldolase, <i>S. cerevisiae</i>) NP_012863.1	e-136 (65/76)	
PbCC003-C05 (<i>C101</i>)			
PbCC001-C10 (<i>C77</i>)			
Signal transduction			
SSHCS002-C06 (PbCC002-F06)	<i>RHO1</i> (GTPase Rho1 <i>A. fumigatus</i>) AY007297	4e-61 (99/99)	Strong signal in the SSH hybridization
Growth and morphogenesis			
SSHCC001-H01 (PbCC002-F03)	<i>SEPI</i> (septin B, <i>A. nidulans</i>) AAB41233.1	6e-63 (92/95)	Strong signal in the SSH hybridization
SSHCS002-A02 (PbCC006-E07)	<i>FLB1</i> (developmental regulator <i>flbA</i> , <i>A. nidulans</i>) P38093	2e-38 (77/86)	Strong signal in the SSH hybridization
Oxidative stress			
SSHCS001-G01 (PbCC015-F08)	<i>TSA1</i> (Putative thiol-specific antioxidant protein, <i>Ajellomyces capsulatus</i>) AAG31645.1	7e-33 (89/95)	Strong signal in the SSH hybridization
Sulfur metabolism			
RST1363 ^b	<i>MEP1</i> , Methionine permease		Sulfur metabolism
PbCS004-F04	<i>APS1</i> , Adenylsulfate kinase		Sulfur metabolism
PbCS006-F07	<i>CHS1</i> , Choline sulfatase		Sulfur metabolism
PbCS001-E12	<i>GCL1</i> , Glutamate-cysteine ligase		Sulfur metabolism
SSHCS001-F05 (PbCC004-C05)	<i>CDII</i> (cysteine dioxygenase, <i>C. elegans</i>) U51993	2e-36 (48/65)	1.7-fold induction in the macroarray; > 3 SSH fragments; strong signal in the SSH hybridization; sulfur metabolism

^aA contig derived from a random sequencing tag collection (Nobrega and Goldman, manuscript in preparation)

the *P. brasiliensis* yeast phase in vitro, because they were more highly expressed in the yeast cells grown under three different conditions in different media. The unknown cluster *C101* was not validated as a yeast-specific gene in our experimental conditions, since the ratio of yeast/mycelium ratio was consistently less than 1. Other gene homologues yielded low ratios or ratios near 1 in one or two culture conditions, as seen for *ILV5*, *CIT1*, *GSP1*, *PAL2*, *FBA1*, *FLB1* and *GCL1*. The most impressive results were seen with the homologues of α -1,3-glucan synthase (*AGS1*) and choline sulfatase (*CHS1*), which gave very high yeast/mycelium ratios under all three conditions studied, up to 82.18 and 61.10 times, respectively more highly expressed in the yeast phase (in mYPD). The *LUC7* homologue (predicted to encode an RNA splicing factor) was also abundantly expressed in the yeast phase in all three media (ratios above 6), followed by *PAL1*, which encodes a serine proteinase. The ratios were comparable in at least two conditions for *ERG25*, *FBA1*, *RHO1*, *SEPI*, *MEPI*, *GCL1* and *CDII*. As previously shown by da Fonseca et al. (2001), we also see increased expression of the *FBA1* and *ENO1* genes that code for fructose-1,6-bisphosphate aldolase and enolase, respectively, during the yeast phase.

It is interesting to note that some genes showed much higher yeast/mycelium ratios when the cells were grown in minimal medium than in rich media. This was the case for *ILV5*, *CIT1*, *C77*, *RHO1*, *FLB1* and *MEPI*. The *RHO1* gene, whose product is involved in signal trans-

duction, and *MEPI*, involved in sulfur metabolism, were from 4 to 15 times more highly expressed in minimal medium than in rich medium (Table 4). The fluctuations in the ratios probably reflect different growth stages and responses to diverse stimuli in the culture environments (Hwang et al. 2003).

Discussion

The experiments reported here represent the first application of high-throughput SSH and macroarray analysis to examine gene expression in the pathogenic phase of *P. brasiliensis*. Both procedures enabled us to identify genes that are preferentially expressed in the yeast phase. Recent studies have demonstrated that dimorphic switching is directly associated with pathogenicity (Braun and Johnson 1997; Csank et al. 1998; Leberer et al. 1997; Lo et al. 1997; Stoldt et al. 1997; Yaar et al. 1997; Borges-Walmsley et al. 2002; Romani et al. 2003). Accordingly, the identification of genes preferentially expressed in the yeast phase can be expected to provide information about the pathobiology of *P. brasiliensis*. The SSH strategy overcomes the problem of differences in mRNA abundance by incorporating a hybridization step that normalizes (equalizes) sequence abundance during the course of subtraction by standard hybridization kinetics (Diatchenko et al. 1996). This advantage is reflected in the fact that most of the SSH clusters obtained here (about 68%) were not already present in

Table 4 Relative transcript levels of *P. brasiliensis* genes measured by real-time RT-PCR in different conditions

<i>P. brasiliensis</i> gene Cluster ID (EST cluster)	Relative transcript level in yeast phase ^a		
	Both yeast and mycelium grown in YPD	Both yeast and mycelium grown in Sabouraud medium	Both yeast and mycelium grown in minimal medium
Metabolism			
<i>ILV5</i>	2.95	0.70	5.60
<i>LUC7</i>	27.77	7.02	6.70
<i>CIT1</i>	1.06	3.26	7.30
<i>CTR3</i>	7.37	4.36	2.00
<i>GSP1</i>	3.98	1.40	1.20
<i>ERG25</i>	2.45	2.37	2.00
<i>PAL1</i>	5.07	6.84	6.30
<i>PAL2</i>	4.27	1.68	1.00
<i>AGS1</i>	63.2	82.18	9.20
<i>FBA1</i>	2.7	3.04	1.40
Unknown			
<i>C77</i>	5.09	2.35	11.80
<i>C101</i>	0.06	0.56	0.60
Signal transduction genes			
<i>RHO1</i>	3.17	2.28	30.5
Growth and Morphogenesis			
<i>SEPI</i>	3.75	3.77	3.00
<i>FLB1</i>	1.91	0.94	3.30
Oxidative stress			
<i>TSA1</i>	18.81	7.51	2.70
Sulfur metabolism			
<i>MEPI</i>	4.23	5.62	29.30
<i>APS1</i>	8.95	4.48	3.00
<i>CHS1</i>	61.10	14.52	26.00
<i>GCL1</i>	1.19	2.38	2.00
<i>CDII</i>	2.07	1.81	3.00

^aThe transcript levels of *P. brasiliensis* genes in the yeast phase (YP) were quantitated by RT-PCR, and expressed relative to those of the same transcripts in the mycelial phase (MP). The measured quantity of each mRNA in each sample was normalized using the *C_t* s obtained for the α -tubulin RNA run on the same plate. The relative levels of each transcript and α -tubulin mRNA were determined from a standard curve (i.e., *C_t* s plotted against logarithm of the DNA copy number). The results are averages of three repetitions. The values represent the ratio of the level in YP to that in the mycelial control (set to 1.00)

our EST databank (Goldman et al. 2003). In our hands, both SSH and macroarray analysis proved to be powerful and reliable, since most of the 21 genes selected for validation by real-time RT-PCR were indeed more highly expressed in the yeast phase than in the mycelial phase. Furthermore, several genes were identified by both methods.

We sorted the genes preferentially expressed in the yeast phase into broad functional categories. Based on the function of their orthologues, we have identified genes that are implicated in several cellular processes. A more extensive analysis of these genes will be necessary before their functions can be unequivocally determined. Classical genetic tools, such as DNA-mediated transformation and modulation of gene expression either by deletion/replacement or inactivation by RNA interference or antisense RNA, have not yet been developed for *P. brasiliensis*. The type of genomic approach described here represents an efficient alternative strategy for identifying genes involved in virulence/pathogenicity, and may provide a wider platform for a better understanding of *P. brasiliensis* biology and the role of specific genes during the dimorphic transition and the onset of virulence/pathogenicity in PCM. As soon as gene inactivation techniques for this organism have been developed, the function of the genes identified may be tested directly.

The correlation between the expression of specific genes in the yeast form and the morphological phase is complicated by the fact that a temperature shift must be used to promote the dimorphic transition. In *P. brasiliensis* no attempts have yet been made to induce the switch to the yeast phase at room temperature. Therefore it is not yet possible to follow the expression of specific genes in cultures that have not been submitted to a temperature shift. The growth conditions used to obtain the yeast probe also differ from those used for the source of the mycelial probe, and these differences could themselves affect gene expression. Hence it is possible that genes that are preferentially expressed in the yeast state are not necessarily connected with morphological differentiation, but are required for physiological adaptation to the new environment or related to the growth phase in vitro (Berman and Sudbery 2002; Hwang et al. 2003). We addressed the possible influence of growth conditions by studying the expression of the identified genes in different culture media. In general, although the levels of expression of some of genes, such as *AGS1* and *LUC7*, could be strikingly different in complete and minimal media, there was a consistent tendency for these genes to be more highly expressed in the yeast phase, supporting the idea that most of them are specific for that cell type.

The most obvious biochemical alteration in *P. brasiliensis* as the fungus adopts the yeast form is the switch from β -1,3- to α -1,3-glucan chains in the cell wall (San-Blas and Niño-Vega 2001). Accordingly, we observed a much higher level of expression of the gene for α -1,3-glucan synthase in the yeast phase than in the mycelial phase. Early reports on the mycelium-to-yeast transition

of *H. capsulatum* attributed an important role to sulfhydryl groups, mainly in the form of cysteine (Maresca and Kobayashi 2000). Hwang et al. (2003) extended these results by showing that the gene that encodes cysteine dioxygenase is more highly expressed in yeast than in the mycelium. The *CD11* gene, which encodes a cysteine dioxygenase in *P. brasiliensis*, was also identified in our SSH screen. It has previously been shown that the mycelial phase of *P. brasiliensis* can grow in the presence of both organic and inorganic sources of sulfur, while yeast cells can only grow in the presence of a sulfur-containing amino acid (Paris et al. 1985). This sulfur auxotrophy suggests that, in addition to temperature, organic sulfur is an important stimulus for keeping *P. brasiliensis* in the yeast phase. Accordingly, we verified that several genes involved in the assimilation of sulfur-containing amino acids, such as methionine permease, are more highly expressed in the yeast phase than in the mycelial phase, indicating a possible role for sulfur metabolism in the maintenance of the yeast state, as in the case of *H. capsulatum*.

The *CIT1* gene encodes citrate synthase, which catalyzes the first reaction in the citric acid cycle, the condensation of acetyl-CoA with oxalacetate to form citrate. The high specific activity of citrate synthase causes an accumulation of isocitrate and citrate in the mitochondria. Both acids are transported into the cytosol where citrate is cleaved by ATP:citrate lyase, giving rise to acetyl-CoA, the precursor of fatty acid synthesis (Lehninger et al. 1993). It is possible that the high-level expression of the *CIT1* gene in *P. brasiliensis* yeast cells could reflect an increased requirement for fatty acid synthesis. The same kind of observation can be made regarding the increase in expression of the *ERG25* gene in the yeast phase. The *ERG25* gene encodes the C-4 sterol methyl oxidase that performs the first of the three enzymatic steps leading to ergosterol biosynthesis in fungi. It is also likely that increased *ERG25* activity is involved in cell membrane turnover. As we have suggested previously, changes in the cell membrane are important factors for the morphological transition at 37°C (Goldman et al. 2003); the high turnover of lipids in the cell membrane may contribute to this. It has already been demonstrated by genome-wide expression analysis of *S. cerevisiae* that the genes involved in the ergosterol biosynthesis pathway show increased expression in the presence of stressing conditions, such as oxidative stress, and a number of antifungal agents target ergosterol biosynthesis (Higgins et al. 2003; Bammert and Fostel 2000). It remains to be determined whether all the genes for ergosterol biosynthesis genes are induced at 37°C, or if the effect is specific for the *ERG25* gene.

Lethal heat stress generates oxidative stress in *S. cerevisiae*, and anaerobic cells are several orders of magnitude more resistant than aerobic cells to a 50°C heat shock (Davidson and Schiestl 2001). The transition from 25°C to 37°C in *P. brasiliensis* (and also in *H. capsulatum* and *Blastomyces dermatitidis*) is accompanied by

marked changes in respiration and a rapid decline in intracellular ATP levels, due to uncoupling of oxidative phosphorylation (Medoff et al. 1987; Maresca and Kobayashi 2000). Although the level of ATP synthesis decreases, growth in the yeast phase at 37°C should give rise to increased production of reactive oxygen species (ROS) due to the higher temperature. These observations suggest that antioxidant systems in *P. brasiliensis* could be extremely important, not only for survival in macrophages, but also for detoxification of ROS. Accordingly, the preferential expression in the yeast phase of the *P. brasiliensis* *TSA1* (thiol-specific antioxidant) and *GST1* genes, which code for a thioredoxin peroxidase and a glutathione-S-transferase, respectively, may serve to provide antioxidants to reduce the levels of reactive oxygen species during growth of the yeast. Curiously, Choi et al. (2003) identified *Tsa1p* as highly expressed in the hyphae of *C. albicans* by two-dimensional gel electrophoresis.

The signaling pathways that control the morphological changes in *P. brasiliensis* are poorly understood, but the involvement of both cAMP and mitogen-activated protein kinase (MAPK) signal transduction pathways has been reported in other dimorphic fungi (Lengeler et al. 2000). Exogenous cAMP inhibits the yeast-to-mycelium (Y → M) transition in *P. brasiliensis*, thus maintaining the pathogenic yeast form (Borges-Walmsley and Walmsley 2000). In *Candida albicans*, on the other hand, the Y → M transition is controlled by cAMP and exogenous cAMP stimulates the formation of pseudohyphae (Rocha et al. 2001). By using SSH, we have identified four genes, *RHO1*, *SEP1*, *FLB1* and *PCK1*, which encode proteins involved in cell signaling and polarity establishment (Lengeler et al. 2000; Borges-Walmsley and Walmsley 2000; Momany 2002). Interestingly *RHO1* was expressed at 10- to 15-fold higher levels in minimal medium than in the complete media. Boyce et al. (2001) have observed that the *CDC42* homologue in *Penicillium marneffei*, *cflA*, is also required for polarization and determination of cell shape during yeast-like growth, and active CflA is required for the separation of yeast cells. However, correct *cflA* function is not required for dimorphic switching.

In summary, our work has succeeded in identifying several genes that could be involved in the maintenance of the pathogenic yeast phase of *P. brasiliensis*.

Acknowledgements We would like to thank the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and MCT-PRONEX, Brazil, for financial support. We also thank the two anonymous reviewers for useful comments

References

- 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* 25:3389–3402
- Bammert GF, Fostel JM (2000) Genome-wide expression patterns in *S. cerevisiae*: comparison of drug treatments and genetic alterations affecting biosynthesis of ergosterol. *Antimicrob Agents Chemother* 44:1255–1265
- Berman J, Sudbery PE (2002) *Candida albicans*: a molecular revolution built on lessons from budding yeast. *Nat Genet* 3:918–930
- Bialek R, Ibricevic A, Fothergill A, Begerow D (2000) Small subunit ribosomal DNA sequences shows *Paracoccidioides brasiliensis* closely related to *Blastomyces dermatitidis*. *J Clin Microbiol* 38:3190–3193
- Borges-Walmsley MI, Walmsley AR (2000) cAMP signaling in pathogenic fungi: control of dimorphic switching and pathogenicity. *Trends Microbiol* 8:133–141
- Borges-Walmsley MI, Chen D, Shu X, Walmsley AR (2002) The pathobiology of *Paracoccidioides brasiliensis*. *Trends Microbiol* 10:80–87
- Boyce KJ, Hymes MJ, Andrianopoulos A (2001) The *CDC42* homolog of the dimorphic fungus *Penicillium marneffei* is required for correct cell polarization during growth but not development. *J Bacteriol* 183:3447–3457
- Braun BR, Johnson AD (1997) Control of filament formation in *Candida albicans* by the transcriptional repressor *TUP1*. *Science* 277:105–109
- Choi W, Yoo YJ, Kim M, Shin D, Jeon HB, Choi W (2003) Identification of proteins highly expressed in the hyphae of *Candida albicans* by two-dimensional electrophoresis. *Yeast* 20:1053–1060
- Cisalpino PS, Puccia R, Yamauchi LM, Cano MIN, da Silveira JF, Travassos LR (1996) Cloning, characterization, and epitope expression of the major diagnostic antigen of *Paracoccidioides brasiliensis*. *J Biol Chem* 271:4553–4560
- Csank C, Schröppel K, Leberer E, Marcus D, Mohamed O, Meloche S, Thomas DY, Whiteway M (1998) Roles of the *Candida albicans* mitogen-activated protein kinase homolog, Cek1p, in hyphal development and systemic candidiasis. *Infect Immun* 66:2713–2721
- Da Fonseca CA, Jesuino RSA, Felipe MS, Cunha DA, Brito WA, Soares CMA (2001) Two-dimensional electrophoresis and characterization of antigens from *Paracoccidioides brasiliensis*. *Microbes Infection* 3:1–8
- Da Silva SP, Borges-Walmsley MI, Pereira IS, Soares CM, Walmsley AR, Felipe MS (1999) Differential expression of an hsp70 gene during transition from the mycelial to the infective yeast form of the human pathogenic fungus *Paracoccidioides brasiliensis*. *Mol Microbiol* 31:1039–1050
- Davidson JF, Schiestl RH (2001) Cytotoxic and genotoxic consequences of heat stress are dependent on the presence of oxygen in *Saccharomyces cerevisiae*. *J Bacteriol* 183:4580–4587
- Diatchenko L, Lau Y-F C, Campbell AP, Chenchik A, Moqadam F, Huang B, Lukyanov S, Lukyanov K, Gurskaya N, Sverdlov ED, Siebert PD (1996) Suppression subtraction hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proc Natl Acad Sci USA* 93:6025–6030
- Ewing B, Green P (1998) Base calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res* 8:196–194
- Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res* 8:175–185
- Goldman GH, et al (2003) EST analysis of the human pathogen *Paracoccidioides brasiliensis* yeast phase: identification of putative homologues of *Candida albicans* virulence/pathogenicity genes. *Eukaryotic Cell* 2:34–48
- Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. *Genome Res* 8:195–202
- Higgins VJ, Beckhouse AG, Oliver AD, Roger PJ, Dawes IW (2003) Yeast genome-wide expression analysis identifies a strong ergosterol and oxidative stress response during the initial stages of an industrial lager fermentation. *Appl Environ Microbiol* 69:4777–4787

- Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. *Genome Res* 9:868–877
- Hwang L, Hocking-Murray D, Bahrami AK, Andersson M, Rine J, Sil A (2003) Identifying phase-specific genes in the fungal pathogen *Histoplasma capsulatum* by using a genomic shotgun microarray. *Mol Biol Cell* 14:2314–2326
- Leberer E, Ziegelbauer K, Schmidt A, Marcus D, Dignard D, Ash J, Johnson L, Thomas DY (1997) Virulence and hyphal formation of *Candida albicans* require the Ste20p-like protein kinase CaCl4p. *Curr Biol* 7:539–546
- Lehninger AL, Nelson DL, Cox MM (1993) Principles of biochemistry (2nd edn). Worth Publishers, New York
- Lengeler KB, Davidson RC, D'Souza C, Harshima T, Shen W-C, Wang P, Pan X, Waugh M, Heitman J (2000) Signal transduction cascades regulating fungal development and virulence. *Microbiol Mol Biol Rev* 64:746–785
- Leverly SB, Toledo MS, Straus AH, Takahashi HK (1998) Structure elucidation of sphingolipids from the mycopathogen *Paracoccidioides brasiliensis*: an immunodominant β -galactofuranose residue is carried by a novel glycosylinositol phosphorylceramide antigen. *Biochemistry* 37:8764–8775
- Lo H-S, Kohler JR, DiDomenico B, Loeberberg D, Cacciapuoti A, Fink GR (1997) Nonfilamentous *C. albicans* mutants are avirulent. *Cell* 90:939–949
- Maresca B, Kobayashi GS (2000) Dimorphism in *Histoplasma capsulatum* and *Blastomyces dermatitidis*. In: Ernst JF, Schimdt R (eds) Dimorphism in human pathogenic and apathogenic yeasts (Contributions to microbiology, vol 5). Karger, Basel, pp 201–216
- Medoff G, Audrey P, Kobayashi G (1987) Mycelial- to yeast-phase transitions of the dimorphic fungi *Blastomyces dermatitidis* and *Paracoccidioides brasiliensis*. *J Bacteriol* 169:4055–4060
- Momany M (2002) Polarity in filamentous fungi: establishment, maintenance and new axes. *Curr Opin Microbiol* 5:580–585
- Paris S, Duran-Gonzalez S, Mariat F (1985) Nutritional studies on *Paracoccidioides brasiliensis*: the role of organic sulfur in dimorphism. *J Med Vet Mycol* 23:85–89
- Restrepo A., Jiménez BE (1980) Growth of *Paracoccidioides brasiliensis* yeast phase in a chemically defined culture medium. *J Clin Microbiol* 12:279–281
- Restrepo A, McEwen JG, Castañeda E (2001) The habitat of *Paracoccidioides brasiliensis*: how far from solving the riddle? *Med Mycol* 39:233–241
- Rocha CRC, Schröppel K, Marcus D, Macil A, Dignard D, Taylor BN, Thomas DY, Whiteway M, Leberer E (2001) Signaling through adenylyl cyclase is essential for hyphal growth and virulence in the pathogenic fungus *Candida albicans*. *Mol Biol Cell* 12:3631–3643
- Romani L, Bitoni F, Puccetti P (2003) Adaptation of *Candida albicans* to the host environment: the role of morphogenesis in virulence and survival in mammalian hosts. *Curr Opin Microbiol* 6:338–343
- Rooney PJ, Klein BS (2002) Linking fungal morphogenesis with virulence. *Cell Microbiol* 4:127–137
- Sambrook J, Russell DW (2001) Molecular cloning: a laboratory manual (3rd edn). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
- San-Blas G (1982) The cell wall of fungal human pathogens: its possible role in host-parasite relationship. A review. *Mycopathologia* 79:159–184
- San-Blas G, Niño-Veja G (2001) *Paracoccidioides brasiliensis*: virulence and host response. In: Cihlar RL, Calderone RA (eds). Fungal pathogenesis: principles and clinical applications. Marcel Dekker, New York
- Schummer M, Ng VLV, Baumgarner RE, Nelson PS, Schummer B, Bednarski DW, Hassell L, Baldwin RL, Karlan BY, Hood L (1999) Comparative hybridization of an array of 21500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. *Gene* 238:375–385
- Semighini CP, Marins M, Goldman MHS, Goldman GH (2002) Quantitative analysis of the relative transcript levels of ABC transporter *Atr* genes in *Aspergillus nidulans* by Real-Time Reverse Transcription-PCR assay. *Appl Environ Microbiol* 68:1351–1357
- Stoldt VR, Sonnenborn A, Leuker CE, Ernst JF (1997) Efg1p, an essential regulator of morphogenesis of the human pathogen *Candida albicans*, is a member of a conserved class of bHLH proteins regulating morphogenetic processes in fungi. *EMBO J* 16:1982–1991
- Toledo MS, Leverly SB, Straus AH, Suzuki E, Momany M, Glushka J, Moulton JM, Takahashi HK (1999) Characterization of sphingolipids from mycopathogens: factors correlating with expression of 2-hydroxy fatty acyl (E)- Δ [3]-unsaturation in cerebroside of *Paracoccidioides brasiliensis* and *Aspergillus fumigatus*. *Biochemistry* 38:7294–7306
- Venâncio EJ, Kyaw CM, Mello V, Silva SP, Soares CMA, Felipe MSS, Silva-Pereira I (2002) Identification of differentially expressed transcripts in the human pathogenic fungus *Paracoccidioides brasiliensis* by differential display. *Med Mycol* 40:45–51
- Vigh L, Maresca B, Horwood JL (1998) Does the membrane physical state control the expression of heat shock and other genes? *Trends Biochem Sci* 23:369–374
- Yaar L, Mavarech M, Koltin Y (1997) RAS-related gene (*CaRSRI*) is involved in budding, cell morphogenesis and hypha development. *Microbiology* 143:3033–3044

Transcriptionally active transposable elements in recent hybrid sugarcane

Paula G. de Araujo^{1,†,*}, Magdalena Rossi^{1,†}, Erika M. de Jesus¹, Nilo L. Saccaro Jr¹, Daniela Kajihara¹, Rogério Massa², Juliana Maria de Felix³, Rodrigo Duarte Drummond³, Maria C. Falco⁴, Sabrina M. Chabregas⁴, Eugênio C. Ulian⁴, Marcelo Menossi³ and Marie-Anne Van Sluys^{1,*}

¹Departamento de Botânica, IB-USP, Rua do Matão 277, São Paulo, CEP 055080-090, SP, Brazil,

²Agilent Technologies Brasil, Al. Araguaia, 1142, Alphaville, CEP 06455-940 Barueri, SP, Brazil,

³Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, C.P.6010, CEP 13083-875, Campinas, SP, Brazil, and

⁴Centro de Tecnologia Canaveira, Caixa Postal 162, CEP 13400-970 Piracicaba, SP, Brazil

Received 3 July 2005; revised 10 August 2005; accepted 16 August 2005.

*For correspondence (fax 55 11 3091 7547; e-mail mavsluys@usp.br).

†Both authors contributed equally to this paper and are joint first co-authors.

*Present address: Alellyx Applied Genomics, Rodovia Anhanguera Km 104, Campinas, CEP 13067-850 SP, Brazil.

Summary

Transposable elements (TEs) are considered to be important components of the maintenance and diversification of genomes. The recent increase in genome sequence data has created an opportunity to evaluate the impact of these active mobile elements on the evolution of plant genomes. Analysis of the sugarcane transcriptome identified 267 clones with significant similarity to previously described plant TEs. After full cDNA sequencing, 68 sugarcane TE clones were assigned to 11 families according to their best sequence alignment against a fully characterized element. Expression was further investigated through a combined study utilizing electronic Northern, microarray, transient and stable sugarcane transformation. Newly synthesized cDNA probes from flower, leaf roll, apical meristem and callus tissues confirm previous results. Callus was identified as the tissue with the highest number of TEs being expressed, revealing that tissue culture drastically induced the expression of different elements. No tissue-specific family was identified. Different representatives within a TE family displayed differential expression patterns, showing that each family presented expression in almost every tissue. Transformation experiments demonstrated that most *Hopscotch* clone-derived U3 regions are, indeed, active promoters, although under a strong transcriptional regulation. This is a large-scale study about the expression pattern of TEs and indicates that mobile genetic elements are transcriptionally active in the highly polyploid and complex sugarcane genome.

Keywords: expression profile, *Hopscotch*, microarrays, promoters, *Saccharum* spp, sugarcane, transposable elements.

Introduction

Transposable elements (TEs) are discrete fragments of DNA that move within the genome and often make duplicate copies of themselves in the process. With the advent of large-scale DNA sequencing it became evident that TEs are the largest component of the genetic material of most eukaryotes. They are present in very high copy number, averaging 45% of the human genome and between 50 and 80% of some grass genomes (Feschotte *et al.*, 2002). These elements are mutagenic agents and their activity produces

sequence changes in single genes as well as large genome rearrangements, both of which can alter patterns of gene expression and gene function (Bennetzen, 2000). Transposable elements generate variability and may have an adaptive function in eukaryote evolution, as originally proposed by McClintock (Feschotte *et al.*, 2002; McClintock, 1984).

Eukaryotic TEs are divided into two classes, according to the intermediate substrate for transposition: an RNA molecule for class I and a DNA molecule for class II. Each class of

TE contains autonomous elements that encode functional products required for transposition and non-autonomous elements that only retain the *cis* sequences necessary for recognition by the transposition machinery. Regardless of the transposition mechanism, the integration of most types of TE results in the duplication of a host genome sequence at the site of insertion, although the size of the duplication varies among TE families (Feschotte *et al.*, 2002).

Class I elements can be divided into two groups on the basis of structure. The first group comprises the retrotransposons with long terminal repeats (LTRs). Autonomous elements of this group encode a capsid-like protein (*gag*), and a polyprotein (*pol*) that is processed to yield protease, reverse transcriptase, RNase H and integrase activities. The LTR non-autonomous elements, called TRIM (terminal-repeat retrotransposons in miniature), lack most or all coding sequence and require proteins in *trans* for mobilization (Witte *et al.*, 2001). The other group involves the non-LTR retrotransposons, or retroposons, that are divided into LINEs (long interspersed nuclear elements) and SINEs (short interspersed nuclear elements) (Feschotte *et al.*, 2002).

Deoxyribonucleic acid transposons, or class II elements, are characterized by terminal inverted repeats (TIRs) and they encode a transposase (Tpase), the protein responsible for excision/insertion of the element. This class comprises the best known plant transposons like *Ac* (Doring *et al.*, 1984), *En/Spm* (Gierl and Saedler, 1989) and *MuDR* (Barker *et al.*, 1984) of maize. The small, non-autonomous elements called MITES (miniature inverted-repeat transposable elements) are structurally reminiscent of active class II non-autonomous elements (Feschotte *et al.*, 2002).

Probably due to the mutagenic activity of TEs, eukaryote genomes have developed efficient mechanisms to silence them (Casacuberta and Santiago, 2003). The transcriptional activity of class I elements is tightly controlled, although some retrotransposons are transcribed under stress conditions such as pathogen infection, physical injuries or abiotic stresses (Grandbastien, 1998; Takeda *et al.*, 1998). Both of the *Mutator* transposon's genes are constitutively expressed in meristems, immature organs and fully differentiated tissues in active maize lines. Transposition activity is restricted to the last few cell divisions during development of somatic or germinal tissue and does not correlate with transcript levels (Rudenko and Walbot, 2001; Rudenko *et al.*, 2003). In *Activator* maize lines, transcripts of the *Ac* transposase gene are rare in seedlings, comprising $(2-13) \times 10^{-5}$ of the poly A⁺ RNA, or 1.5–10 transcripts per cell (Fusswinkel *et al.*, 1991). *Ac* transposase protein accumulates in endosperm nuclei in 2- μ m aggregates. Since the presence of these forms does not correlate with excision frequency they may serve a regulatory function by inactivating excess transposase (Heinlein *et al.*, 1994).

The recent advances in gene silencing mechanisms have greatly increased our understanding of TE silencing. There is

evidence that TEs producing double-stranded or aberrant RNA are silenced by a post-transcriptional gene silencing mechanism (PTGS), while active TEs, which are abundant in the genome, are inactivated by a transcriptional gene silencing (TGS). The mechanism of TGS involves chromatin remodeling and results in a condensed heterochromatin region inaccessible to the transcriptional machinery (Okamoto and Hirochika, 2001).

The Brazilian Sugarcane EST Sequencing Project (SU-CEST) identified a surprising diversity of expressed TE-like sequences (Rossi *et al.*, 2001; Vettore *et al.*, 2003; Vincentz *et al.*, 2004). Here, the expression of 68 TE cDNA clones of sugarcane was analyzed by a combined strategy of electronic Northern, macroarrays and transient and stable sugarcane transformation. Callus was identified as the tissue with the highest number of TEs being expressed. Although different representatives displayed differential expression patterns no tissue-specific family was identified. Transformation experiments demonstrated that most *Hopscotch* clone-derived U3 regions are, indeed, active promoters, although under strong transcriptional regulation.

Results

Characterization of sugarcane TEs

The 267 sugarcane TE-like clones previously identified *in silico* (Rossi *et al.*, 2001) were fully sequenced and assigned to a family according to the greatest similarity to a fully characterized element, defined here as the reference element. An exhaustive sequence analysis was carried out in order to select exclusively mRNAs coming from transcription of TEs. Although the production of aberrant poly A⁻ TE mRNAs was reported, and their role in TGS and PTGS extensively investigated, we decided to eliminate from the analysis all poly A⁻ clones avoiding genomic contamination (Rudenko *et al.*, 2003). Redundant clones, which are those clones with identical sequences and identified in the same cDNA library, were also eliminated. The remaining clones were searched for possible read-through cDNA sequences, analyzing their 5' and 3' ends beyond the reference element homologous region by BLASTX and BLASTN. Finally, 68 clones were selected representing 11 TE families and their features are displayed in Table 1. *Mutator* and *Hopscotch* were the transposon and retrotransposon families, respectively most represented within the sugarcane transcriptome. A putative transposase or polyprotein was considered complete when it spanned at least 60% of the reference protein. The presence of stop codons or frame shifts was also analyzed. Four transposon-like clones were shown to be putative complete ORFs without frame shifts or stop codons, two homologous to *Mutator* (TE109 and TE148) and two homologous to *Activator* (TE048 and TE096). Even presenting a frame shift between the integrase and the reverse

Table 1 TE-like cDNA features isolated from sugarcane transcriptome

Reference element ^a	Clone name ^b	Tissue ^c	Size (bp)	Frame shift/stop ^d	Similarity with reference element (aa) (%) ^e	Alignment with reference element ^f	Putative complete transposase or polyprotein ^g
<i>Mutator</i> (823 aa) (gi2130141)	TE019	SD	1734	No/no	43	(1 nt)230–714(305 nt)	Incomplete cDNA
	TE029	SD	788	No/no	40	(16 nt)468–713(82 nt)	Incomplete cDNA
	TE037	LR	1128	Yes/yes	72	(3 nt)364–643(273 nt)	Incomplete cDNA
	TE047	CL	1679	No/yes	59	(29 nt)319–710(507 nt)	Incomplete cDNA
	TE059	FL	1063	No/no	44	(32 nt)444–640(461 nt)	Incomplete cDNA
	TE076	SD	1701	No/no	43	(8 nt)231–714(265 nt)	Incomplete cDNA
	TE084	SB	1373	No/no	43	(2 nt)318–714(198 nt)	Incomplete cDNA
	TE085	SD	645	No/no	40	(5 nt)508–713(67 nt)	Incomplete cDNA
	TE101	RT	1213	No/no	40	(4 nt)395–713(264 nt)	Incomplete cDNA
	TE109	AM	2348	No/no	42	(29 nt)128–704(520 nt)	Yes (70%)
	TE120	SB	1552	No/no	43	(9 nt)289–714(289 nt)	Incomplete cDNA
	TE121	SD	970	No/no	42	(37 nt)456–714(201 nt)	Incomplete cDNA
	TE138	RZ	1102	No/no	49	(1 nt)540–704(633 nt)	Incomplete cDNA
	TE142	SB	1596	Yes/yes	67	(12 nt)343–719(386 nt)	Incomplete cDNA
	TE148	LR	2349	No/no	42	(154 nt)133–714(48 nt)	Yes (70%)
	TE161	RT	1361	No/no	39	(53 nt)430–714(501 nt)	Incomplete cDNA
	TE165	FL	1948	Yes/yes	65	(11 nt)265–668(729 nt)	Incomplete cDNA
	TE172	FL	1176	No/no	44	(5 nt)448–666(538 nt)	Incomplete cDNA
	TE177	AD	1651	Yes/yes	66	(3 nt)338–721(428 nt)	Incomplete cDNA
	TE187	HR	1293	No/no	45	(34 nt)435–666(587 nt)	Incomplete cDNA
	TE194	CL	851	No/yes	60	(12 nt)456–715(2 nt)	Incomplete cDNA
	TE227	SD	703	No/no	41	(36 nt)546–714(244 nt)	Incomplete cDNA
	TE228	SB	1701	No/no	43	(8 nt)231–714(265 nt)	Incomplete cDNA
	TE263	SD	965	No/no	41	(50 nt)468–713(201 nt)	Incomplete cDNA
	TE266	CL	1268	No/no	41	(29 nt)385–713(290 nt)	Incomplete cDNA
	TE274	LB	1213	No/no	42	(1 nt)374–714(201 nt)	Incomplete cDNA
	TE279	ST	689	No/no	41	(56 nt)546–714(210 nt)	Incomplete cDNA
<i>Activator</i> (807 aa) (gi22113)	TE007	LB	372	Yes/no	62	(0 nt)159–573(0 nt)	Incomplete cDNA
	TE017	FL	1229	No/no	46	(24 nt)450–754(257 nt)	Incomplete cDNA
	TE046	CL	1049	No/no	81	(20 nt)158–438(183 nt)	Incomplete cDNA
	TE048	CL	2148	No/no	49	(188 nt)223–762(316 nt)	Yes (67%)
	TE096	LR	1828	No/no	48	(1 nt)259–762(291 nt)	Yes (62%)
	TE203	SD	1005	No/no	45	(11 nt)512–754(226 nt)	Incomplete cDNA
	TE207	AD	1435	No/no	51	(1 nt)397–762(312 nt)	Incomplete cDNA
	TE221	FL	2078	Yes/yes	48	(9 nt)184–774(322 nt)	Yes (73%)
	TE144	AM	1760	Yes/no	51	(27 nt)300–684(491 nt)	Incomplete cDNA
<i>Tam3</i> (749 aa) (gi16064)	TE191	AD	2179	No/yes	50	(44 nt)238–712(689 nt)	Yes (63%)
	TE153	RT	989	No/no	61	(3 nt)538–805(185 nt)	Incomplete cDNA
<i>Tip100</i> (808 aa) (gi4063770)	TE010	FL	1871	No/no	68	(154 nt)925–1425(220 nt)	Incomplete cDNA
<i>Hopscotch</i> (1439 aa) (gi7444442)	TE028	RT	2915	Yes/no	66	(4 nt)540–1425(259 nt)	Yes (62%)
	TE043	CL	1689	No/no	69	(32 nt)964–1425(277 nt)	Incomplete cDNA
	TE070	RT	1502	No/no	68	(14 nt)1008–1425(240 nt)	Incomplete cDNA
	TE083	RT	2464	Yes/no	55	(2 nt)674–1425(217 nt)	Incomplete cDNA
	TE098	RT	1750	No/no	66	(2 nt)927–1425(251 nt)	Incomplete cDNA
	TE174	HR	1532	Yes/no	65	992–1425(231 nt)	Incomplete cDNA
	TE176	ST	707	No/no	67	(4 nt)1264–1425(217 nt)	Incomplete cDNA
	TE186	FL	958	No/no	60	(133 nt)1235–1425(252 nt)	Incomplete cDNA
	TE197	FL	1763	No/no	60	(18 nt)931–1425(233 nt)	Incomplete cDNA
	TE250	RT	1815	No/no	62	(2 nt)600–1167(1 nt)	Incomplete cDNA
	TE273	LB	1326	No/no	64	(4 nt)1069–1425(251 nt)	Incomplete cDNA
	TE276	RT	1717	No/no	76	(569 nt)896–1276(2 nt)	Incomplete cDNA
<i>Maggy</i> (pol.1199 aa) (gi522302)	TE027	LR	1664	No/no	52	(123 nt)747–1168(280 nt)	Incomplete cDNA
	TE053	CL	2607	No/no	50	(39 nt)394–1095(594 nt)	Incomplete cDNA
	TE219	FL	1598	No/no	50	861–1168(680 nt)	Incomplete cDNA
	TE269	CL	2605	Yes/no	50	(39 nt)394–1095(593 nt)	Incomplete cDNA
<i>Opie</i> (pol1068 aa) (gi7489803)	TE018	LV	2222	No/no	61	(142 nt)4–689(4 nt)	Incomplete cDNA
	TE024	FL	1509	No/no	82	(37 nt)768–1066(575 nt)	Incomplete cDNA
	TE057	FL	1119	No/no	83	(1 nt)574–944(2 nt)	Incomplete cDNA

Table 1 Continued

Reference element ^a	Clone name ^b	Tissue ^c	Size (bp)	Frame shift/stop ^d	Similarity with reference element (aa) (%) ^e	Alignment with reference element ^f	Putative complete transposase or polyprotein ^g
<i>Stonor4</i> (1406 aa) (gi7489738)	TE182	FL	663	No/no	80	(1 nt)725–944(2 nt)	Incomplete cDNA
	TE245	FL	933	No/no	69	(1 nt)645–891(191 nt)	Incomplete cDNA
	TE009	RZ	541	Yes/no	53	(28 nt)193–395(2 nt)	Incomplete cDNA
	TE040	AM	1043	Yes/no	61	(130 nt)58–397	Incomplete cDNA
	TE106	SD	733	Yes/no	59	(4 nt)135–412(2 nt)	Incomplete cDNA
	TE258	AM	1444	Yes/no	61	(535 nt)58–395(2 nt)	Incomplete cDNA
<i>Tekay</i> (1470 aa) (gi18092337)	TE270	FL	1890	Yes/yes	69	(2 nt)372–1019(2 nt)	Incomplete cDNA
	TE159	ST	3219	No/yes	74	(2 nt)560–1468(520 nt)	Yes (62%)
<i>Tnt1</i> (1328 aa) (gi130582)	TE049	CL	2718	No/no	55	(2 nt)550–1325(307 nt)	Incomplete cDNA
<i>Tto1</i> (1338 aa) (gi7444414)	TE002	AM	1198	Yes/yes	72	(2 nt)959–1328(91 nt)	Incomplete cDNA

^aReference elements that were used to classify TE-like cDNA clones.

^bSugarcane TE identification, which is different from the SUCST clone name.

^cTissue library where cDNA clones were originally identified. AD, root–shoot transition zone, stalk and apical meristem of plantlets maintained *in vitro* inoculated with *Gluconacetobacter diazotrophicans*. AM, apical meristem and surrounding tissues of mature and immature plants. CL, pool of sugarcane calli. FL, mature and immature inflorescence. HR, root–shoot transition zone, stalk and apical meristem of plantlets maintained *in vitro* inoculated with *Herbaspirillum rubrisubalbicans*. LB, lateral buds from field-grown adult plants and adult plants grown in a greenhouse. LR, leaf roll from field-grown adult plants (large and small insert library). LV, etiolated leaves from plantlets grown *in vitro*. RT, root tips (1–3 cm long) from adult plants and root apex from the stalk bottom of adult plants. RZ, shoot–root transition zone from adult and young plants. SB, stalk bark from undeveloped, upper part of adult plants. SD, developing seeds. ST, first and fourth apical stalk internode of adult plants.

^dPresence of frame shifts and/or stop codons in the nucleotide sequence resulting in a potentially inactive protein.

^eAmino acid similarity along the alignment (BLASTX) with the reference-element.

^fDescription of the alignment against the reference-element. Between brackets, the number of nucleotides (nt) without homology, at the 5' and 3' end respectively. Central numbers indicate the reference element amino acid interval homologous to the cDNA clone.

^gA cDNA was considered completed if it spans at least 60% of the reference element protein and contains possible start and stop codons 5' and 3' respectively. Incomplete cDNAs have been originated from an incomplete reverse transcription.

transcriptase domains, the clone TE028, homologous to the *Hopscotch* retrotransposon, was considered a putative complete ORF because frame shifts are involved in the regulation mechanism for polyprotein production in many LTR retrotransposons (Kumar and Bennetzen, 1999). Most of the TE-derived cDNAs described are 5' end-incomplete most probably due to partial reverse transcription during libraries construction.

Expression pattern of 68 sugarcane TEs

Macroarray experiments were performed using cDNA probes synthesized with freshly extracted RNA from sugarcane callus, apical meristem, leaf roll and flower tissues. The 68 clones were spotted in triplicate onto five membranes. The three most homogeneous filters were selected and first hybridized with callus cDNA, and after stripping rehybridized with apical meristem, leaf roll or flower cDNA, respectively (see Materials and Methods). Two families, *Mutator* (38%) and the retrotransposon *Hopscotch* (13%), account for more than 50% of the expressed TEs, while only five clones (8%) showed no expression in the tested tissues (Figure 1). These results confirm the initial observations that these two families, which were the most abundant TE-like ESTs in the SUCST database (Table 1), are the most highly expressed. *MAGGY* retrotransposon-

like clones displayed a wide range of expression, easily distinguished in Figure 1.

To compare the macroarray expression pattern with the tissue from which clones were originally identified an electronic Northern was constructed from EST sequencing results considering only the tissues evaluated in the macroarray (Figure 2). Both histograms showed a similar distribution pattern, with callus being highest in transcript abundance.

To gain even more confidence about the results, clones of the two most abundantly expressed TE families (*Mutator*-like clone TE109, and *Hopscotch*-like clone TE010) were selected for hybridization studies on Northern blots. Total RNA from independent samples was size-fractionated by gel electrophoresis, transferred to a nylon membrane and hybridized with labeled probes for clones TE109 and TE010. Results presented in Figure 3(a) reveal the same expression pattern as that obtained in the macroarray experiment (Figure 3b).

Transcriptional activity of U3 regions from Hopscotch-like cDNA clones

The LTR retrotransposons are flanked by LTRs in direct orientation. Long terminal repeats can range from few 100 bp to over 5 kbp in size, and are usually divided into three

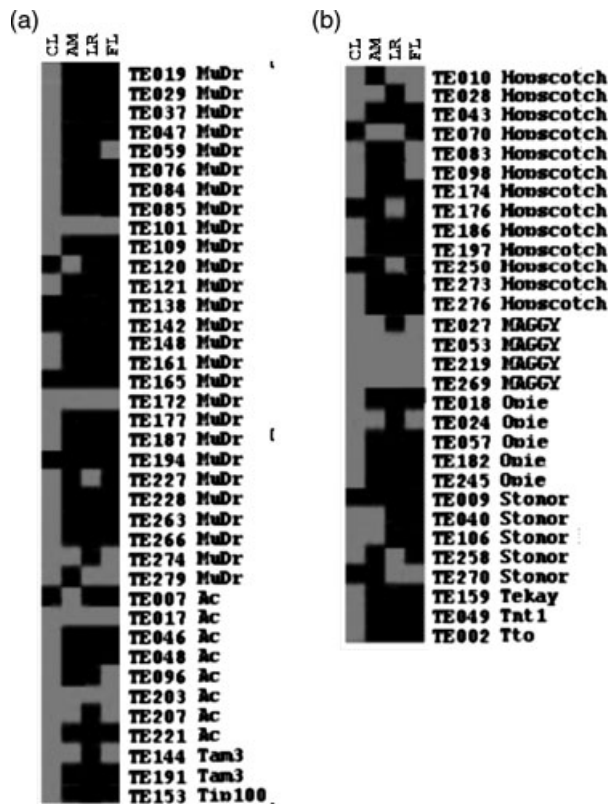


Figure 1. Macroarray diagram from TE expression analysis. Rows represent clones. Columns represent different tissues. CL, callus; AM, apical meristem; LR, leaf roll; FL, flower. Gray and black indicate expressed and non-expressed clones respectively. (a) Class I elements. (b) Class II elements.

regions: U3, R and U5. Transcription initiates at the 5' end of R in the 5'-LTR and terminates in the 3' end of R in the 3'-LTR, which defines U3 as the promoter region of an element. Thus, both, 5'- and 3'-LTR regions contain a promoter that can lead to read-through transcription of sequences downstream from the inserted element (Kumar and Bennetzen,

Figure 2. Distribution frequency of TE expression among the studied tissues.

(a) Macroarray experiment. (b) SUCST libraries. The number of expressed TEs identified in each cDNA library was normalized with respect to the total number of clones sequenced per library (Vettore *et al.*, 2003). CL, callus; AM, apical meristem; LR, leaf roll; FL, flower.

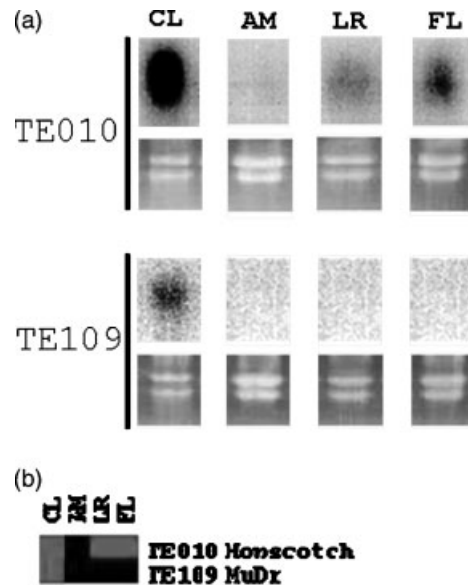
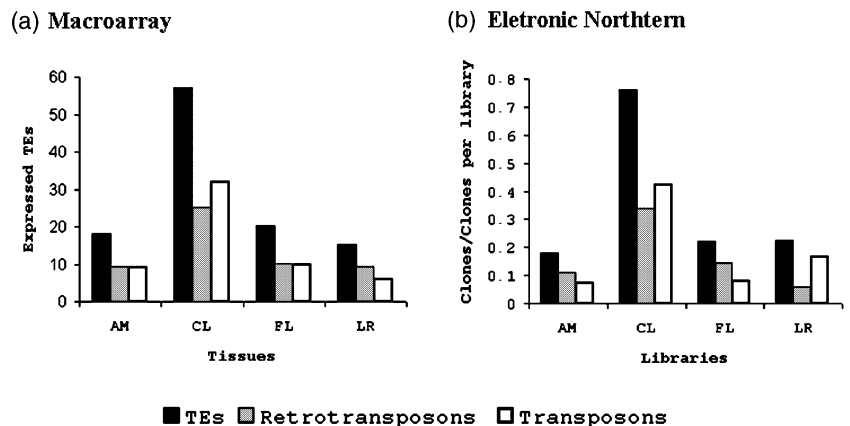


Figure 3. Ribonucleic acid blot hybridization. (a) Hybridization profile for *MuDr* transposon-like clone TE109 probe and *Hopsotch* retrotransposon-like clone TE010 probe. CL, callus; AM, apical meristem; LR, leaf roll; FL, flower. Both gel pictures are shown as control for the amount of RNA. (b) Macroarray detail for both TE109 and TE010 clones.

1999). The transcriptional promoter activity of the U3 region of LTR retrotransposons has been demonstrated for four plant retrotransposons: *Tnt1* (Grandbastien, 1998), *Tto1* (Takeda *et al.*, 1998), *Tos17* (Takeda *et al.*, 1998) and *Retrolyc1* (Araujo *et al.*, 2001). To better understand expression determinants of the most representative retrotransposon in the sugarcane transcriptome we analyzed the 13 *Hopsotch*-like cDNA clones searching for putative U3 regions. For four of the eleven that harbored putative LTRs (Table 1), we constructed U3-β-glucuronidase (GUS) reporter gene fusions. Figure 4 presents the nucleotide alignment of the predicted U3 region located in the 3' end of *Hopsotch*-like cDNA clones TE010, TE186, TE197 and TE273,

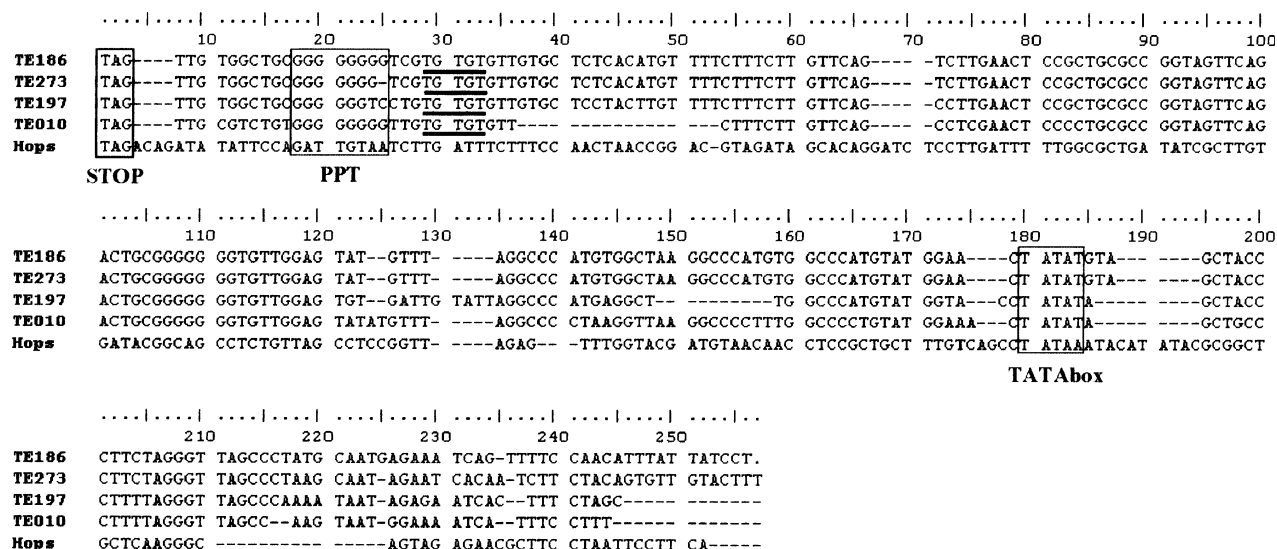


Figure 4. 3'-UTR region alignment of *Hopscotch*-like cDNA clones.

CLUSTALX alignment of four *Hopscotch*-like U3-putative promoters from clones TE010, TE186, TE197 and TE273. Black boxes indicate reading sequence 5' to 3': reverse transcriptase stop codon, the putative polypurine track and the putative TATA box. Underlined is the consensus sequence for the LTR 5' end. The aligned sequences of TE010, TE186, TE197 and TE273 span the 3' U3 region from the reverse transcriptase stop codon (black box) until the last clone base right before the poly A⁺ tail. The maize *Hopscotch* fragment encompasses the last 231 nucleotides of the GenBank sequence (gi:531388) from the reverse transcriptase stop codon.

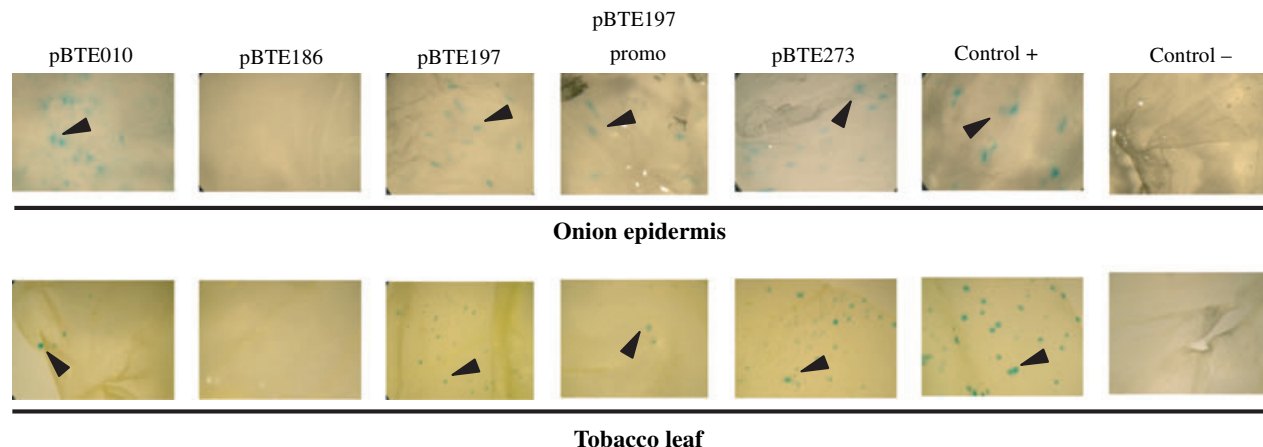


Figure 5. Transient expression analyses of *Hopscotch*-like U3-putative promoters/GUS fusions.

Onion epidermis and tobacco leaves were bombarded with the following plasmids: pBTE010, pBTE186, pBTE197, pBTE197promo and pBTE273. Arrows indicate blue points due to GUS activity. Positive controls for onion epidermis and tobacco leaves are the 35S promoter/GUS fusion and ubiquitin promoter/GUS fusion, respectively. The negative controls are the same tissues bombarded with particles without DNA.

from the polyprotein stop codon to the end of the clones right before the poly A⁺ tail. With the exception of the construct pBTE197promo, the other constructs used for U3-GUS fusions not only contain the U3 region but also part of the polyprotein and even the whole cDNA clone (Table 2). The pBTE197promo construct encompasses only 209 bp of the 3' end of the cDNA clone and was used as a control to assess whether the promoter function was affected by the polyprotein sequence.

Microprojectile bombardment of the constructs into *Allium cepa* epidermal cells and tobacco leaves revealed

that four out of the five constructs (pBTE010, pBTE197, pBTE273 and pBTE197promo) were capable of supporting GUS transcription (Figure 5). There was no qualitative difference in GUS expression between pBTE197 and pBTE197promo indicating that no transcriptional determinants are present in the polyprotein sequence. Also, while pBTE273 supported transient expression of GUS, pBTE186 did not, although the two putative promoters differ by only few nucleotides at the very 3' end of the U3 sequence after the putative TATA box (Figure 4). Although few sequence divergences could explain the drastic differences

Table 2 Putative promoter-GUS fusions

Construct	cDNA full size (bp)	Construct fragment size (bp) ^a	Cloning strategy
pBTE010	1872	1365	Clone <i>Pst</i> I site/vector 3' polylinker <i>Bam</i> HI site
pBTE186	957	957	Vector 5' polylinker <i>Pst</i> I site/vector 3' polylinker <i>Bam</i> HI site
pBTE197	1737	1200	Clone <i>Pst</i> I site/vector 3' polylinker <i>Bam</i> HI site
pBTE197promo	1737	209	PCR with <i>Pst</i> I/ <i>Bam</i> HI containing primers followed by <i>Pst</i> I/ <i>Bam</i> HI restriction
pBTE273	1329	1274	Clone <i>Pst</i> I site/vector 3' polylinker <i>Bam</i> HI site

^aSize of the cDNA 3' end fragment fused to the *uidA* reporter gene.

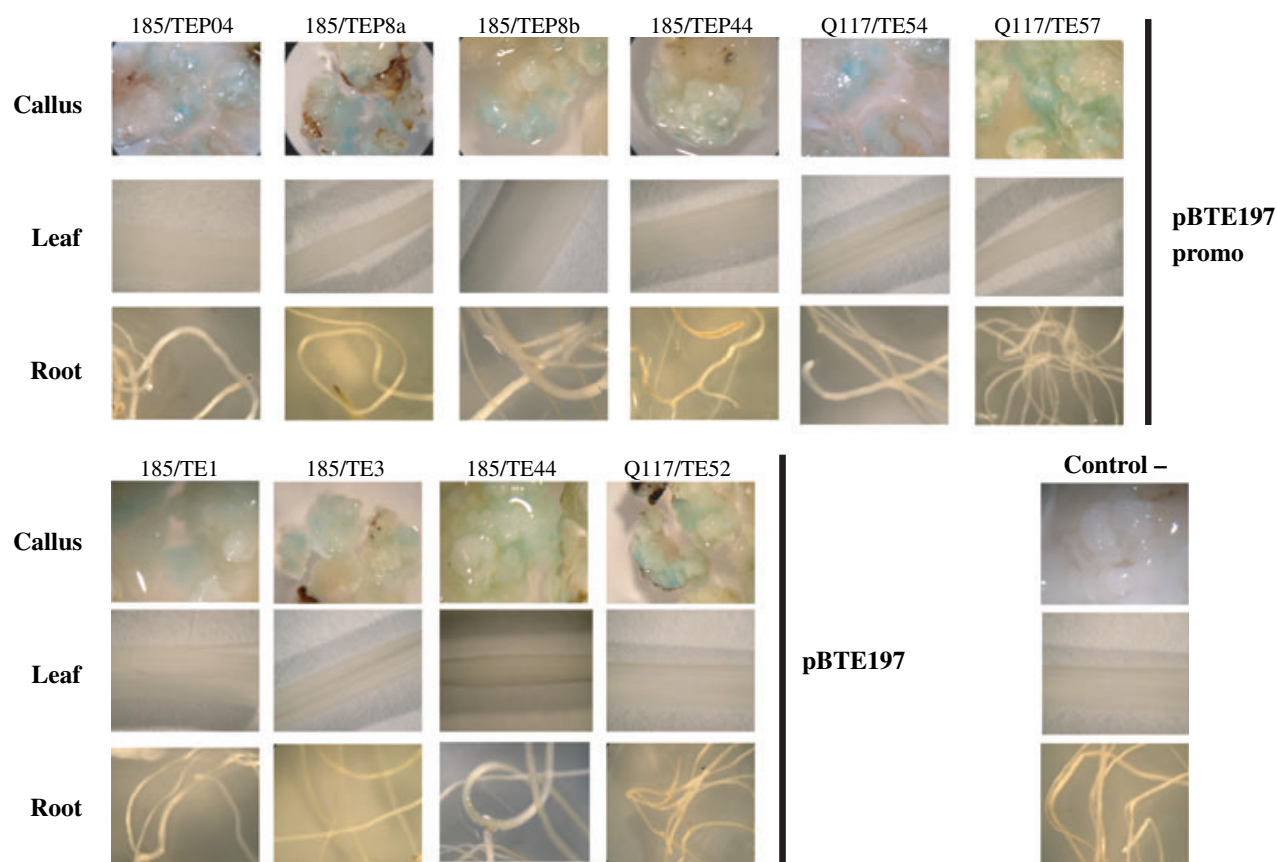
in promoter function, the absence of GUS activity upon bombardment with pBTE186 does not necessary reflect a failure in transcription. This could be due to an artifact in fusion construct that prevents translation or produces an out-of-frame protein.

Stable expression assays were carried out in sugarcane transgenic lines transformed with pBTE197 and pBTE197-promo constructs. A total of ten independent lines were generated, six for the pBTE197promo construct and four for the pBTE197, in two different sugarcane cultivars (SP80-185 and Q117). Different tissues of *in vitro* cultured transgenic plants were tested for GUS activity. Neither leaves nor roots displayed blue staining, but when calli were induced from the apical meristem of transgenic lines, transcription from the U3 region of the TE197 clone was induced (Figure 6). Clear tissue specificity was revealed for the transcriptional activity of the U3 region of the TE197 clone, and the U3 upstream sequence does not seem to modify the expression pattern. No differences were detected in the expression pattern between the two sugarcane cultivars (SP80-185 and Q117).

Discussion

Callus induction and TEs transcriptional activity

The high frequency of mutations found among plants regenerated from undifferentiated calli cultures is called somaclonal variation and includes a vast quantity of genetic

**Figure 6.** Stable expression analyses of *Hopscotch*-like U3-putative promoters/GUS fusion.

Histochemical GUS expression analysis of TE197. Different tissues of transgenic sugarcane lines for pBTE197 or pBTE197promo constructs were tested for GUS activity. The negative control is an untransformed plant.

changes. Genetic modifications include chromosomal rearrangements, single-gene mutations and changes in DNA methylation (Phillips *et al.*, 1994). Callus induction produces a genome resetting that does not necessarily mimic the same sequence of events that occur under natural conditions, giving rise to a broadly altered gene expression pattern (McClintock, 1984). Tissue culture-induced mutations have been identified in every species studied, and sugarcane is no exception. Several reports have described somaclonal variation in sugarcane plants regenerated from callus culture (Irvine, 1984; Irvine *et al.*, 1991; Sreenivasan *et al.*, 1987; Taylor *et al.*, 1995).

Although no somaclonal phenotypic mutants have been associated with TEs, transpositional activation was demonstrated for some LTR retrotransposons in callus-regenerated plants. This is the case for retrotransposons *Tnt1* and *Tto1* from tobacco and *Tos10* and *Tos17* from rice (Courtial *et al.*, 2001; Melayah *et al.*, 2001). Concerning this evidence, our results reveal that a large number of TE families are expressed in callus (Figure 2) and might be responsible for some of the somaclonal variation phenomena described in sugarcane. Moreover, data showed an increased transcriptional activity of classical DNA transposons with tissue culture (Figures 1 and 2). To precisely address this point we are conducting experiments to evaluate possible new TE insertions in somaclonal mutant plants, and during callus induction and regenerated plants.

LTR-derived promoters

The activity of the U3-GUS fusions confirms that those sugarcane TEs are transcriptionally active and their own promoters drive their transcription. These results reinforce our sequence analysis data showing that the presented cDNA clones are indeed TE-derived mRNAs and not read-through chimeric transcripts.

The GUS activity in transgenic sugarcane plants reflected the results obtained from macroarrays. Clone TE197 was expressed in calli (Figure 1), and only after callus induction was it possible to activate GUS transcription in all transgenic lines (Figure 6). Our observations demonstrated, once again, that TEs are submitted to extraordinarily fine mechanisms of transcriptional regulation.

Genome instability and transcriptional activity of TEs

Sugarcane is an economically important crop that belongs to a family of monocotyledonous angiosperms. Modern sugarcane cultivars are mostly allopolyploids derived from an interspecific hybridization between the wild species *Saccharum spontaneum* ($2n = 36-128$), and the cultivated *Saccharum officinarum* ($2n = 70-140$). The resulting large and complex genome has between 100 and 130 chromo-

somes, of which 15–25% are contributed by *S. spontaneum*. *Saccharum officinarum* and *S. spontaneum* monoploid genomes contain around 930 and 750 Mbp respectively (D'Hont and Glaszmann, 2001).

It is estimated that up to 70% of angiosperms have suffered polyploidization during their evolutionary history (He *et al.*, 2003). Since most polyploid species are derived from genetically different diploid parents, they maintain higher levels of heterozygosity than their progenitors, supporting wider genetic diversity (Soltis and Soltis, 2000). Immediately after genome duplication the genome undergoes a complex process of reorganization that includes deletions, translocations, inversions and duplications resulting in altered patterns of gene expression and even new gene functions (He *et al.*, 2003). It has been recently demonstrated that retrotransposons are also activated by interspecific hybridization (Kashkush *et al.*, 2002) and consequently contribute to chromosomal rearrangements and the modification of adjacent gene expression (Kashkush *et al.*, 2003). Sugarcane is a perfect example of this genome stress phenomenon since it is undergoing the hybridization process between two species with their corresponding polyploid origins. In sugarcane, the 'recent hybrid' condition is enforced by its vegetative propagation. The analysis of the genealogy of modern sugarcane cultivars indicates that only eight to nine meioses have occurred since the original interspecific crossing (E. Ulian, personal communication).

An interesting example for comparison is the hexaploid *Triticum aestivum* (AABBDD). Dot blot experiments with 17 TE families for diploid and polyploid wheat revealed that the polyploid wheat genome has a higher copy number of TEs than does diploid wheat. Moreover, polyploid wheats presented more copies than the sum of their parental species genomes. Contrary to that, when synthetic amphiploids [AASS (*Aegilops speltoides* × *Triticum urartu*), A^mA^mS^{sh} (*Aegilops sharonensis* × *Triticum monoccocum*), SSDD (*A. speltoides* × *Aegilops tauschii*) and AABBDD (*Triticum turgidum* × *A. tauschii*)] are examined, no significant change is detected (Li *et al.*, 2004). Southern blot experiments for several of the TEs studied in this paper presented similar results, that is, no significant differences in copy number were revealed between *S. spontaneum* and *S. officinarum* and sugarcane hybrid (data not shown). If expression levels are compared among crops at different stages of polyploidization and domestication it is detected that from 407 000 *Zea mays* ($2n$, 2500 Mbp) ESTs, 0.014% displayed similarity to retrotransposons (Meyers *et al.*, 2001), while the polyploids *T. aestivum* ($6n$, 17 000 Mbp) and sugarcane ($8n$, 3150 Mbp) presented 2.4% and 2.3%, respectively of their transcriptomes to be composed of TEs (Li *et al.*, 2004; Vettore *et al.*, 2003). These data suggest that changes in copy number do not occur immediately after polyploidization but progressively after transcriptional activation.

Final considerations

In this paper we present results showing the transcription pattern of 68 TE-derived mRNAs with three different approaches: macroarrays, Northern and transient and stable sugarcane transformation. It is important to point out that this transcriptional activity does not necessarily correlate with transpositional activity. Even though, transcription is the first step allowing transposition to occur. For example, cytosine methylation is correlated with transcriptional and transpositional activity of DNA transposons. For LTR retrotransposons, transcription and subsequent transposition is only detectable under certain conditions like biotic and abiotic stresses and tissue culture (Casacuberta and Santiago, 2003; Okamoto and Hirochika, 2001).

Callus is the tissue where most TE clones are expressed, and although no direct correlation has been presented the involvement of TEs in the occurrence of somaclonal variation cannot be excluded. Finally, we isolated active U3-derived promoters that showed well-defined tissue specificity. These results open an unexplored source of biotechnological tools for gene expression.

Experimental procedures

Clones, DNA sequencing and analysis

The sugarcane clones homologous to TEs (Rossi *et al.*, 2001) were named TE (for transposable element) plus three digits. The 68 clones studied in this paper were submitted to GenBank under accession numbers DQ115016 to DQ115083.

The complete list of SUCEST project libraries and their respective sugarcane varieties can be found in Vettore *et al.* (2003).

Sequencing was performed with BigDye terminators on a Perkin Elmer ABI3700 sequencer. The BLAST program was used for similarity searches (Altschul *et al.*, 1997). The sequences were assembled using the Phred + Phrap + Consed package (Gordon *et al.*, 1998) and bases had a Phred quality of at least 20. Hopscotch-like LTR clones were aligned using CLUSTALX (1.81) for Macintosh.

Macroarray experiment

The 68 sugarcane clones homologous to TEs were arranged in 96-well microtiter plates. Deoxyribonucleic acid miniprep was performed as previously described in Vettore *et al.* (2003). Plasmid DNA was denatured in 0.2 N NaOH for 15 min at 37°C and then spotted twice in the same position onto Hybond-N filters (Hybond-N; Amersham Biosciences, Piscataway, NJ, USA) with a hand-held 96-pin printhead tool (V&P Scientific, San Diego, CA, USA). This replicator usually deposited 0.1 µl of DNA solution (approximately 5 ng of DNA). To decrease the variation in the amount of DNA among spots and filters, the set of 220 clones were spotted in triplicate onto five nylon filters (85 × 125 mm) in a 3 × 3 array. This procedure reduced the coefficient of variation among spots by 50% (Nogueira *et al.* 2003). Additionally, 12 spots containing DNA of the empty plasmid vector pSPORT1 (Invitrogen, Carlsbad, CA, USA) were

applied to the filters as a negative hybridization control. Deoxyribonucleic acid was fixed to the filters by baking at 80°C for 2 h and storing the filters at room temperature. In order to estimate variation in the amount of DNA in the spots, all filters were first hybridized with an oligonucleotide probe recognizing the sequence of the *Amp^r* gene in the pSPORT1 vector. Based on this hybridization result we chose the three most homogeneous filters to perform the following hybridizations. After stripping, the chosen filters were hybridized with cDNA sample probes.

For probe preparation, total RNA was extracted from R5 generation calli, apical meristem and leaves (leaf + 1, first leaf with a visible dewlap) from 14-month-old plants of cultivar SP80-3280. Immature flowers were collected from plants of cultivar SP87-432. Total RNA was extracted from 1 g of tissue using TRIzol Reagent (Invitrogen) according to the manufacturer's recommendations. Labeling and hybridization protocols are described in Nogueira *et al.* (2003). Thirty micrograms of total RNA was reverse transcribed with 250 U of Superscript II (Invitrogen), 100 mM DTT, 80 U of RNAGuard (Amersham Biosciences), 3 µM of oligo-dT18V, 50 µCi [α -³²P] deoxycytidine triphosphate (dCTP) and unlabeled deoxy-ATP, deoxyguanosine triphosphate (dGTP) and deoxythymidine triphosphate (dTTP) (1 mM each) for 20 min at 42°C. Then, unlabeled dCTP was added to a final concentration of 1 mM. The reaction was continued for another hour. Complementary DNA was purified using ProbeQuant G50 microcolumns (Amersham Biosciences) according to the manufacturer's instructions. Filters were hybridized with cDNA probes for 18 h at 42°C. Membranes were washed for 15 min twice with 2× SSC/0.1% SDS, twice with 1× SSC/0.1% SDS and twice with 0.1× SSC/0.1% SDS at 65°C. Exposure was in imaging plates for 96 h and then scanned in a phosphorimager FLA3000-G (Fujifilm, Tokyo, Japan).

All signals were quantified by using Array Vision software (Imaging Research, St Catharines, ON, Canada). Each spot signal and its corresponding local background were quantified individually. The paired Student's *t*-test was used to verify whether the emission of the spots corresponding to each clone was significantly different from the background emission around the same spots. A cut-off value of 2.92 was adopted, corresponding to a significance level of 5%, and those clones showing a *t*-test score higher than this value were considered as expressed. The final data were analyzed with packaged CLUSTAL and TreeView software (Eisen *et al.*, 1998).

Northern blot analysis

New total RNA extractions were done from the same tissue samples used for macroarray probe preparation as previously described. Ten micrograms of RNA were electrophoresed in 1% (w/v) agarose gel containing formaldehyde and transferred to Hybond-N membrane (Amersham Biosciences). The hybridization procedure was according to Papini-Terzi *et al.* (2003). To get rid of the poly A⁺ tail that could give cross-hybridization, the probes for clones TE010 and TE109 were obtained by PCR using the following internal primers: TE010g (5'-TGTAAGTCAGTCAGCCGGC-3'), TE010b (5'-CCATGGGCGAAG-TGCGGG-3'), TE109g (5'-GGGAGGATGGCTACGAGCAGCT-3') and TE109b (5'-CGACGAGACCCACCGAGGA-3'). PCR conditions were: 94°C 5 min, 94°C 30 sec, 58°C 30 sec, 72°C 1.30 min for 35 cycles and 72°C 10 min. The resulting PCR reaction was subjected to DNA agarose gel 1% in Tris-borate/EDTA electrophoresis buffer (TBE) 1× and purified in a Qiagen Column KIT (Qiagen, Valencia, CA, USA). Probes were radioactively labeled with an Invitrogen Random Primer Labeling System Kit using specific internal primers to increase stringency.

Constructs

The putative promoter regions (cDNA-3' end after the reverse transcriptase stop codon) of four selected *Hopscotch*-like cDNA clones were introduced upstream of the reporter gene *uidA* encoding for β -glucuronidase (GUS). SUCEST cDNA libraries were constructed into a pSport vector (Invitrogen) between *SaI* and *NotI* restriction sites. Insert fragments were removed from the pSport vector by *PstI*/*Bam*HI restriction (constructs: pTE010, pTE186, pTE197, pTE273) or PCR with *PstI* and *Bam*HI containing specific primers followed by *PstI*/*Bam*HI restriction (construct pTE197promo). On the other hand the pBI221 vector, which contains the *uidA* gene under 35S promoter control (Clontech), was opened by *PstI*/*Bam*HI restriction, removing the 35S promoter, and ligated to the previously obtained fragments. Table 2 summarizes the features of the constructs.

Transient expression

Particle bombardment of tobacco leaves and *A. cepa* epidermis was done according to Brasileiro and Carneiro (1998), using a helium version of the PDS-1000 device (Sanford *et al.*, 1991) developed by EMBRAPA (Brasília, Brazil). Deoxyribonucleic acid constructs for particle preparation were isolated with a QiAfilter™ Plasmid Midi Kit (Qiagen, no. 12243).

Two independent experiments were carried out in triplicate. Positive controls were the pBI221 vector containing the 35S::*uidA* cassette (Clontech, Palo Alto, CA, USA) and pUbi (ubiquitin promoter *uidA*, Centro de Tecnologia Canavieira, Caixa Postal 162, CEP 13400-970 Piracicaba, SP, Brazil) for tobacco and *A. cepa*, respectively. Negative controls were bombarded tissues with no DNA-containing particles.

Plant expression

Transgenic plants were recovered from embryogenic callus transformed using a modified biolistic protocol. Callus was initiated from sugarcane varieties SP80-185 and Q117 using CI-3 medium containing 3 mg l⁻¹ 2,4-dichlorophenoxyacetic acid (2,4-D), 5% coconut water, 150 mg l⁻¹ citric acid, 250 mg l⁻¹ Clavulin Beecham, and 7 g l⁻¹ agar. Young leaf rolls of 6–12-month-old plants were cultured in the dark at 27°C, and selected embryogenic callus was subcultured on the same medium every 3 weeks.

Calli were co-bombarded with plasmid vectors pTE197 or pTE197promo together with pHA9, which carries the geneticin resistance gene. Vector DNA preparation used a variation of the coprecipitation method of Klein *et al.* (1988a,b). After bombardment, calli were kept for 1 week on CI-3 medium, without selection, for recovery. Transgenic calli were selected on CI-3 medium containing 35 mg l⁻¹ of geneticin for 6 weeks in the dark. Resistant calli were placed on the same medium without 2,4-D to regenerate plants. After approximately 3 months rooted plants were obtained and kept *in vitro* by micropropagation. Non-transformed control plants were obtained by regeneration from the same callus type going through the same tissue culture steps without bombardment. Transgenic plants were tested for vector genomic insertion by PCR using primers designed on the promoter region (5'-AACTG-CAGTTGTGGCTGCGGGGGTCTG-3') and on the GUS coding region (5'-CGGGTTGGGTTTCTACAGGACG-3'). Four transgenic lines were obtained for the pBTE197 construct: Q117/TE52 for the Q117 variety, and 185/TE1, 185/TE3 and 185/TE44 for SP80-185. For the pBTE197promo construct six independent lines were regenerated: Q117/TEP54 and Q117/TEP57 for the Q117 variety, and 185/TEP4, 185/TEP8a, 185/TEP8b and 185/TEP44 for SP80-185.

For expression analysis, calli were induced from the apical meristem of plants from all transgenic lines by culturing on MS (Murashige and Skoog, 1962) basal medium supplemented with 3 mg l⁻¹ 2,4-D. After 7 weeks the calli presented somatic embryos and they were assayed for GUS activity. The culture room was maintained at 24 ± 2°C with a 12-h photoperiod.

GUS activity assays

For transient expression experiments, histochemical GUS activity was assayed 24 h after bombardment. The GUS activity in transgenic sugarcane lines was analyzed in leaves and roots from 1-month-old plants maintained *in vitro* and embryogenic calli. Staining of GUS activity was carried out as described by McCabe *et al.* (1988) using X-GLU (5-bromo-4-chloro-3-indolyl- β -D-glucuronide) as substrate.

Acknowledgements

The authors would like to thank Dr Nina Fedoroff for helpful comments on the manuscript. M. Rossi, E.M. de Jesus, P.G. Araujo and J.M. Félix were recipients of FAPESP fellowships. N.L. Saccaro Jr and D. Kajihara were recipients of CNPq fellowships. R.D. Drummond was a recipient of a CAPES fellowship. This work was supported with grants from FAPESP (Brazil) and CNPq (Brazil) to M.A. Van Sluys.

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.
- Araujo, P.G., Casacuberta, J.M., Costa, A.P.P., Hashimoto, R.Y., Grandbastien, M.A. and Van Sluys, M.A. (2001) Retrocyl1 subfamilies defined by different U3 LTR regulatory regions in the *Lycopersicon* genus. *Mol. Genet. Genomics*, **266**, 35–41.
- Barker, R.F., Thompson, D.V., Talbot, D.R., Swanson, J. and Bennetzen, J.L. (1984) Nucleotide sequence of the maize transposable element *Mu*. *Nucleic Acids Res.* **12**, 5955–5967.
- Bennetzen, J.L. (2000) Transposable element contributions to plant genome evolution. *Plant Mol. Biol.* **42**, 251–269.
- Brasileiro, A.C.M. and Carneiro, V.T.C. (1998) *Manual de transformação genética de plantas*. Brasília: Ed. Embrapa.
- Casacuberta, J.M. and Santiago, N. (2003) Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. *Gene*, **311**, 1–11.
- Courtial, B., Feuerbach, F., Eberhard, S., Rohmer, L., Chiapello, H., Camilleri, C. and Lucas, H. (2001) Tnt1 transposition events are induced by *in vitro* transformation of *Arabidopsis thaliana*, and transposed copies integrate into genes. *Mol. Genet. Genomics*, **265**, 32–42.
- D'Hont, A. and Glaszman, J.C. (2001) Sugarcane genome analysis with molecular markers, a first decade of research. *Proc. Int. Soc. Sugarcane Technol.* **24**, 556–559.
- Doring, H.P., Tillmann, E. and Starlinger, P. (1984) DNA sequence of the maize transposable element *Dissociation*. *Nature*, **307**, 127–130.
- Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

- Feschotte, C., Jiang, N. and Wessler, S.R. (2002) Plant transposable elements: where genetics meets genomics. *Nat. Genet.* **3**, 329–341.
- Fusswinkel, H., Schein, S., Courage, U., Starlinger, P. and Kunze, R. (1991) Detection and abundance of mRNA and protein encoded by transposable element activator (Ac) in maize. *Mol. Gen. Genet.* **225**, 186–192.
- Gierl, A. and Saedler, H. (1989) The *Em/Spm* transposable element of *Zea mays*. *Plant Mol. Biol.* **13**, 261–266.
- Gordon, D., Abajian, C. and Green, P. (1998) Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195–202.
- Grandbastien, M.A. (1998) Activation of plant retrotransposons under stress conditions. *Trends Plant Sci.* **3**, 181–187.
- He, P., Friebe, B.R., Gill, B.S. and Zhou, J.-M. (2003) Allopolyploidy alters gene expression in the highly stable hexaploid wheat. *Plant Mol. Biol.* **52**, 401–414.
- Heinlein, M., Brattig, T. and Kunze, R. (1994) In vivo aggregation of maize Activator (Ac) transposase in nuclei of maize endosperm and *Petunia* protoplasts. *Plant J.* **5**, 705–714.
- Irvine, J.E. (1984) The frequency of marker changes in sugarcane plants regenerated from callus culture. *Plant Cell Tissue Organ Cult.* **3**, 201–209.
- Irvine, J.E., Benda, G.T.A., Legendre, B.L. and Machada, G.R. (1991) The frequency of marker changes in sugarcane plants regenerated from callus culture: evidence for vegetative and genetic transmission, epigenetic effects and chimera disruption. *Plant Cell Tissue Organ Cult.* **26**, 115–125.
- Kashkush, K., Feldman, M. and Levy, A.A. (2002) Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics*, **160**, 1651–1659.
- Kashkush, K., Feldman, M. and Levy, A.A. (2003) Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nat. Genet.* **33**, 102–106.
- Klein, T.M., Fromm, M., Weissinger, A., Tomes, D., Schaff, S., Sletten, M. and Sanford, J.C. (1988a) Transfer of foreign genes into intact maize cells with high-velocity microprojectiles. *Proc. Natl Acad. Sci. USA*, **85**, 4305–4309.
- Klein, T.M., Gradziel, T., Fromm, M.E. and Sanford, J.C. (1988b) Factors influencing gene delivery into *Zea mays* cells by high-velocity microprojectiles. *Biotechnology*, **6**, 559–563.
- Kumar, A. and Bennetzen, J.L. (1999) Plant retrotransposons. *Annu. Rev. Genet.* **33**, 479–532.
- Li, W., Zhang, P., Fellers, J.P., Fribe, B. and Gill, B. (2004) Sequence composition, organization, and evolution of the core Triticeae genome. *Plant J.* **40**, 500–511.
- McCabe, D.E., Swain, W.F., Martinell, B.J. and Christou, P. (1988) Stable transformation of soybean (*Glycine max*) by particle acceleration. *Biotechnology*, **6**, 923–926.
- McClintock, B. (1984) The significance of responses of the genome to challenge. *Science*, **226**, 792–801.
- Melayah, D., Bonnard, E., Chalhouh, B., Audeon, C. and Grandbastien, M.A. (2001) The mobility of the tobacco Tnt1 retrotransposon correlates with its transcriptional activation by fungal factors. *Plant J.* **28**, 159–168.
- Meyers, B.C., Tingey, S.V. and Morgante, M. (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res.* **11**, 1660–1676.
- Murashige, T. and Skoog, F. (1962) A revised medium for rapid growth and bioassays with tobacco tissue culture. *Physiol. Plant.* **15**, 473–497.
- Nogueira, F.T., De Rosa, 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.
- Okamoto, H. and Hirochika, H. (2001) Silencing of transposable elements in plants. *Trends Plant Sci.* **6**, 527–534.
- Papini-Terzi, F.S., Galhardo da Silva, R.G., Farias, L.P., Menck, C.F. and Van Sluys, M.A. (2003) Point mutation is responsible for *Arabidopsis* *tz-201* mutant phenotype affecting thiamin biosynthesis. *Plant Cell Physiol.* **44**, 856–860.
- Phillips, L.R., Kaeppler, S.M. and Olhoft, P. (1994) Genetic instability of plant tissue cultures: breakdown of normal controls. *Proc. Natl Acad. Sci. USA*, **91**, 5222–5226.
- Rossi, M., Araujo, P.G. and Van Sluys, M.A. (2001) Survey of transposable elements in sugarcane expressed sequence tags (ESTs). *Genet. Mol. Biol.* **24**, 147–154.
- Rudenko, G.N. and Walbot, V. (2001) Expression and post-transcriptional regulation of maize transposable elements *MuDR* and its derivatives. *Plant Cell*, **13**, 553–570.
- Rudenko, G.N., Ono, A. and Walbot, V. (2003) Initiation of silencing of maize *MuDR/Mu* transposable elements. *Plant J.* **33**, 1013–1025.
- Sanford, J.C., Devit, M.J., Russell, J.A., Smith, F.D., Harpending, P.R., Roy, M.K. and Johnson, S.A. (1991) An improved helium-driven biolistic device. *Technique*, **3**, 13–16.
- Soltis, P.S. and Soltis, D.E. (2000) The role of genetic and genomic attributes in the success of polyploids. *Proc. Natl Acad. Sci. USA*, **97**, 7051–7057.
- Sreenivasan, J.T., Sreenivasan, T.V. and Alexander, K.C. (1987) Somaclonal variation for rust resistance in sugarcane. *Indian J. Genet.* **47**, 109–114.
- Takeda, S., Sugimoto, K., Otsuki, H. and Hirochika, H. (1998) Transcriptional activation of the tobacco retrotransposon *Tto1* by wounding and methyl jasmonate. *Plant Mol. Biol.* **36**, 365–376.
- Taylor, P.W.J., Geijskes, J.R., Ko, H.L., Fraser, T.A., Henry, R.J. and Birch, R.G. (1995) Sensitivity of random amplified polymorphism DNA analysis to detect genetic change in sugarcane during tissue culture. *Theor. Appl. Genet.* **90**, 1169–1173.
- Vettore, A.L., Da Silva, F.R., Kemper, E.L. et al. (2003) Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. *Genome Res.* **12**, 2725–2735.
- Vincentz, M., Cara, F.A.A., Okura, V.K. et al. (2004) Evaluation of monocot and eudicot divergence using the sugarcane transcriptome. *Plant Physiol.* **134**, 951–959.
- Witte, C.P., Le, Q.H., Bureau, T. and Kumar, A. (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proc. Natl Acad. Sci. USA*, **98**, 13778–13783.

GenBank: DQ115016 to DQ115083.

Isolation and characterization of *Coffea* genes induced during coffee leaf miner (*Leucoptera coffeella*) infestation

Jorge Maurício Costa Mondego^{a,b}, Oliveira Guerreiro-Filho^c, Mário Henrique Bengtson^d,
Rodrigo Duarte Drummond^{a,b}, Juliana de Maria Felix^{a,b}, Melina Pasini Duarte^{a,b},
Daniel Ramiro^c, Mirian Perez Maluf^c, Mari Cleide Sogayar^d, Marcelo Menossi^{a,b,*}

^a Centro de Biologia Molecular e Engenharia Genética, Laboratório de Genoma Funcional, Universidade Estadual de Campinas (UNICAMP), CP 6010, 13083-875 Campinas, SP, Brazil

^b Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas (UNICAMP), CP 6109, 13084-971 Campinas, SP, Brazil

^c Centro de Café Alcides Carvalho, Instituto Agrônomo de Campinas, 13020-902 Campinas, SP, Brazil

^d Instituto de Química, Universidade de São Paulo (USP), Laboratório de Biologia Celular e Molecular, 05508-900 São Paulo, SP, Brazil

Received 7 January 2005; received in revised form 15 March 2005

Available online 2 April 2005

Abstract

Coffea arabica, one of the most important breeding species in the world, is susceptible to the leaf miner *Leucoptera coffeella*, which causes severe damage to coffee plantations. A closely related coffee species (*Coffea racemosa*) resistant to this insect was crossed with *C. arabica* and resulted in segregating progenies with resistance or susceptibility to coffee leaf miner. The aim of this work was to isolate and characterize the genes involved in coffee resistance to this pest. Subtracted cDNA libraries enriched in genes preferentially expressed in coffee plants resistant to *L. coffeella* were constructed. Approximately 1500 clones were spotted on nylon membranes and hybridized to cDNA probes derived from RNA samples from infestation experiments. Several genes were differentially expressed. We selected expressed ESTs with the most interesting expression profiles and confirmed the up-regulation of five of them (Class III Chitinase PR-8, signal peptidase complex subunit SPC25, photosystem gene *psaH*, a putative calcium exchanger similar to CAX9 and a homeotic gene *BEL*) by RNA blot. The possible functions of these genes in coffee resistance and coffee development, and a hypothetical defense mechanism against *L. coffeella* are discussed.

© 2005 Elsevier Ireland Ltd. All rights reserved.

Keywords: Biotic stress; Coffee; DNA arrays; Plant-insect interactions; Subtracted cDNA libraries

1. Introduction

Coffee (*Coffea* sp.) is one of the most important tropical crops in the world. *Coffea arabica*, the main cultivated species, is susceptible to infestation by the coffee leaf miner (*Leucoptera coffeella*) (Guérin Méneville) (Lepidoptera: Lyonetiidae). This insect causes severe damage to coffee crops, with losses that may reach 50% of the total production [1]. The *L. coffeella* moth deposits its eggs on the adaxial

leaf surface. After the embryonic stage (4–6 days), the larva ecloses from the egg towards the mesophyll and forms mines during its feeding. Subsequently, the larva leaves the mine and becomes a chrysalis that adheres to the abaxial leaf surface. The transformation of the chrysalis into a new moth completes the insect's life cycle (Fig. 1).

Different levels of resistance to this pest have been observed among *Coffea* species. In resistant species, the larva dies within the mine, or may not even manage to eclose from the egg. The genes of one of these naturally resistant species, *C. racemosa*, have been used to develop pest-resistant coffee varieties because of the ease with

* Corresponding author. Tel.: +55 193788 1118; fax: +55 193788 1089.
E-mail address: menossi@unicamp.br (M. Menossi).

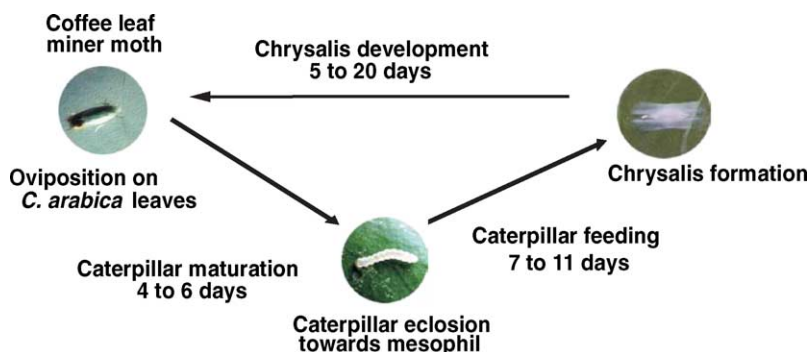


Fig. 1. Life cycle of the coffee leaf miner (*Leucoptera coffeella*).

which this species can be crossed with *C. arabica* [2]. The resistance to coffee leaf miner in plants obtained by crosses between *C. arabica* and *C. racemosa* apparently derives from two complementary and dominant genes [3,4]. Despite the importance of the damage caused by the coffee leaf miner to *C. arabica* crops, little is known about the defense mechanisms that resistant plants activate against *L. coffeella*.

Plants have developed several defense mechanisms to respond to biotic stress, such as caused by bacteria, fungi, nematodes and arthropods. Some signaling pathways have been identified as being important in this response. The salicylic acid (SA)-dependent cascade plays a central role in defense against pathogens [5,6], and the jasmonic acid (JA)-dependent cascade is related to the activation of defenses against pathogens and wounding, such as caused by herbivore chewing [5–7]. Most of the studies describing arthropod infestation have dealt with species that extensively injure the foliage [7]. There is little information on how plants deal with more sedentary herbivores, such as aphids, spider mites, galling insects and leaf miners, which cause less damage to the leaves [7].

The advent of genome projects has led to the development of new methodologies that have extended our ability to analyze gene expression. Among these techniques, DNA arrays on high-density nylon filters have been used to identify genes induced by various factors [8,9]. Additionally, the construction of subtractive cDNA libraries has also been used to isolate genes that are differentially expressed between two cDNA populations. Suppressive subtraction hybridization (SSH) [10] and Representational difference analysis (RDA) [11] are the most commonly used techniques to prepare subtracted cDNA libraries. The combination of DNA arrays and subtracted cDNA libraries has been applied to the identification and cloning of genes in various biological models [12–14].

In this work, we have used a combination of subtracted cDNA libraries (RDA and SSH) and DNA arrays to clone and analyze the expression of coffee plant genes induced during coffee leaf miner infestation. To our knowledge, this is the first report on gene expression in coffee plants attacked by *L. coffeella*.

2. Materials and methods

2.1. Insect rearing

Leucoptera coffeella were reared on seedlings of *C. arabica* on a 16:8 light/dark photoperiod, at a temperature of $29 \pm 2^\circ\text{C}$ and a relative humidity of $85 \pm 5\%$. The adult insects were fed with a 10% sucrose solution.

2.2. Plant material and evaluation of coffee plant resistance to *L. coffeella*

Coffee plants were obtained from the Agronomic Institute of Campinas (IAC), Campinas, São Paulo, Brazil. The experiments to assess leaf miner resistance were done with progeny H14954-29 plants, from the fifth backcross generation of the $(C. racemosa \times C. arabica) \times C. arabica$ breeding program [4]. Isolated leaves were infested in rearing cages [4]. On the following day, 1.8 cm leaf discs were taken from the leaves using a cork bore. The discs were placed on damp plastic foam and maintained in a plastic box for two weeks. *Coffea arabica* cultivar Catuaí vermelho leaves were used as a susceptible control. *Coffea arabica* cultivar Mundo Novo plants were used to assess the organ-specific expression of coffee genes. Leaves, roots, different phases of flower development (green flower buds, white flower buds and open flowers) and fruits in different stages of development (early green, green, green–yellow, yellow, yellow–red and red) were collected and immediately frozen in nitrogen liquid.

2.3. Leaf miner infestation

Two infestation experiments were done. In each experiment, 60 young leaves from highly susceptible plants (S) and 60 leaves from highly resistant plants (R) were detached from the corresponding plants. Each leaf was inserted by the petiole into a 2 ml microcentrifuge tube filled with water. The tubes were then fixed to a plastic support and 40 S and 40 R leaves were placed in *L. coffeella* rearing cages. The other support containing 20 S and 20 R leaves was incubated in a plastic box without coffee leaf miners. Cork

bores 3–8 cm in diameter were used to collect leaf discs with eggs or with mines on the leaf surface. Three days after infestation, *L. coffeella* eggs were removed from 20 S leaves (So: susceptible after oviposition) and 20 R leaves (Ro: resistant after oviposition), and the respective leaf discs were collected. After egg hatching (i.e. four days after Ro/So collection), the larvae were removed from the mines of the remaining S leaves and leaf discs from S leaves (Se: susceptible after larval eclosion) and R leaves (Re: resistant after eclosion) were collected. Leaf discs from control S (Sc: susceptible control) and control R (resistant control) leaves were also collected. All leaf discs were immediately frozen in liquid nitrogen after collection.

2.4. RNA extraction

Total RNA was isolated as previously described [15], with slight modifications. Leaf discs were ground in liquid nitrogen and 0.2 g of tissue was resuspended in 700 μ l of RNA extraction buffer (8 M guanidine-HCl, 50 mM Tris-HCl pH 8.0, 20 mM EDTA pH 8.0, 50 mM β -mercaptoethanol) and 700 μ l of phenol/chloroform/isoamyl alcohol (25:24:1, v/v/v). The aqueous phase was recovered and extracted in 700 μ l of phenol/chloroform/isoamyl alcohol (25:24:1, v/v/v). The supernatant was then resuspended in 700 μ l of chloroform/isoamyl alcohol (24:1, v/v). The last supernatant was precipitated with 3 M sodium acetate, pH 5.2 (0.2 v) and ethanol absolute (1 v) at -20°C for 16 h. After centrifugation, the pellets were washed in 70% ethanol and resuspended in diethylpyrocarbonate-treated water.

2.5. Construction of subtracted cDNA libraries

Double-stranded cDNA from tester (Ro + Re) and driver (Rc + So + Se) RNA samples was obtained using the Smart cDNA synthesis kit (Clontech, USA). RDA was done as described previously [16] and the double-stranded cDNA was digested with *Mbo*I. Three rounds of subtraction (DP1, DP2 and DP3) were executed [16]. SSH was done with the PCR-select subtraction kit, according to the manufacturer's recommendations (Clontech, USA). cDNA fragments were purified with GFX columns (ApBiotech, USA). Both SSH and RDA cDNAs were cloned into the SK II⁻ pBluescript vector and transformed in DH5 α *E. coli* cells. Recombinant clones identified on X-gal-containing plates were picked into individual wells of 96-well trays containing 0.15 ml of Circle Grow broth (BIO 101, USA) in 8% glycerol and allowed to grow at 37°C . Plasmidial DNA was purified, resuspended in water and stored in 96-well trays at -20°C .

2.6. DNA array preparation, hybridization and analysis

Sixteen 96-well plates containing plasmid clones of subtracted cDNA libraries were selected (eight from RDA

and eight from SSH). The plasmid DNA was denatured in 0.2 N NaOH for 15 min at 37°C and spotted onto Hybond-N filters (Amersham, USA) with a hand-held 96-pin printhead tool (V&P Scientific, San Diego, USA), which applies 0.1 μ l of DNA solution (approximately 5 ng of DNA). The set of eight 96-well plates was spotted on different nylon filters in a 4×4 array configuration (768 ESTs per filter). The DNA was fixed to the filters by baking at 80°C for 2 h and the filters were then stored at room temperature until hybridization. Variations in the amount of DNA on the spots were estimated by hybridizing the filters with a [α - ^{32}P]dCTP labeled oligonucleotide probe that recognized the sequence of the *Amp*^r gene of the plasmid vector, according to McPherson (<http://www.tree.caltech.edu/protocols/overgo.html>). The filters were then exposed to imaging plates for 96 h and scanned in a Phosphorimager FLA3000-G (Fujifilm, Japan). The oligo vector probe was stripped from the filters by boiling in 0.1% SDS. cDNA probes were produced as described previously [17]. Fifteen micrograms of total RNA was reverse transcribed in a final volume of 25 μ l with Superscript II RT (Invitrogen, USA) using 3 μ M of an oligo-dT18V primer, unlabeled dATP, dGTP, and dTTP (1 mM each) and 3000 Ci mmol⁻¹ of [α - ^{32}P]dCTP at 42°C for 20 min. Subsequently, unlabeled dCTP was added to a final concentration of 1 mM and, after 40 min, the probes were purified using ProbeQuant G-50 microcolumns (ApBiotech, USA).

Filter arrays were prehybridized in a solution containing 50% formamide, 5 $\times$ SSC, 5 $\times$ Denhardt's solution, 50 mM sodium phosphate pH 6.8, 1% SDS, and 100 μ g/ml of denatured salmon sperm DNA, for 4 h at 42°C . Thereafter, the filters were hybridized with denatured purified cDNA probes in hybridization solution (50% formamide, 5 $\times$ SSC, 50 mM sodium phosphate, pH 6.8, 1% SDS, 5% dextran) for 16 h at 42°C . The arrays were washed in 0.2 $\times$ SSC with 0.5% SDS for 20 min at room temperature, twice in 0.5 $\times$ SSC with 0.5% SDS for 20 min at 65°C , and 0.2 $\times$ SSC with 0.5% SDS for 20 min at 65°C . The filters were sealed with a 29 μ m thick plastic film, and then exposed to imaging plates for 96 h and scanned in a Phosphorimager. The signal intensities were quantified using Array Vision software (Imaging Research, Canada).

The local background was subtracted from each spot. The median of the background intensities of each membrane was calculated by measuring the intensity of 12 spots positioned on filter areas without cDNA spots. This value was used to normalize the signals from each gene (see text for details). Normalized signal intensities were log₂-transformed, and for each gene a one-way analysis of variance (ANOVA) was applied to the transformed data with a Bonferroni adjusted *p*-value of 3.26E-05 (0.05/1536 = *P*/number of ESTs) to correct for multiple tests. The intensity of duplicate DNA spots was averaged and used to calculate the expression ratios between treatments. Selected ESTs were assembled using a hierarchical clusterization method [18] according to their expression profiles.

2.7. RNA blot

Ten micrograms of total RNA was incubated in denaturing RNA buffer [19] for 15 min at 56 °C, fractionated in 1% agarose gel containing formaldehyde, and blotted onto a Hybond-N+ filter (ApBiotec, USA). The blots were hybridized at 42 °C with the cDNA inserts labeled with [α -³²P]dCTP [19]. The filters were washed under high stringency conditions and then exposed to imaging plates. Digitalized images of the RNA blots signals were quantified using Image Gauge software (Fujifilm, Japan).

2.8. DNA sequencing and sequence analysis

cDNA clones displaying differential expression were sequenced in an ABI Prism 3700 sequencer (Applied Biosystems, USA). Virtual translations of the cDNA clone sequences were analyzed by BLASTX [20] to identify similarities in the GenBank database. *E*-values higher than 1.0E–5 were not considered statistically significant (no significant match). For conserved domain analysis of virtual translations, the Pfam [21] and SMART [22] databases were used. The accession numbers of the coffee ESTs (Expressed Sequence Tags) cited in this manuscript are available in GenBank: (CK484622–CK484627; CV998038–CV998046; CV998048–CV99050; CX068758–CX068760). The array data described in this manuscript have been submitted to Gene Expression Omnibus (GEO), under the accession numbers GPL1730 (platform), GSM37179 and GSM37206 to GSM37210 (samples), and GSE2045 (series).

3. Results and discussion

3.1. Evaluation of leaf miner resistant phenotypes

Four hundred plants from the fifth backcross generation of H14954-29 progeny were evaluated. One leaf from each plant was detached and incubated in a rearing cage containing leaf miner moths. Two weeks after egg hatching, four types of phenotypes were found (Fig. 2). In the first, leaves showed occasional dead eggs, with small lesions below them. In the second phenotype, the leaf miner had managed to eclose, but died within thin, irregular, filiform mines. Since these leaf miners had not completed their life cycle, we considered that these phenotypes were derived from highly resistant and moderately resistant plants, respectively. In the third phenotype, the leaves contained mines with large irregular shapes in which the caterpillar managed to pupate; such plants were considered as moderately susceptible. In the fourth phenotype, leaves with a very similar phenotype to that found in susceptible *C. arabica* control cultivars were considered as being derived from highly susceptible plants. A second infestation experiment produced the same pattern of plant phenotypes found in the first experiment.

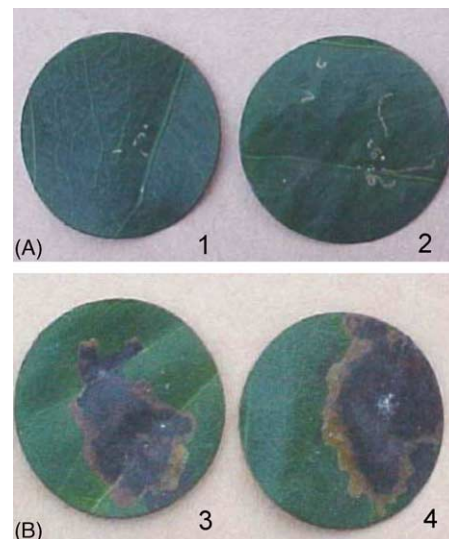


Fig. 2. Evaluation of the resistance of coffee leaves infested by *Leucoptera coffeella*; (A) infested leaves showing total (1) or moderate (2) resistance to coffee leaf miner; (B) infested leaves showing moderate (3) or total (4) susceptibility to coffee leaf miner.

3.2. *Leucoptera coffeella* infestation and construction of cDNA subtracted libraries

After assessing the resistance to leaf miners, highly resistant (R) and highly susceptible (S) coffee plants were chosen for further experiments. Based on stipulated treatments (see Section 2), leaf discs from the area surrounding the site of leaf miner oviposition or leaf damage were collected and frozen in liquid nitrogen. The RNAs were assembled into two pools to construct cDNA-subtracted libraries enriched in genes preferentially up-regulated in resistant plants following infestation by *L. coffeella*. Thus, the tester pool contained RNA from Ro and Re, and the driver pool contained RNA from Rc, So and Se (see Section 2).

3.3. Macroarray hybridization and selection of up-regulated genes

Two sets of high-density nylon filters were constructed, one with ESTs derived from the RDA library and one with ESTs from the SSH library. Each filter contained 768 ESTs that were spotted at two positions on the filters to verify spotting reproducibility. The high-density filters were hybridized with a [α -³²P]dCTP labeled probe corresponding to the plasmid vector. Replicate spots that varied in intensity by more than two-fold were discarded. A 0.9 correlation coefficient between the signals of replicate spots from different filters was obtained, indicating that most of the replicate spots had little variation in the amount of DNA of applied. In most studies, the median of the spot intensities is used as a factor for normalization. However, in subtracted cDNA libraries, some treatments will produce intensities

Table 1

Sequence similarities and normalized intensities of coffee ESTs up-regulated during *L. coffeella* infestation

ESTs	Expression ^b						Best Blast hit ^c	E-value	Categorization
	Rc	Ro	Re	Sc	So	Se			
Group 1 ^a									
In Ro and So									
SSH101B04 (CK484622) ^d	12.20	44.44	8.53	8.51	18.60	3.40	Tumor-related protein Miraculin-like <i>Nicotiana tabacum</i> (T03803)	3.00E-11	Defense and stress
SSH104G04 (CK484625)	11.81	37.13	7.48	5.40	15.73	3.07	Putative potassium-dependent sodium calcium exchanger -like protein <i>Oryza sativa</i> (XP_470536)	2.00E-18	Membrane exchanger
SSH202A01 (CV998043)	14.64	36.49	7.88	4.52	17.63	3.43	Cytochrome b6 <i>Psilotum nudum</i> (NP_569658)	3.00E-25	Photosynthesis
SSH201D05 (CX068758)	40.65	84.42	25.42	5.72	72.32	20.44	No protein match	–	Unknown function
In Ro									
SSH101H05 (CX068759)	17.16	67.71	11.37	15.58	20.90	2.63	No protein match	–	Unknown function
SSH201D09 (CK484624)	34.82	125.33	17.15	26.60	41.11	4.03	Probable microsomal signal peptidase 25 kDa subunit <i>Arabidopsis thaliana</i> (P58684)	4.00E-13	Protein regulation
SSH202H06 (CK484626)	42.59	144.85	21.79	35.05	51.97	4.66	Photosystem I psaH precursor <i>Nicotiana sylvestris</i> (T16958)	1.00E-12	Photosynthesis
SSH203H02 (CV998044)	45.22	146.11	32.72	35.30	56.99	20.05	4-Diphosphocytidyl-2C-methyl-D-erythritol synthase <i>Arabidopsis thaliana</i> (AAF61714)	8.00E-38	Secondary metabolism
In So									
DP201C11 (CX068760)	14.75	11.04	9.33	2.20	5.58	3.22	Putative glycine rich protein <i>Arabidopsis thaliana</i> (BAC42653)	1.00E-14	Defense and stress
DP201D12 (CV998046)	5.87	10.85	3.89	0.87	5.58	3.15	Acid phosphatase <i>Glycine Max</i> (CAA11075)	4.00E-06	Defense and stress
SSH202C10 (CK484627)	4.60	5.18	3.90	5.57	16.05	3.75	BEL1-related homeotic protein 29 <i>Solanum tuberosum</i> (AAN03626)	2.00E-74	Nuclear factor
SSH202G10 (CV998045)	1.82	1.70	1.96	1.08	5.02	1.09	Expressed protein <i>Arabidopsis thaliana</i> (NP_568392)	4.00E-29	Unknown function
SSH202G11 (CV998048)	8.70	11.71	11.83	3.20	16.63	3.81	RRS1 Ribosome biogenesis regulatory protein homolog <i>Arabidopsis thaliana</i> (AAD32772)	5.00E-47	Protein regulation
Group 2									
In Se and So									
DP103H03 (CV998049)	2.70	3.06	1.79	0.35	2.94	3.28	Putative nodule membrane protein <i>Medicago sativa</i> (AAL57201)	4.00E-31	Unknown function
DP104C02 (CV998041)	9.97	13.21	9.45	2.05	8.86	25.03	Lipoxygenase LOX-3 <i>Solanum tuberosum</i> (T07775)	4.00E-43	Defense and stress
DP301B04 (CV998038)	4.37	5.99	3.68	0.90	3.60	11.14	Lea1P <i>Daucus carota</i> (AAG24641)	2.00E-21	Defense and stress
DP302A12 (CV998050)	1.70	2.57	1.33	0.22	3.10	3.37	Lipid transfer/trypsin inhibitor/seed storage protein <i>Arabidopsis thaliana</i> (NP_190966)	6.00E-18	Defense and stress
In Se									
SSH101G06 (CV998039)	16.94	28.65	20.32	10.34	13.84	61.18	Phospholipase D <i>Lycopersicon esculentum</i> (AAG48162)	2.00E-54	Defense and stress
SSH201B02 (CV998040)	15.44	22.37	13.44	7.27	13.84	46.96	Putative transcriptional regulator FACT-like <i>Arabidopsis thaliana</i> (NP_192809)	7.00E-29	Nuclear factor
In Se, Ro and So									
DP302B11 (CK484623)	4.92	19.08	7.02	1.58	5.52	23.01	Class III chitinase <i>Lupinus albus</i> (CAA76203)	1.00E-30	Defense and stress

^a ESTs groups refer to hierarchical clustering of ESTs according to their induction ratio. group 1: (Ro, So) > (Rc, Re, Se); group 2: (Se) > (Rc, Ro, Re, So).^b Expression is the ratio between the average of the intensity of replica spots and the median of the membrane background. ESTs were considered up-regulated when the ratio between the expression values in the treatments and control in resistant or susceptible plants (Ro/Rc, Re/Rc, So/Sc, Se/Sc) was >2.0 in at least one treatment.^c Indicates the best BLAST hit of each EST sequence obtained using the BLASTX algorithm.^d Refers to accession numbers in Genbank.

much higher than the others because the genes were not randomly selected. Therefore, the median of the background intensities of each membrane was used to normalize the spot intensities [23].

In selecting of differentially expressed genes, we used $\log_2$ transformed normalized signal intensities assessed by ANOVA. Using this approach, 157 ESTs (around 10% of the subtracted libraries ESTs; 12.3% of RDA ESTs and 7.5% of SSH ESTs) were assumed to be differentially expressed in at least one treatment. Additionally, we selected ESTs for which the ratio between the average of the normalized intensity of these genes in each treatment and the respective intensities in the resistant control (Ro or Re/Rc) or susceptible control (So or Se/Sc) was >2.0 . These ESTs were assembled by hierarchical clustering, according to their expression profiles (data not shown). Two groups contained ESTs that had interesting patterns of expression: group 1 contained ESTs up-regulated in Ro and So, whereas group 2 contained ESTs preferentially induced in Se. Finally, these ESTs were sequenced and analyzed by BLASTX. As a result, 13 ESTs from group 1 and 8 ESTs from group 2 were selected as differentially expressed, with 2 of them (EST SSH201D05 and EST SSH101H05) having no matches in GenBank (Table 1).

Surprisingly, many ESTs were induced in the susceptible genotype, which could indicate that our subtraction was not accurate. However, our results may reflect the ineffectiveness of RDA and SSH in profiling subtle differences in gene expression [16,24]. Other reports that have used the strategy of complex tester/driver subtraction have also documented this inefficacy [23,25].

3.4. Genes modulated by coffee leaf miner infestation

Based on BLASTX and protein domain analyses, the selected ESTs were divided into eight functional categories: defense and stress, membrane exchanger, protein regulation, nuclear factors, photosynthesis, secondary metabolism and unknown function. To validate the macroarray data, selected ESTs were used as probes in RNA blot analysis, with RNA from an independent replicate experiment (Fig. 3). One EST was similar to a gene previously described as defense-stress up-regulated (Class III chitinase), and four ESTs were similar to genes not previously described to be stress-induced (*CAX9*, *SPC25*, *BEL*, and *psaH*). Comparison of the expression patterns obtained using macroarrays and RNA blot analysis yielded a good correlation (Fig. 3), which confirmed the efficiency of macroarray analysis in identifying differentially expressed genes. We also assessed the expression pattern of these genes induced during coffee leaf miner infestation in different organs of plants from *C. arabica* cultivar Mundo Novo (Fig. 4), which is used as a parental genotype in the selection of many Brazilian cultivars; including Catuaí, the cultivar that originated the plants of the progeny analyzed in this work [4]. Since the genetic variability of *C. arabica* plants is extremely reduced

[26], we believe that the information obtained with the RNA from Mundo Novo plants can be informative on the gene expression pattern in *Coffea* plants.

EST DP302B11 encoded a polypeptide similar to PR-8 proteins, which are class III chitinases. Other reports have shown the up-regulation of PR-8 proteins by fungi [27,28], and SA and wounding [28]. Fig. 3 shows the up-regulation of EST DP302B11 in Ro and Se. This gene was most expressed in leaves, green and white flower buds, and only slightly in roots (Fig. 4A). Additionally, it was constitutively expressed throughout fruit development (Fig. 4B). As well as being important proteins in plant defense, chitinases are also involved in embryonic development [29].

A protein similar to *Arabidopsis* 25 kDa subunit of a probable microsomal signal peptidase complex SPC25 was encoded by EST SSH201D09. This protein complex cleaves the signal peptides of secretory proteins within the co-translational translocation pathway [30]. Coffee *SPC25* was up-regulated in Ro and almost maintained a steady-state level in So (Fig. 3). Curiously, polypeptides derived from defense genes up-regulated in Ro, EST DP302B11 and EST SSH101B04 (PR-8 and miraculin), contain signal peptides [28,31]. In addition, *SPC25* was expressed in young organs like green flower buds and green fruits (Fig. 4A and B).

EST SSH202H06 was up-regulated after oviposition in resistant plants (Fig. 3) and was similar to the *psaH* gene, whose product stabilizes the photosystem I (PSI) complex during electron transport [32]. *psaH* was expressed in leaves, green and white flower buds, and in early green fruits (Fig. 4A and B). The presence of a transcript in white flower buds may be related with the presence of photosystem proteins in mature flower buds [33]. EST SSH202H06 transcripts were detected only in early green fruits, which agrees with the restricted expression of photosynthetic genes in immature tomato fruits [34].

EST SSH104G04 EST was similar to a rice putative K^+ -dependent Na^+/Ca^{2+} exchanger-like gene and to *A. thaliana* putative cation exchanger *CAX9* gene. The four *Arabidopsis* CAX proteins characterized so far (CAX 1–4) drive the efflux of Ca^{2+} from the cytoplasm [35,36]. On the other hand, CAX 7–11 genes are more similar to animal and fungal Na^2+/Ca^{2+} exchangers, which can drive Ca^{2+} into cells [37]. In fact, Ca^{2+} acts as a secondary messenger in plant signal transduction responses to biotic stress [38]. *CAX9* EST was predominantly expressed in leaves and slightly expressed in early green fruits (Fig. 4A and B).

We also identified EST SSH202C10, that was similar to the potato homeotic nuclear factor *BEL1*, which is an important component in determining the fate of leaf cells [39]. This EST was induced in the So treatment (Fig. 3). EST SSH202C10 was expressed with higher levels in white flower buds (Fig. 4A) and in early green, yellow–red and red fruits (Fig. 4B). This result agrees with the role of *BEL* genes during flower [40] and fruit development [41,42].

Apart from EST DP302B11, which encoded a class III chitinase, five other ESTs identified by DNA array analysis

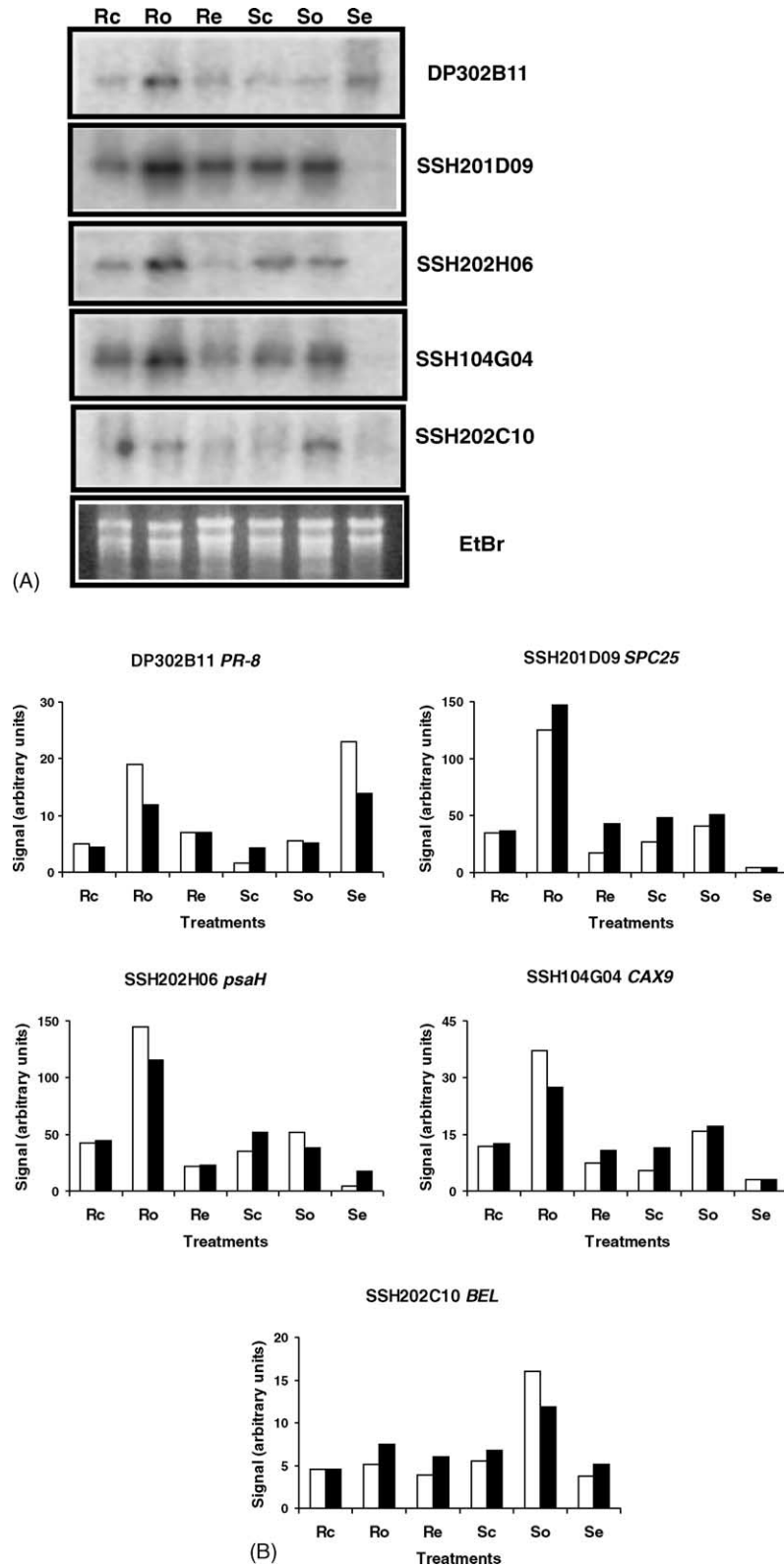


Fig. 3. Expression profiles of genes induced by leaf miner infestation; (A) RNA blot analysis of selected genes containing RNAs from the second *L. coffeella* infestation experiment; The membrane was hybridized with the probes indicated at the right of the panel. Ethidium bromide-stained rRNA (EtBr) was visualized as a control for RNA loading; (B) comparison of the expression patterns between DNA arrays (white columns) and RNA blots from the second infestation experiment (black columns). The bars represent signals detected (arbitrary units) in the hybridizations. Rc, resistant plants not infested; Ro, resistant plants after oviposition; Re, resistant plants after larval eclosion; Sc, susceptible plants not infested; So, susceptible plants after oviposition; Se, susceptible plants after larval eclosion.

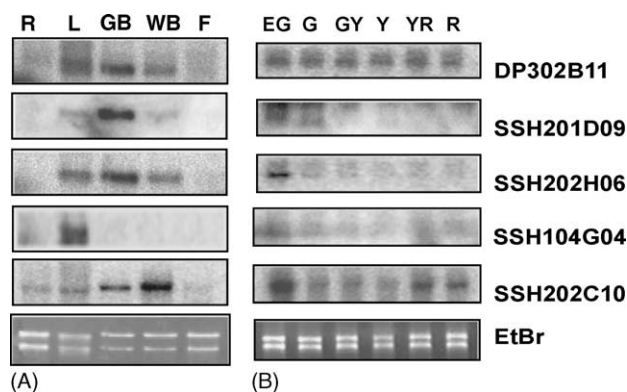


Fig. 4. Expression profiles of genes induced by leaf miner infestation in different organs of *Coffea arabica*; (A) RNA blot analysis of selected genes containing RNAs from roots (R), leaves (L), green flower buds (GB), white flower buds (WB) and open flowers (F); (B) RNA blot analysis of selected genes containing RNAs from different stages of coffee fruit development: early green (EG), green (G), green-yellow (GY), yellow (Y), yellow-red (YR) and red (R). Ethidium bromide-stained rRNA (EtBr) was visualized as a control for RNA loading.

as up-regulated during *L. coffeella* infestation, have previously been related to plant defense and biotic stress: EST SSH101G06 (similar to phospholipase D) [43], EST DP104C02 (similar to lipoxygenase LOX3) [44], EST DP301B04 (similar to LEA1) [45,46], EST DP201D12 (similar to acid phosphatase VSP) [47,48], EST DP302A12 (similar to proteins containing lipid transfer protein/trypsin inhibitor/seed storage domain) [49], EST SSH101B04 (similar to putative serine protease inhibitor miraculin) [50], and EST DP201C11 (similar to a putative GRP) [51,52].

The regulation of protein synthesis and degradation is crucial during plant stress. In addition to EST SSH201D09 (similar to signal peptidase subunit SPC25), we identified two other ESTs related to protein turnover: EST DP104G07, which was similar to tobacco α 6 20S proteasome subunit gene α -din6 [53] and EST SSH202G11 that was similar to the *Arabidopsis* ribosome regulator synthesis gene *RRS1* [54].

Besides isolating EST SSH202C10 (*BEL1*), we identified another nuclear factor not previously associated with biotic stress. ESTs SSH201B02 was similar to *Arabidopsis* nuclear factor FACT, involved in chromatin rearrangement [55].

Apart from *psaH* (EST SSH202H06), another photosynthetic gene was up-regulated after *L. coffeella* oviposition in resistant and susceptible plants and down-regulated during leaf miner herbivory (Re and Se) (Table 1). EST SSH202A01 product was similar to cytochrome b6-f [56]. Interestingly, photosynthetic genes are down-regulated by JA/herbivory and up-regulated by SA [57,58].

Additionally, we isolated EST SSH203H02 that is similar to 4-diphosphocytidyl-2C-methyl-D-erythritol synthase (plant ispD), an enzyme of the deoxyxylulose phosphate pathway that produces terpenoids, which are secondary metabolites used in defense [59]. Finally, we identified two ESTs whose products were similar to proteins of unknown function: EST SSH202G10 that is similar to an *Arabidopsis*

protein that has no domain listed in the PFAM or SMART databanks, and EST DP103H03 that is similar to a *Medicago sativa* protein that contains a START-domain, which allows proteins to attach to lipids [60].

3.5. A hypothesis to coffee leaf miner resistance

The expression profiles of the genes included in the groups chosen suggest a possible defense mechanism triggered by the *L. coffeella* attack in coffee plants. Group 1 contained ESTs up-regulated by oviposition treatments. Some of these ESTs (*psaH* and *SPC25*, for example) were only up-regulated in resistant plants after oviposition (Ro) and may provide a genotype-specific resistance after oviposition by *L. coffeella*. Curiously, no ESTs were induced in resistant plants after eclosion (Re), which reinforces the possible importance of genes induced during oviposition in arresting the development of *L. coffeella* larvae in resistant plants. On the other hand, some ESTs from group 1 (*BEL* i.e.) were up-regulated in susceptible plants after oviposition (So) (Table 1). This intriguing expression profile suggests that susceptible plants may activate a hypothetical “lack of resistance” signal after moth oviposition. Group 2 contained ESTs that were more induced in susceptible plants after larval eclosion and leaf miner herbivory (Se) (*PR-8* i.e.). Most of group 2 ESTs were also up-regulated in susceptible plants after oviposition (So) (Table 1). These results suggest that although susceptible plants activate defense genes after oviposition, this signaling is insufficient to arrest larval development.

The majority of ESTs selected had a higher steady-state level in non-infested resistant plants (Rc) than in non-infested susceptible plants (Sc) (Table 1). These ESTs were more expressed in resistant plants after oviposition (Ro) than in susceptible plants after oviposition (So) (Table 1). The higher rate of *L. coffeella* egg antibiosis seen in resistant plants may be resulted from the up-regulation of defense genes (*PR-8* i.e.) after oviposition, combined with a higher basal expression of stress-related genes in this genotype.

In conclusion, using a combination of subtracted cDNA libraries and DNA arrays we have identified 21 ESTs that are differentially expressed during the infestation of coffee plants by coffee leaf miners. Our results provide a starting point in the research of molecular mechanisms of *Coffea* resistance against the coffee leaf miner. Further studies are needed to unravel the signaling pathways involved in the resistance of coffee plants to *L. coffeella* oviposition.

Acknowledgements

We thank Edna Santos, the Plant Genomics Laboratory staff and Christian Colin for excellent technical assistance. J.M.C.M., M.H.B., J.M.F. and M.P.D. were recipients of fellowships from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). R.D.D. and D.R. were recipients of

fellowships from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Consórcio Nacional do Café (CNC), respectively. This work was supported by Consórcio Brasileiro de Pesquisa e Desenvolvimento do Café Fundo de Apoio ao Ensino e à Pesquisa (FAEP): project no. 040504, Financiadora de Estudos e Projetos (FINEP) and Pró Reitoria de Pesquisa da Universidade de São Paulo (PRP-USP).

References

- [1] A.E. Paulini, P.R. Andrade, J.B. Matiello, Z. Mansk, A.J. Paulino, Eficiência do controle do bicho mineiro do café (*Perileucoptera coffeella*, Guer-Men, 1842) e sua relação com a produtividade. In: Resumos do Congresso Brasileiro de Pesquisas Cafeeiras, 3 Curitiba, IBC/GERCA, Rio de Janeiro (1975) p. 1.
- [2] A. Carvalho, L.C. Mônico, Relaciones genéticas de espécies seleccionadas de *Coffea*, Café 9 (1968) 1–19.
- [3] H.P. Medina Filho, A.P. Carvalho, L.C. Mônico, Melhoramento do cafeeiro. XXXVII. Observações sobre a resistência do cafeeiro ao bicho mineiro, Bragantia 36 (1977) 131–137.
- [4] O. Guerreiro, M.B. Silvarolla, A.B. Eskes, Expression and mode of inheritance of resistance in coffee to leaf miner *Perileucoptera coffeella*, Euphytica 105 (1999) 7–15.
- [5] B.J. Feys, J.E. Parker, Interplay of signaling pathways in plant disease resistance, Trends Genet. 16 (2000) 449–455.
- [6] K. Maleck, R.A. Dietrich, Defense on multiple fronts: how do plants cope with diverse enemies? Trends Plant Sci. 4 (1999) 215–219.
- [7] L.L. Walling, The myriad plant responses to herbivores, J. Plant Growth Regul. 19 (2000) 195–216.
- [8] T. Desprez, J. Amselem, M. Caboche, H. Hofte, Differential gene expression in *Arabidopsis* monitored using cDNA arrays, Plant J. 14 (1998) 643–652.
- [9] M. Tamaoki, N. Nakajima, A. Kubo, M. Aono, T. Matsuyama, H. Saji, Transcriptome analysis of O3-exposed *Arabidopsis* reveals that multiple signal pathways act mutually antagonistically to induce gene expression, Plant Mol. Biol. 53 (2003) 443–456.
- [10] L. Diatchenko, Y.F.C. Lau, A.P. Campbell, A. Chenchik, F. Moqadam, B. Huang, S. Lukyanov, K. Lukyanov, N. Gurskaya, E.D. Sverdlov, P.D. Siebert, Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries, Proc. Natl. Acad. Sci. U.S.A. 93 (1996) 6025–6030.
- [11] M. Hubank, D.G. Schatz, Identifying differences in mRNA expression by representational difference analysis of cDNA, Nucleic Acids Res. 22 (1994) 5640–5648.
- [12] M. Matvienko, M.J. Torres, J.I. Yoder, Transcriptional responses in the hemiparasitic plant *Triphysaria versicolor* to host plant signals, Plant Physiol. 127 (2001) 272–282.
- [13] L. Xiong, M.W. Lee, M. Qi, Y. Yang, Identification of defense-related rice genes by suppression subtractive hybridization and differential screening, Mol. Plant Microbe Interact. 14 (2001) 685–692.
- [14] J. Zheng, J. Zhao, Y. Tao, J. Wang, Y. Liu, J. Fu, Y. Jin, P. Gao, J. Zhang, Y. Bai, G. Wang, Isolation and analysis of water stress induced genes in maize seedlings by subtractive PCR and cDNA macroarray, Plant Mol. Biol. 55 (2004) 807–823.
- [15] J. Logemann, J. Schell, L. Willmitzer, Improved method for the isolation of RNA from plant tissues, Anal. Biochem. 163 (1987) 16–20.
- [16] M. Hubank, D.G. Schatz, Representational difference analysis of cDNA, in: S.P. Hunt, R. Levesey (Eds.), Functional Genomics: A Practical Approach, Oxford University Press, 2000, pp. 46–80.
- [17] M. Schummer, V.L.V. Ng, R.E. Baumgarner, P.S. Nelson, B. Schummer, D.W. Bednarski, L. Hassell, R.L. Baldwin, B.Y. Karlan, L. Hood, Comparative hybridization of an array of 21,500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas, Gene 238 (1999) 375–385.
- [18] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. U.S.A. 95 (1998) 14863–14868.
- [19] J. Sambrook, E.F. Fritsch, T. Maniatis, Molecular Cloning: A Laboratory Manual, second ed., Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1989.
- [20] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
- [21] A. Bateman, E. Birney, R. Durbin, S.R. Eddy, K.L. Howe, E.L. Sonnhammer, The Pfam protein families database, Nucleic Acids Res. 28 (2000) 263–266.
- [22] J. Schultz, R.R. Copley, T. Doerks, C.P. Ponting, P. Bork, SMART: a web-based tool for the study of genetically mobile domains, Nucleic Acids Res. 28 (2000) 231–234.
- [23] M. Frohme, B. Scharm, H. Delius, R. Knecht, J.D. Hoheisel, Use of representational difference analysis and cDNA arrays for transcriptional profiling of tumor tissue, Ann. N. Y. Acad. Sci. 910 (2000) 85–104.
- [24] W. Ji, M.B. Wright, L. Cai, A. Flament, K. Lindpaintner, Efficacy of SSH PCR in isolating differentially expressed genes, BMC Genomics 3 (2002) 12.
- [25] K. Schroeyers, C. Chaparro, S. Goormachtig, M. Holsters, Nodulation-enhanced sequences from the water stress-tolerant tropical legume *Sesbania rostrata*, Plant Sci. 167 (2004) 207–216.
- [26] P. Lashermes, M.C. Combes, J. Robert, P. Trouslot, A. D'Hont, F. Anthony, A. Charrier, Molecular characterisation and origin of the *Coffea arabica* L. genome, Mol. Gen. Genet. 261 (1999) 259–266.
- [27] K.K. Nielsen, J.D. Mikkelsen, K.M. Kragh, K. Bojsen, An acidic class III chitinase in sugar beet: induction by *Cercospora beticola*, characterization, and expression in transgenic tobacco plants, Mol. Plant Microbe Interact. 6 (1993) 495–506.
- [28] A.P. Regalado, C. Pinheiro, S. Vidal, I. Chaves, C.P. Ricardo, C. Rodrigues-Pousada, The *Lupinus albus* class-III chitinase gene, IF3, is constitutively expressed in vegetative organs and developing seeds, Planta 210 (2000) 543–550.
- [29] A.J. van Hengel, F. Guzzo, A. van Kammen, S.C. de Vries, Expression pattern of the carrot EP3 endochitinase genes in suspension cultures and in developing seeds, Plant Physiol. 117 (1998) 43–53.
- [30] M. Paetzel, A. Karla, N.C. Strynadka, R.E. Dalbey, Signal peptidases, Chem. Rev. 102 (2002) 4549–4580.
- [31] E.D. Brenner, K.N. Lambert, I. Kaloshian, V.M. Williamson, Characterization of *LeMir*, a root-knot nematode-induced gene in tomato with an encoded product secreted from the root, Plant Physiol. 118 (1998) 237–247.
- [32] H.V. Scheller, P.E. Jensen, A. Haldrup, C. Lunde, J. Knoetzel, Role of subunits in eukaryotic photosystem I, Biochim. Biophys. Acta 1507 (2001) 41–60.
- [33] Y. Yamamoto, H. Tsuji, J. Obokata, Structure and expression of a nuclear gene for the PSI-D subunit of photosystem I in *Nicotiana sylvestris*, Plant Mol. Biol. 22 (1993) 985–994.
- [34] M. Sugita, W. Gruissem, Developmental, organ-specific, and light-dependent expression of the tomato ribulose-1,5-bisphosphate carboxylase small subunit gene family, Proc. Natl. Acad. Sci. U.S.A. 84 (1987) 7104–7108.
- [35] T. Shigaki, K. Hirschi, Characterization of CAX-like genes in plants: implications for functional diversity, Gene 257 (2000) 291–298.
- [36] N.H. Cheng, J.K. Pittman, T. Shigaki, K.D. Hirschi, Characterization of CAX4, an *Arabidopsis* H(+)/cation antiporter, Plant Physiol. 128 (2002) 1245–1254.
- [37] K.D. Philipson, D.A. Nicoll, Sodium-calcium exchange: a molecular perspective, Ann. Rev. Physiol. 62 (2000) 111–133.
- [38] D. Sanders, J. Pelloux, C. Brownlee, J.F. Harper, Calcium at the crossroads of signaling, Plant Cell 14 (2002) 401–417.

- [39] H. Chen, F.M. Rosin, S. Prat, D.J. Hannapel, Interacting transcription factors from the three-amino acid loop extension superclass regulate tuber formation, *Plant Physiol.* 132 (2003) 1391–1404.
- [40] L. Reiser, Z. Modrusan, L. Margossian, A. Samach, N. Ohad, G.W. Haughn, R.L. Fischer, The *BELL1* gene encodes a homeodomain protein involved in pattern formation in the *Arabidopsis* ovule primordium, *Cell* 83 (1995) 735–742.
- [41] Y.H. Dong, J.L. Yao, R.G. Atkinson, J.J. Putterill, B.A. Morris, R.C. Gardner, *MDH1*: an apple homeobox gene belonging to the BEL1 family, *Plant Mol. Biol.* 42 (2000) 623–633.
- [42] A.H. Roeder, C. Ferrandiz, M.F. Yanofsky, The role of the REPLUMLESS homeodomain protein in patterning the *Arabidopsis* fruit, *Curr. Biol.* 13 (2003) 1630–1635.
- [43] X. Wang, Phospholipase D in hormonal and stress signaling, *Curr. Opin. Plant Biol.* 5 (2002) 408–414.
- [44] G.A. Howe, A.L. Schillmiller, Oxylin metabolism in response to stress, *Curr. Opin. Plant Biol.* 5 (2002) 230–236.
- [45] L.M. Weaver, S. Gan, B. Quirino, R.M. Amasino, A comparison of the expression patterns of several senescence-associated genes in response to stress and hormone treatment, *Plant Mol. Biol.* 37 (1998) 455–469.
- [46] S.E. Keates, T.A. Kostman, J.D. Anderson, B.A. Bailey, Altered gene expression in three plant species in response to treatment with Nep1, a fungal protein that causes necrosis, *Plant Physiol.* 132 (2003) 1610–1622.
- [47] S. Berger, T. Mitchell-Olds, H.U. Stotz, Local and differential control of vegetative storage protein expression in response to herbivore damage in *Arabidopsis thaliana*, *Physiol. Plant.* 114 (2002) 85–91.
- [48] H.S. Mason, D.B. DeWald, R.A. Creelman, J.E. Mullet, Coregulation of soybean vegetative storage protein gene expression by methyl jasmonate and soluble sugars, *Plant Physiol.* 98 (1992) 859–867.
- [49] J.P. Blein, P. Coutos-Thevenot, D. Marion, M. Ponchet, From elicitors to lipid-transfer proteins: a new insight in cell signalling involved in plant defence mechanisms, *Trends Plant Sci.* 7 (2002) 293–296.
- [50] S. Theerasilp, Y. Kurihara, Complete purification and characterization of the taste-modifying protein, miraculin, from miracle fruit, *J. Biol. Chem.* 263 (1988) 11536–11539.
- [51] A. Molina, M. Mena, P. Carbonero, F. Garcia-Olmedo, Differential expression of pathogen-responsive genes encoding two types of glycine-rich proteins in barley, *Plant Mol. Biol.* 33 (1997) 803–810.
- [52] E. Yasuda, H. Ebinuma, H. Wabiko, A novel glycine-rich/hydrophobic 16 kDa polypeptide gene from tobacco: similarity to proline-rich protein genes and its wound-inducible and developmentally regulated expression, *Plant Mol. Biol.* 33 (1997) 667–678.
- [53] J. Dahan, P. Etienne, A.S. Petitot, V. Houot, J.P. Blein, L. Suty, Cryptogein affects expression of alpha3, alpha6 and beta1 20S proteasome subunits encoding genes in tobacco, *J. Exp. Bot.* 52 (2001) 1947–1948.
- [54] A. Tsuno, K. Miyoshi, R. Tsujii, T. Miyakawa, K. Mizuta, *RRS1*, a conserved essential gene, encodes a novel regulatory protein required for ribosome biogenesis in *Saccharomyces cerevisiae*, *Mol. Cell. Biol.* 20 (2000) 2066–2074.
- [55] R. Belotserkovskaya, D. Reinberg, Facts about FACT and transcript elongation through chromatin, *Curr. Opin. Genet. Dev.* 14 (2004) 139–146.
- [56] M. Hervás, J.A. Navarro, M.A. De La Rosa, Electron transfer between membrane complexes and soluble proteins in photosynthesis, *Acc. Chem. Res.* 36 (2003) 798–805.
- [57] P.M. Schenk, K. Kazan, I. Wilson, J.P. Anderson, T. Richmond, S.C. Somerville, J.M. Manners, Coordinated plant defense responses in *Arabidopsis* revealed by microarray analysis, *Proc. Natl. Acad. Sci. U.S.A.* 97 (2000) 11655–11660.
- [58] D. Hermsmeier, U. Schittko, I.T. Baldwin, Molecular interactions between the specialist herbivore *Manduca sexta* (Lepidoptera, Sphingidae) and its natural host *Nicotiana attenuata*. I. Large-scale changes in the accumulation of growth- and defense-related plant mRNAs, *Plant Physiol.* 125 (2001) 683–700.
- [59] W. Eisenreich, F. Rohdich, A. Bacher, Deoxyxylulose phosphate pathway to terpenoids, *Trends Plant Sci.* 6 (2001) 78–84.
- [60] K. Schrick, D. Nguyen, W.M. Karlowski, K.F. Mayer, START lipid/sterol-binding domains are amplified in plants and are predominantly associated with homeodomain transcription factors, *Genome Biol.* 5 (2004) R41.

Capítulo 4: Testes estatísticos específicos para dados de arranjos de DNA

O grande avanço alcançado com o desenvolvimento da tecnologia de arranjos de DNA é a capacidade de analisar a expressão de milhares de genes em paralelo. A grande quantidade de dados gerados, porém, torna-se um problema no momento de analisá-los a procura de genes diferencialmente expressos. Como visto no capítulo anterior, mesmo adotando-se uma pequena margem de erro na aplicação de testes estatísticos pode-se encontrar uma grande quantidade de falsos-positivos. Além disso, a análise dos dados de arranjos de DNA costuma ser muito dificultada pela pequena quantidade de replicações, usual nos experimentos. Tornou-se então necessário desenvolver testes estatísticos mais adequados aos dados de arranjos de DNA, que tirassem proveito das suas características e assim fossem mais eficientes na seleção de genes diferencialmente expressos.

4.1 - Seleção de genes baseada na distribuição global dos logaritmos das razões de expressão

Uma suposição geralmente empregada na análise dos dados de arranjos é que a grande maioria dos genes presentes nos arranjos apresenta níveis de expressão constantes nas amostras estudadas. Essa hipótese pode ser considerada válida na maior parte dos experimentos, quando os genes nos arranjos representam todo o genoma (ou todo o transcrito) do organismo

estudado ou quando são selecionados ao acaso. Como visto no Capítulo 2, essa suposição pode ser usada na normalização dos dados pela média, mediana ou pela aplicação do algoritmo *lowess*. Além disso, ela pode desempenhar um papel fundamental na seleção de genes diferencialmente expressos.

Quando é válido assumir que a maioria dos genes representados nos arranjos não é diferencialmente regulada entre as amostras estudadas, pode-se supor também que as razões de expressão obtidas a partir de uma hibridação estão centradas em torno de 1, ou seja, que o logaritmos dessas razões estão centrados em torno de 0. Como esses logaritmos em geral apresentam distribuição normal (Friddle *et al.*, 2000), basta então calcular seu desvio padrão para obter uma boa aproximação de sua distribuição ($N(\mu, \sigma)$, onde μ e σ são respectivamente a média e o desvio padrão obtidos dos dados). Supondo-se ainda que, como a fração de genes diferencialmente expressos é pequena, ela não chega a interferir significativamente nessa distribuição, podemos tomá-la como sendo a distribuição sob a hipótese nula (isto é, a distribuição sob a hipótese de expressão constante entre as amostras comparadas). Assim, os genes que se encontram nas caudas superior e inferior dessa distribuição podem ser selecionados como diferencialmente expressos. Por exemplo, adotando uma margem de erro de 5% podemos selecionar como induzidos aqueles genes cujos logaritmos da razão de expressão fica acima de $\mu + 1,96\sigma$, e como reprimidos aqueles cujos logaritmos ficam abaixo de $\mu - 1,96\sigma$, onde o valor 1,96 é obtido a partir de uma tabela de valores críticos da distribuição normal. No caso de haver

réplicas do experimento, é possível selecionar os genes que se destacam desta forma em todas as hibridações.

Vale observar que a estratégia de análise descrita acima é um tanto ao quanto simplista, não levando em consideração as variações na média e no desvio dos logaritmos dependentes da intensidade média dos sinais dos genes. Na verdade, essa estratégia é apenas uma sofisticação do critério de análise de arranjos de DNA que consiste simplesmente em definir um valor limite para as razões de expressão, acima do qual os genes são selecionados como diferencialmente expressos, com a diferença que nela os valores de corte não são arbitrados *a priori*, e sim determinados a partir dos dados. No entanto, quando há poucas réplicas do experimento (ou nenhuma réplica), a aplicação de outros testes estatísticos aos dados se torna inadequada, e essa estratégia configura uma boa opção para extrair conclusões dos mesmos. Nogueira et al. (2003) adotou essa estratégia para analisar dados provenientes de arranjos de DNA em membranas de *nylon*, num estudo voltado para a identificação de genes de cana-de-açúcar regulados durante a exposição a baixas temperaturas.

4.2 - Estratégias de controle do número de falsos-positivos

Devido ao grande número de genes usualmente presentes em arranjos de DNA, a aplicação de testes estatísticos aos dados provenientes desses experimentos pode levar à seleção de um grande número de falsos-positivos, isto é, de genes erroneamente selecionados como diferencialmente expressos. Como descrito no Capítulo 3, uma maneira de controlar a ocorrência de falsos positivos

em múltiplos testes feitos em paralelo é através da correção de Bonferroni, que consiste em tornar os testes estatísticos tão estritos que a probabilidade de selecionar erroneamente qualquer gene torna-se menor que a margem de erro escolhida (essa probabilidade costuma ser chamada de *family-wise error rate* - *FWER*). No entanto, esse procedimento ainda é pouco adequado aos dados de arranjos de DNA, nos quais, devido ao grande número de genes analisados e à presença de diversas fontes de ruído, sua aplicação torna os testes tão estritos que apenas um pequeno número de genes é selecionado. Como uma alternativa ao controle do FWER, Benjamini e Hochberg (1995) propuseram o conceito de *false discovery rate* (FDR), definido como sendo a fração esperada de falsos-positivos entre os genes selecionados. Para tornar mais claro, suponhamos que estamos testando m genes em paralelo, dos quais m_0 estão sob hipótese nula, isto é, têm expressão constante entre as amostras analisadas e $m_1 = m - m_0$ são diferencialmente expressos. Suponhamos também que os resultados encontrados após a aplicação de um teste estatístico sejam como na Tabela 1 abaixo.

Tabela1: Número de genes selecionados corretamente e falsos-positivos

	Declarados não significantes	Declarados significantes	Total
Genes sob hipótese nula	U	V	m_0
Genes diferencialmente expressos	T	S	m_1
Total	$m - R$	R	m

De acordo com os valores na Tabela 1, o FWER seria a probabilidade de $V \geq 1$, ou seja, a probabilidade de haver genes erroneamente declarados como significantes dentre os selecionados pelo teste. Por outro lado, o FDR seria o valor esperado de V/R , isto é, a frequência esperada de falsos-positivos entre os genes selecionados. A grande vantagem de se procurar controlar o FDR, ao invés do FWER, na análise de dados de arranjos de DNA é que usualmente os genes selecionados terão seu perfil de expressão confirmado em experimentos subseqüentes (como *Northern blot* ou PCR em tempo real), e portanto pode-se tolerar a seleção incorreta de um pequeno número de genes. Por exemplo, é bem mais interessante num experimento desse tipo obter uma seleção de vinte genes, dos quais espera-se que dois estejam erroneamente classificados, do que uma seleção de apenas cinco genes onde espera-se (com 95% de certeza) que não haja erros. No primeiro caso, os vinte genes irão passar por validação e então os dois falsos-positivos serão descartados, restando dezoito genes corretamente selecionados, enquanto no segundo caso, na melhor das hipóteses, os cinco genes serão confirmados e teremos apenas essa pequena seleção de diferencialmente expressos.

4.3 - Significance Analysis of Microarrays (SAM)

Assim com a escolha do teste estatístico a ser aplicado aos dados de arranjos de DNA, a formulação de estratégias para controlar o número de falsos-positivos selecionados é um ponto delicado da análise. Apesar de já ser opinião

corrente entre a comunidade científica que o controle do FDR é mais adequado aos dados de arranjos de DNA que o controle do FWER (Pounds e Morris, 2003), calcular o FDR não é uma tarefa trivial. O seu valor depende de uma estimativa para o número de genes erroneamente classificados, o que a princípio não é conhecido. Tusher *et al.*, (2001) propôs um método estatístico (*Significance analysis of microarrays* - SAM) adaptado especificamente para a seleção de genes diferencialmente expressos em dados de arranjos de DNA, baseado numa versão modificada do teste t de Student e no uso de permutações dos dados para estimar o FDR.

Para tornar mais claro o algoritmo SAM, suponhamos que os dados estão organizados numa matriz, na qual cada linha representa um gene presente nos arranjos e cada coluna os resultados de uma hibridação. Vamos supor também que as colunas estão divididas em dois grupos, cada um contendo dados provenientes de um estado diferente (tratamento, tecido, indivíduo etc.) que deseja-se comparar. A estatística aplicada pelo SAM a cada gene i é a "diferença relativa" $d(i)$, definida como:

$$d(i) = \frac{m_2(i) - m_1(i)}{s(i) + s_0},$$

em que $m_1(i)$ e $m_2(i)$ são respectivamente as médias entre os sinais detectados (usualmente em escala logarítmica) para o gene i nos estados 1 e 2, enquanto $s(i)$ é o desvio padrão das medidas de expressão repetidas,

$$s(i) = \sqrt{a \left\{ \sum_{n_1} [x_{n_1}(i) - m_1(i)]^2 + \sum_{n_2} [x_{n_2}(i) - m_2(i)]^2 \right\}},$$

com n_1 e n_2 os números de medidas dos estados 1 e 2, respectivamente, e $a=(1/n_1 + 1/n_2)/(n_1+n_2-2)$. Vale observar que se $s_0=0$ e $n_1=n_2$, a diferença relativa se torna equivalente ao teste t. O valor s_0 é adicionado ao denominador de $d(i)$ para evitar que grandes valores de $d(i)$ surjam apenas devido a valores de $s(i)$ muito pequenos. Além disso, s_0 é escolhido de maneira a minimizar a variância dos valores de $d(i)$ ao longo dos genes.

Por fim, para selecionar os genes diferencialmente expressos o algoritmo SAM precisa fixar valores de corte para os $d(i)$ e estimar os respectivos valores de FDR. Para isso, são utilizadas permutações aleatórias dos dados (Dudoit *et al.*, 2000), ou seja, as colunas da matriz de dados são permutadas ao acaso, misturando os valores provenientes dos dois estados, e os valores de $d(i)$ são computados para os dados permutados. Isso é repetido um grande número de vezes e então, para cada valor de corte fixado para o $d(i)$, estima-se o número de genes erroneamente selecionados pela média do número de valores de $d(i)$ acima do valor de corte a cada conjunto de dados permutado. O SAM define então um q-valor para cada gene, que é o FDR esperado se tomarmos o valor de corte estritamente necessário para o gene ser selecionado como diferencialmente expresso.

Nosso grupo de pesquisa aplicou o SAM na análise de dados de arranjos de DNA de um experimento voltado para a identificação de genes de milho relacionados com a tolerância a toxidez do alumínio (Al^{3+}). Plantas de duas variedades de milho, uma sensível e outra tolerante ao alumínio, foram cultivadas em hidroponia e tratadas com quatro concentrações diferentes do íon Al^{3+} (0, 15, 50 e 75 μM). Amostras de RNA total de raiz foram extraídas e utilizadas na

hibridação de arranjos de DNA em membranas de *nylon* contendo genes de cana-de-açúcar selecionados ao acaso, valendo-se do fato de que a grande homologia existente entre os genes de milho e cana possibilita a sua hibridação heteróloga. A aplicação do algoritmo SAM aos dados gerados identificou diversos genes de milho induzidos ou reprimidos nas variedades analisadas (Cançado *et al.*, 2006, artigo submetido ao Journal of Experimental Botany, em anexo).

RUNNING TITLE: Genes modulated by Al in maize

TITLE: Gene expression profiling of aluminum stress in maize roots.

KEYWORDS: abiotic stress, heterologous hybridization, *Saccharum* sp, *Zea mays*

Geraldo M. A. Cançado^{1,2}, Fábio T. S. Nogueira^{1,2}, Rodrigo D. Drummond^{1,2},
Juliana Felix^{1,2}, George Bezerra³, Fernando Von Zuben³, Renato A. Jorge⁴,
Marcelo Menossi^{1,2*}.

¹*Functional genomics Laboratory, Center for Molecular Biology and Genetic Engineering, State University of Campinas, P.O.Box 6010, 13083-970 Campinas, SP, Brazil.*

²*Department of Genetics and Evolution, Biology Institute, State University of Campinas, Campinas, SP, Brazil.*

³*Bioinformatics and Bio-inspired computation Laboratory, Faculty of Electrical Engineering, P.O.Box 6101, 13083-852 Campinas, SP, Brazil.*

⁴*Chemistry Institute, State University of Campinas, P.O.Box 6154, 13083-970 Campinas, SP, Brazil.*

*Address for Correspondence:

Centro de Biologia Molecular e Engenharia Genética, State University of Campinas (Unicamp), CP 6010, 13083-970, Campinas, SP, Brazil

Telephone number: +55 (19) 35211143

Fax number: +55 (19) 35211089

E-mail menossi@unicamp.br

Abstract

To understand the mechanisms responsible for aluminum (Al) toxicity and tolerance in plants, a DNA macroarray approach was used to analyze Al-induced gene expression changes in the roots of two maize lines (Cat100-6, Al-tolerant, and S1587-17, Al-sensitive). Due to the high degree of conservation observed among plant species within the same family, we analyzed the expression profile of maize genes using 2,304 sugarcane (*Saccharum* sp.) expressed sequence tags (ESTs) from different libraries. We have identified 85 genes in Al stressed maize root tips with significantly altered expression patterns, demonstrating that cross-species hybridization is a suitable strategy for overall gene expression profiling. Among the ESTs up-regulated by Al, genes encoding for previously identified Al-induced proteins, such as phenyl ammonia-lyase, chitinase, Bowman-Birk proteinase inhibitor and wali7, were identified. Several novel genes up- and down-regulated by Al-stress were also identified in Cat100-6 and S1587-17, providing new light upon the understanding of the Al-toxicity and -tolerance in plants.

Introduction

Aluminum (Al) toxicity is one of the major limiting factors to plant growth and development in many acid soils, thus the elucidation of the genetic and physiological basis of Al-stress responses would be of great interest to agriculture around the world. Al toxicity problems are of enormous importance for maize production in developing countries located in tropical areas of Asia, Africa and Latin America. Most maize cultivars currently being used are susceptible to toxic levels of Al in the soil, and thus decreases in yield of up to 80% result from Al toxicity (Herrera-Estrella, 1999).

The most dramatic symptom of Al toxicity is the inhibition of root growth, which has become a widely accepted measure of Al stress in plants. Although Al toxicity primarily restricts root growth, given sufficient exposure, a myriad of different symptoms appear on both roots and shoots that are often mistaken for soil nutrient deficiencies. At the cellular level, Al has been shown to affect a large number of biochemical processes, including lipid peroxidation (Yamamoto *et al.*, 2001), inositol 1,4,5-triphosphate signaling transduction (Jones and Kochian, 1995), cytoplasmic calcium homeostasis (Zhang and Rengel, 1999), microtubules and actin organization in cell elongation (Blancaflor *et al.* 1998), and callose formation and deposition (Horst *et al.*, 1997).

Although Al is responsible for promoting serious metabolic dysfunctions, some plants have evolved Al tolerance mechanisms that enable them to grow in Al-toxic, acid soil environments (for a review, see Kochian *et al.*, 2004; Ma *et al.*, 2001). The source of this tolerance is a genetic buffering capacity that permits

plants to adapt to the most diverse terrestrial environments. Therefore, a comprehensive understanding of the molecular mechanisms underlying Al toxicity and tolerance in plants could provide important insights into the development of new varieties with improved Al-tolerance.

In fact, Al toxicity can promote profound changes in gene expression, altering the control of normal physiological processes. It has been repeatedly observed that Al induces transcription of several genes, including those significantly associated with pathogen-, wounding-, and oxidative stress-induced proteins (Cruz-Ortega and Ownby, 1993; Snowden and Gardner, 1993; Snowden *et al.*, 1995; Cruz-Ortega *et al.*, 1997; Hamel *et al.* 1998; Richards *et al.*, 1998; Ezaki *et al.*, 2000, Watt 2003, Xiao *et al.*, 2005). These studies have been restricted to wheat and *Arabidopsis*. In the case of maize, a crop with a wide range of Al tolerant germplasms, gene expression changes induced by Al stress are almost unknown, except for a gene encoding a glutathione S-transferase recently identified (Cançado *et al.*, 2005).

In recent years, genomic approaches have been successfully used to examine global gene expression changes. Although the information derived from such studies undoubtedly provide a powerful means for studying the molecular mechanisms involved in biotic and abiotic stresses, the availability of cDNA sequences necessary to produce the arrays for large-scale gene expression profiling remains as a major limitation to several plants species. The use of heterologous hybridization is a well known tool to provide genetic information. And recently it has been used in large scale gene expression profiling. cDNA arrays from *Pinus* have been used to assess gene expression in *Picea* and *Nicotiana*

(Van Zil *et al.* 2002) species. Renn *et al.* (2004) found that stringent statistical analysis can allow gene expression profiling using heterologous hybridization even in distantly related fish species.

Taking into account the high degree of conservation among gene sequences from closely related species such as maize (*Zea mays* L.) and sugarcane (*Saccharum* sp.) (Kellogg, 2001) and the availability of filter DNA arrays containing genes from the sugarcane EST project (Nogueira *et al.*, 2003) , we decided to test whether these DNA arrays could be used to identify expression profiles in maize. Although comparative data between maize and sugarcane has not been available until now, Vincentz *et al.* (2004) observed 70.5% of ESTs sequences match between sugarcane and *Arabidopsis*. The same authors concluded that the evolutionary proximity of rice and sugarcane means that a large number of homologous sequences are to be expected from these two species. As maize is closer to sugarcane than *Arabidopsis* or even rice (Draye *et al.* 2001), a large number of homologous sequences between maize and sugarcane might be expected. Therefore, if the heterologous hybridization between maize and sugarcane held true, more than forty thousands sugarcane genes identified by the SUCEST project would become a valuable resource to study gene expression in maize and other monocots species with agronomical importance, such as sorghum, rye, barley and wheat, and whose genome database is not currently available.

In the present work, we have proven the suitability of the cross-species hybridization between maize and sugarcane as a large-scale expression profile approach. We have identified novel genes up- and down-regulated by AI stress in

two Al-contrasting maize lines, giving new insights on the way maize plants respond to Al.

Material and Methods

Plant material, seedling growth and Al-treatment

Maize (*Zea mays* L.) inbred lines Cat 100-6 (Al-tolerant) and S1587-17 (Al-sensitive) (Moon *et al.*, 1997), were obtained from the germoplasm collection of Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, Campinas, Brazil. Plants were grown in the field and self-pollinated. The sugarcane (*Saccharum* sp.) genotypes growth conditions, and the SUCEST libraries production are described in Vettore *et al.* (2001).

Maize seeds were surface-sterilized with 70% (v/v) ethanol for 1 min, 0.5% (v/v) sodium hypochloride for 20 min, rinsed four times with sterile water and germinated at 28 °C between two layers of moist filter paper for 3 days. Seedlings were grown in nutrient solution at pH 4.2 continuously aerated as described by Jorge *et al.* (2001) and with 16 h of light at 80 to 100 $\mu\text{E m}^{-2}\text{s}^{-1}$ and temperatures of 26 °C during the day and 20 °C at night. The root growth inhibition (RGI) calculations were carried out according to Moon *et al.* (1997).

In macroarray experiments, 3 days-old seedlings from Cat100-6 and S1587-17 were exposed to 0 (control), 75 and 283 μM of $\text{AlK}_3(\text{SO}_4)_3$ for 24 h. Two additional concentrations of Al (15 and 520 μM) were used for the RNA-gel blot analyses. The DNA macroarray experiment was independently replicated three times while RNA-gel blot experiments were independently replicated at least two times.

Macroarrays and Probe Preparation

Twenty four 96-well plates containing EST plasmid clones were randomly sampled from the following sugarcane cDNA libraries: heat- and cold-treated and untreated *callus* (CL6), sugarcane plantlets infected with *H. rubrisubalbicans* (HR1) or *A.diazotrophicans* (AD1), and leaf row tissue (LR1; Vettore *et al.*, 2001). Nylon filters containing EST plasmids were prepared as described in Nogueira *et al.* (2003). Three sets of filters were used, each one containing 768 ESTs, totaling 2,304 ESTs. Each EST was spotted twice on the same nylon filter and twice on the same spot to reduce experimental variation (Nogueira *et al.*, 2003).

RNA isolation and cDNA synthesis

Total RNA was isolated from roots according to Logemann *et al.* (1987) with minor modifications. Thirty root tips under each treatment were excised (approximately 5 mm), frozen in liquid nitrogen and grounded in extraction buffer (8 M Guanidine-HCl, 50 mM Tris-HCl (pH 8.0), 20 mM EDTA (pH 8.0) and 2% [v/v] 2-mercaptoethanol). After extraction with 1 volume phenol:chloroform:isoamyl alcohol (25:24:1 v/v/v), the suspension was separated by centrifugation (5000 *g*, 15 min), the aqueous phase was recovered by ethanol precipitation and the pellet was resuspended in DEPC-treated water.

The cDNA probes were produced as described by Schummer *et al.* (1999) with slight modifications. Approximately 30 µg of total RNA was reverse transcribed with Superscript II (Invitrogen, USA) using an oligo-dT18V (3µM) primer, with 3,000 Ci mmol⁻¹[α-³³P]dCTP and unlabeled dATP, dGTP, and dTTP (1 mM each) for 20

min at 42°C. Unlabeled dCTP was then added to a final concentration of 1 mM and the reaction continued for another 40 min. The cDNA probes were purified by using ProbeQuant G-50 microcolumns according to the manufacturer's instructions (Amersham Biosciences, USA) and the radiolabeled probe intensity was normalized with the aid of a 1217 Rack Beta liquid scintillation counter (LKB Wallac, Finland).

Variation of the amount of spotted DNA was previously estimated by hybridizing the filters with an oligonucleotide probe that recognizes the sequence of the *Amp^r* gene of the pSPORT1 vector (Invitrogen, USA). This probe was synthesized with the primers 5'-GTGGTCCTGCAACTTTATCCGC-3' and 5'-TAGACTGGATGGAGCGGATAA-3' in the presence of [α -³³P]dCTP, according to the protocol described by McPherson (2000; <http://www.tree.caltech.edu/protocols/overgo.html>).

Macroarrays Analysis

Filters were initially hybridized with the oligo *Amp^r* probe as described in Nogueira *et al.*, (2003). After hybridization and washing, the filters were exposed to imaging plates for 96 h and then scanned in a phosphorimager FLA3000-G (Fujifilm, Tokyo). Further, *Amp^r* probe was removed from the filters and then hybridized with cDNA probes and washed as described by Schummer *et al.* (1997). The filters were sealed in plastic film, immediately exposed to imaging plates for 96 h, and scanned as above. Signal was quantified using Array Vision software (Imaging Research, St. Catherines, ON, Canada). Grids were predefined and

manually adjusted to obtain optimal spot recognition, and spots were then quantified individually.

Filtered and normalized data was analyzed with SAM software (Tusher *et al.*, 2001). Treatment (75 μ M and 283 μ M of Al) comparisons against its control (absence of Al) were performed within each maize genotype, and genotype comparisons were performed within each Al concentration. For all comparisons, SAM parameters were set as follows: minimum fold-change was set to 1.5 at least for one comparison and the delta value was chosen as the minimum value that leads to an estimated median False Discovery Rate (estimate of the percentage of spots that will be falsely determined to be significant) of 1% or less.

RNA- gel blot analysis

Samples of total RNA (10 μ g) extracted from maize root tips were separated in 1% (w/v) formaldehyde-agarose gels. RNA blotting and filter hybridization were performed according to Sambrook *et al.* (1989). After hybridization and washing, RNA filters were exposed to imaging plates for 24 h and then scanned in a phosphorimager. RNA filters were further hybridized with a 26S rRNA probe to confirm equal RNA loading.

Bioinformatics analysis

Comparative sequence analysis was performed with BLASTx and BLASTn algorithms (Altschul *et al.*, 1997) against GenBank database (<http://www.ncbi.nlm.nih.gov>). Matches were considered significant when *E* values

were below 10^{-5} and PAM120 similarity scores were above 80 (Newman *et al.*, 1994). Data clustering of the gene expression profiles was performed using the algorithm HaiNet (Hierarchial Artificial Immune NETwork, Bezerra *et al.*, 2005), which provides an automatic detection of the optimized number of clusters and a hierarchical structure of the correlation among clusters. The average expression from the macroarray replicates was used for clustering the 85 genes previously selected by the SAM analysis. Before the clustering with HaiNet the expression intensity from each treatment was normalized by the mean obtained from all treatments to set the sensitivity of expression into the same scale.

Results

Effect of Al on roots of Cat100-6 and S1587-17

The percentage of root growth inhibition (RGI) for Cat100-6 (Al-tolerant) and S1587-17 (Al-sensitive) in complete nutrient solutions containing different concentrations of Al is shown in Fig. 1. The two genotypes showed inhibition of the root growth under presence of Al after 24 h. The most evident effect of Al was observed in the highest doses of 283 and 520 μM of Al. Although Al affected the root growth of Cat100-6 (reductions of almost 20 and 25% in the root growth in 283 and 520 μM of Al, respectively), S1587-17 showed a much more pronounced Al toxic effect. To the same Al doses, S1587-17 suffered reductions higher than 40 and 55%, respectively (Fig. 1).

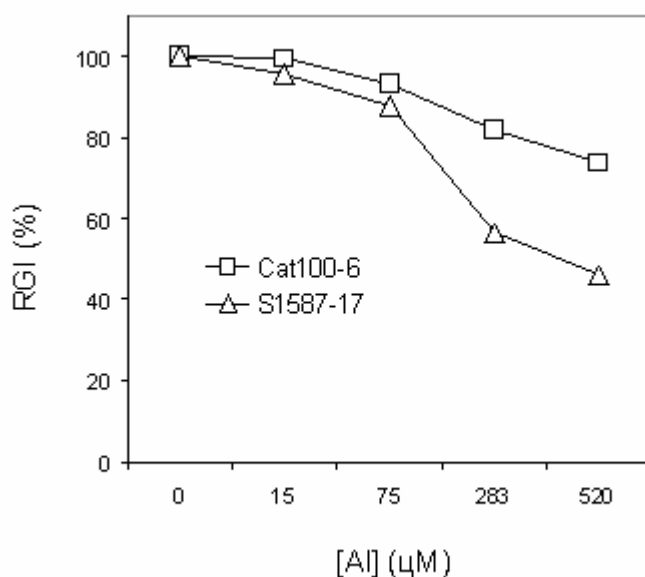


Figure 1: Dose–response curves for root growth inhibition (RGI) of Cat100-6 (square) and S1587-17 (triangle) after 24 of exposition in nutrient solution containing 0, 25, 75, 283, and 520 μM of Al. The data are the mean $\pm$ SE obtained from 15 replicates.

Identification Al-responsive genes using cross-species hybridization of macroarrays

Comparative maps among different grass species demonstrate that the information from one species can be extrapolated onto other ones, for breeding, ecology, evolution and molecular biology purposes (Guimarães et al., 1997). The close relationship between maize and sugarcane (Bennetzen and Freeling, 1997; Draye et al., 2001) prompted us to test whether gene expression in maize roots under Al stress could be outlined using macroarrays containing sugarcane ESTs. Macroarray data were analyzed by the SAM software (Tusher et al., 2001) with parameters chosen in order to lead to conservative selections of differentially expressed genes (see Materials and Methods). Each nylon array containing sugarcane ESTS was tested with different sets of cDNA probes synthesized from maize total RNA from three independent experiments to verify reproducibility. The normalized signal intensity of each spot was determined after subtracting the local background intensity. Thus 85 high quality sugarcane ESTs showing at least a 1.5-fold induction or repression were effectively selected, as shown in Table 1. As expected, the absolute-fold induction values were not identical among the biological samples, but the expression profiles were very similar, confirming the reproducibility of our macroarray data (data not shown). The worth of our macroarray experiment for screening Al-altered genes was demonstrated by the identification of several Al-inducible genes that had already been reported in other plant species and even in maize.

Table I. Average expression ratios and sequence similarities of Al-responsive sugarcane ESTs

Accession .	Class ^c	Blast Hit (E-value)	Description ^d	Ratios ^a				Cluster ^e	Ratios ^b	
				C75	C283	S75	S283		C/S 75	C/S 283
<i>Induced in both Cat100-6 and S1587-17</i>										
CA095811	III	JC5843 (2E-46)	Chitinase III	1.3	5.0	1.3	1.7	4	2.6	7.8
CA101236	VI	NP_922793 (2E-66)	Unknown protein	0.4	1.9	1.5	2.3	4	0.3	3.0
CA102633	VI	XP_469468 (E-129)	Unknown protein	0.7	1.9	3.4	2.5	4	0.9	3.0
CA098848	III	BAB63915 (0)	ERD protein	3.0	1.7	0.2	2.1	6	4.4	1.2
CA095754	III	1OM0A (E-119)	Xylanase inhibitor protein	1.5	1.7	3.1	1.4	7	1.9	4.5
CA102690	II	CAA27681 (0)	Alcohol dehydrogenase 1	0.7	2.3	1.5	1.3	12	0.3	1.0
CA096776	I	AAC69625 (E-115)	WD-40 protein	1.8	1.3	1.5	1.4	13	1.6	1.2
CA102731	II	AAF85966 (E-104)	Sucrose synthase	3.1	3.2	1.9	1.8	13	1.1	0.9
CA064600	III	BAA97804 (4E-46)	β-Glucuronidase precursor	1.0	1.9	0.7	1.7	16	1.7	1.6
CA064719	III	BAB19963	ASR protein	4.9	6.6	1.7	2.3	16	1.4	2.2
CA095602	VI		No hit	1.4	2.2	1.3	1.5	18	1.2	2.1
CA095678	III	P81713 (2E-07)	Bowman-birk proteinase inhibitor	1.5	1.3	18.7	46.6	20	1.4	1.1
CA097100	VI		No hit	3.2	1.1	2.4	65.6	20	2.5	0.7
CA095919	I	P49625 (E-116)	60S Ribosomal protein L5	1.9	1.2	2.2	2.0	22	1.4	1.0
CA097212	II	2008300A (0)	Sucrose synthase	1.7	1.4	3.2	3.0	22	1.7	1.1
<i>Induced only in Cat100-6</i>										
CA102689	VI		No hit	0.7	1.7	0.9	0.5	1	0.3	1.3
CA120042	III	BAD14927 (0.0)	DAHP synthetase	2.9	4.9	1.1	1.1	1	0.3	1.0
CA064763	III	AAL40137 (0)	Phenylalanine ammonia-lyase	1.3	3.7	1.1	1.1	4	1.0	3.0
CA064602		CAA13177 (7E-62)	Cinnamyl alcohol dehydrogenase	3.0	1.6	0.7	0.8	5	3.4	1.9
CA097041	VI		No hit	1.9	1.1	0.5	0.6	5	3.3	2.0
CA064780	VI	NP_683323 (E-14)	Unknown protein	1.0	1.5	0.4	0.5	6	2.3	1.8
CA064608	II	CAB87248 (9E-33)	Glycerol 3-phosphate permease	2.3	2.7	0.8	1.4	7	5.9	3.8
CA064605	IV	AAB97114 (2E-89)	Small GTP-binding protein	2.3	1.7	0.9	0.3	10	1.2	3.0
CA064810	VI		No hit	1.7	1.5	0.8	0.8	10	1.4	1.3
CA095938	VI		No hit	1.6	1.2	0.7	0.6	10	1.9	1.5
CA096090	III	AAC49972 (2E-28)	Hypersensitive response protein	1.8	2.0	1.4	1.0	10	1.0	1.6
CA096097	III	NP_077728 (9E-67)	Putative ferredoxin reductase	1.8	1.7	1.0	0.8	10	1.3	1.5
CA064787	III	AAK71314 (E-102)	Papain-like cysteine peptidase	1.6	1.3	1.4	0.6	11	1.5	3.6
CA095663	I	P05621 (E-38)	Histone H2B.2	1.5	1.1	1.2	1.3	11	0.7	0.5
CA096797	VI	XP_467727 (7E-48)	Unknown protein	0.8	2.2	0.8	0.4	12	0.7	4.1
CA097462	I	NP_958815 (2E-14)	LUC7-Like protein	0.6	2.0	0.4	0.3	12	1.7	5.2
CA101237	III	Q40977 (4E-26)	Monodehydroascorbate reductase	1.0	2.3	1.3	1.0	12	0.5	2.0
CA097457	V	S72526 (0)	Vacuolar H ⁺ -pyrophosphatase	1.9	1.5	1.2	1.4	13	1.3	0.8
CA064742	III	AAM75139 (5E-39)	Alkaline α-galactosidase	8.2	5.0	0.5	1.0	14	6.0	2.9
CA097200	VI		No hit	1.6	1.3	0.5	0.8	15	2.3	2.0
CA102716	III	AAC37416 (7E-50)	Wali7	1.1	1.6	1.2	1.0	15	1.1	2.1
CA096094	I	AAF76167 (E-174)	Nuclear cap-binding protein	1.2	1.9	1.0	0.7	18	0.9	2.0
CA097210	V	S52030 (E-24)	Oleosin 17	0.8	1.5	0.6	0.7	18	1.2	1.7
CA095623	III	CAD27730 (5E-41)	Xylanase inhibitor	0.9	1.8	1.0	1.3	19	0.7	1.4
CA095852	I	CAA58669 (2E-28)	Ribosomal protein S27	1.4	1.7	0.9	1.0	19	0.9	1.1
CA096135	I	AAG60059 (E-173)	Acetyltransferase-related protein	0.9	1.8	1.1	0.9	19	0.7	1.6

Table continues

Accession .	Class ^c	Blast Hit (E-value)	Description ^d	Ratios ^a				Cluster ^e	Ratios ^b	
				C75	C283	S75	S283		C/S 75	C/S 28
<i>Induced only in S1587-17</i>										
CA064736	V	AAB72111 (6E-57)	BP-80 vacuolar sorting receptor	0.9	0.9	3.4	0.8	2	0.1	0.7
CA102674	I	AAN15557 (4E-52)	ABI3-interacting protein 2	0.9	1.3	2.4	0.6	2	0.3	1.4
CA102687	VI	XP_475538 (8E-38)	Unknown protein	0.5	1.3	4.0	0.7	2	0.3	1.9
CA102772	II	CAA54609 (7E-75)	UTP-glucose glucosyltransferase	1.1	0.6	7.6	2.0	2	0.3	1.1
CA120045	I	NP_919951 (E-115)	Putative retroelement	0.4	0.5	9.1	0.8	2	0.0	0.4
CA101178	VI	AAF01557 (2E-28)	Unknown protein	1.5	1.1	1.8	0.9	3	0.3	0.6
CA097428	III	AAM64219 (2E-07)	Cadmium induced protein Cdl19	0.9	0.5	2.1	1.0	8	0.7	0.8
CA097438	III	CAA31077 (8E-36)	ABA-inducible protein	0.9	0.4	1.7	1.2	8	0.5	0.4
CA101174	VI		No hit	0.6	0.5	2.9	1.6	8	0.6	1.0
CA095885	I	T52344 (2E-87)	OsNAC5 protein	0.3	0.6	1.3	2.8	9	0.7	0.7
CA097119	VI	CAE03862 (3E-91)	Unknown protein	1.0	1.1	1.8	1.3	11	0.6	0.9
CA097111	III	JC5845 (2E-84)	Chitinase III	1.1	1.0	1.5	1.0	14	5.5	5.4
CA064657	V	CAA45024 (5E-94)	Aspartate aminotransferase	0.9	0.8	4.5	2.9	17	1.3	1.5
CA064710	II	AAF70821 (0)	β-Galactosidase	0.3	0.3	3.8	1.4	17	0.7	0.9
CA096803	IV	XP_469542 (3E-99)	Putative ATP-binding protein	0.7	0.6	0.9	2.5	17	1.4	1.0
CA097066	VI	CAE05958 (E-129)	Unknown protein	1.2	1.1	0.9	2.1	17	2.2	1.1
CA097909	VI	XP_476424 (E-22)	Unknown protein	0.2	0.4	40.2	15.0	17	0.2	0.8
CA101149	I	T01996 (6E-45)	Nucleoid DNA-binding protein	0.8	0.4	4.0	5.0	17	2.0	0.5
CA064656	V	P23687 (E-156)	Prolyl oligopeptidase	1.2	1.2	1.8	2.3	20	1.3	0.8
CA096567	II	CAA39454 (0)	Enolase	1.3	0.6	1.6	2.7	20	1.4	0.5
CA096578	I	O15818 (E-105)	Putative eIF-3	0.9	0.6	1.3	1.6	20	1.0	0.6
CA102758	V	O13656 (6E-24)	Mitochondrial import receptor	1.0	0.8	1.2	2.6	20	1.1	0.4
CA119990	III	JC2510 (0)	β-Tubulin	1.1	0.9	1.6	2.1	20	1.3	0.9
CA064621	III	CAA55893 (E-140)	Putative imbibition protein	1.0	1.3	1.1	1.9	21	1.8	1.3
CA064717	VI		No hit	1.2	1.0	1.1	2.1	21	2.9	1.5
CA097411	I	AAC25599 (3E-10)	CRP1 protein	1.2	0.9	0.9	2.3	21	2.2	0.8
CA101244	III	NP_005856 (E-43)	Putative serine peptidase	1.0	0.7	0.8	1.9	21	3.9	1.3
CA119993	IV	CAA10660 (E-151)	Ca ²⁺ -ATPase	1.1	1.1	1.8	2.2	22	1.6	1.5
<i>Repressed in both Cat100-6 and S1587-17</i>										
CA095925	II	AAG28503 (E-151)	Hexokinase	0.6	0.2	0.3	0.8	9	1.2	0.4
<i>Repressed only in Cat100-6</i>										
CA095622	VI		No hit	0.4	0.5	0.8	0.8	9	0.8	0.9
CA095948	VI	AAN41388 (2E-17)	Unknown protein	0.6	0.4	0.9	1.1	9	1.7	0.7
CA097085	VI		No hit	0.7	0.4	0.7	0.6	9	1.4	0.8
CA101148	I	T01996 (6E-45)	Nucleoid DNA-binding protein	0.8	0.4	1.3	1.5	9	1.6	0.8
CA095886	I	CAA63194 (E-155)	Ribonucleotide reductase R2	0.5	0.8	0.9	1.3	11	0.5	0.6
CA096779	I	P12629 (7E-60)	50S ribosomal protein L13	0.3	0.5	1.2	0.8	11	0.2	0.5
CA099555	I	CAB75508 (3E-14)	ABI3-interacting protein	0.5	0.6	1.5	1.1	17	0.7	1.2

Table continues

Accession .	Class ^c	Blast Hit (E-value)	Description ^d	Ratios ^a				Cluster ^e	Ratios ^b	
				C75	C283	S75	S283		C/S 75	C/S 283
<i>Repressed only in S1587-17</i>										
CA101271	III	AAM23263 (E-30)	DnaJ-like protein	1.0	0.6	1.1	0.3	3	0.4	0.9
CA095710	I	BAA95894 (2E-72)	Putative reverse transcriptase	1.0	0.7	0.5	0.2	5	4.0	6.5
CA096690	VI	BAB11002 (E-138)	Unknown protein	0.9	0.7	0.3	0.3	5	4.2	3.6
CA095910	V	AAF74980 (E-121)	Cystathionine b-lyase	1.1	0.9	0.3	0.7	6	2.6	1.0
CA097854	I	BAA31739 (2E-05)	COP1-Interacting protein 7	0.9	0.7	0.4	0.8	6	2.8	1.1
CA098865	I	AAT40500 (E-50)	Putative reverse transcriptase	0.9	1.2	0.5	0.8	6	1.7	1.3
CA064616	I	AAD45720 (2E-31)	Zinc finger protein	1.3	1.0	0.6	0.3	14	10.6	16.3
CA064644	V	AAM60860 (E-136)	Thiol-monophosphate biosynthesis	0.6	1.1	0.4	1.0	21	3.1	1.9

^aExpression ratios of Al-treated intervals (75 and 283 μ M of Al) in relation to its control (0 μ M of Al): C75 = Cat100-6 under 75 μ M of Al; C283 = Cat100-6 under 283 μ M of Al; S75 = S1587-17 under 75 μ M of Al; S283 = S1587-17 under 283 μ M of Al. ^bExpression ratios of Cat100-6 in relation to S1587-17: C/S 75 = Cat100-6/S1587-17 under 75 μ M of Al; C/S 283 = Cat100-6/S1587-17 under 283 μ M of Al. ^cClasses dividing the genes in six general categories: I) Gene regulation; II) Sugar metabolism; III) Plant stress response; IV) Signal transduction; V) Other functions; and VI) Unknown and no hit proteins. ^dDescription indicates the putative functions of the gene products expected from similarity sequences. ^eClustering profile shown in Fig. 4.

The relevant biological functions of identified Al-responsive genes are shown in Table 1. The ESTs encoding known proteins represent a wide range of functions, including transcription, signalling, sugar metabolism, defense and development. These results suggest that several metabolic processes, including perception of stress signals and regulation of gene expression, were altered during-Al stress in both Al-tolerant and Al-sensitive plants.

Among the 85 Al-responsive genes detected in our array data, 41 (48.2%) and 43 (50.5%) were up-regulated Cat100-6 and S1587-17, respectively, noting that 15 (17.6%) were up-regulated in both genotypes (Table 2). This represents 81.2% of the ESTs identified in this work whose expression was altered by Al in both genotypes. The number of ESTs down-regulated in Cat100-6 and S1587-17 were 8 (9.4%) and 9 (10.6%), respectively, while only one was repressed in both representing 18.8% of all ESTs significantly altered by Al (Table 2). The up-regulated sugarcane ESTs were distributed in three groups (Table 2). The first

group contained ESTs encoding proteins previously described as Al-responsive, such as phenylalanine ammonia-lyase, wali7, Bowman-Birk proteinase inhibitor, and β -tubulin. The second group contained ESTs encoding proteins responsive to other biotic and abiotic stresses, but not reported as Al-responsive until now; this group included xylanase inhibitor protein, OsNAC5 protein, cadmium induced protein Cdl19, and ribosomal protein S27. The third group contained ESTs encoding unknown and proteins with no hits in the databases, and proteins not reported as stress-responsive, such as ASR protein, nucleoid DNA-binding protein cnd41, mitochondrial import receptor tom40, BP-80 vacuolar sorting receptor, glycerol 3-phosphate permease, and histone H2B.2.

To estimate the relative contribution of Al-altered genes from each SUCEST library used in the array experiments, we calculated the number of Al-altered ESTs from each cDNA library, as shown in the Fig. 2. Interestingly, the library CL6 accounted for most of the Al-induced ESTs and almost all of the Al-repressed ESTs identified in Cat100-6 and S1587-17. The CL6 library was constructed from a mixture of heat- and cold- treated and untreated sugarcane *callus* tissues (Vettore *et al.*, 2001) and was therefore a good source of biotic and abiotic stress-related ESTs (Vettore *et al.*, 2003).

Table 2. *Aluminum-regulated sugarcane ESTs identified by macroarray expression profiling*

Summary	Cat100-6		S1587-17		TOTAL	
	N ^o . of ESTs	% ESTs	N ^o . of ESTs	% ESTs	N ^o . of ESTs	% ESTs
Al-responsive ESTs						
Up-regulated ESTs	41	48.2	43	50.6	69	81.3
(Known Al-responsive ^a)	(5)	(5.9)	(6)	(7.0)	(8)	(9.4)
(Other stress-responsive ^b)	(17)	(20)	(14)	(16.5)	(25)	(29.4)
(Novel Al-responsive ^c)	(19)	(22.3)	(23)	(27.1)	(36)	(42.3)
Down-regulated ESTs	8	9.4	9	10.6	16	18.8
Unknown ESTs						
Al up-regulated	4	4.6	7	8.1	9	10.4
Al down-regulated	1	1.2	1	1.2	2	2.3
No hit ESTs						
Al up-regulated	7	8.1	4	4.6	9	10.4
Al down-regulated	2	2.4	0	0	2	2.4

^aESTs homologous to previously described Al-induced genes. ^bESTs homologous to previously described general-stress responsive genes. ^cESTs encoding proteins that have not been described previously as Al-induced or other stress-induced.

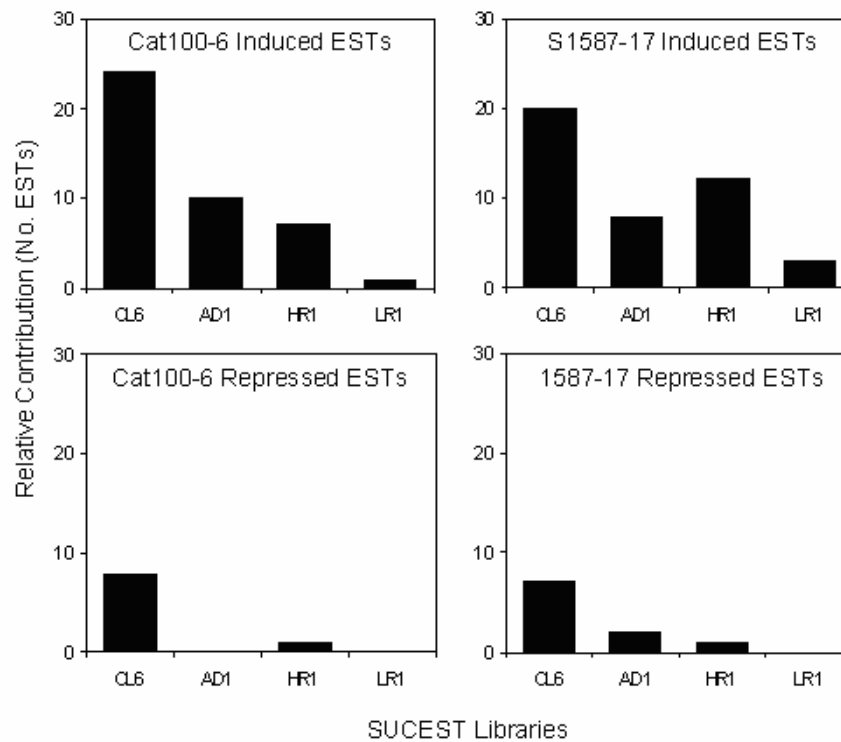


Figure 2: Relative contribution of the SUCEST libraries used in the macroarray experiments to identify AI-inducible genes. The values represent the percentage of normalized AI-altered genes identified from each cDNA library (see “Results”) in relation to the total number of ESTs identified as AI-altered in our arrays. CL6, Heat and cold-treated and untreated *callus*; AD1, sugarcane plantlets infected with *Acetobacter diazotrophicans*; HR1, plantlets infected with *Herbasperillum rubrisubalbicans*; and LR1, leaf row tissue.

RNA-gel blotting Analysis

RNA-gel blotting analysis using total RNA from at least two independent biological replicates were performed for specific genes in order to validate the macroarray data. Fig. 3A shows the expression profile of six genes with high similarity to a phenylalanine ammonia-lyase (accession CA064763), an abscisic acid- and stress-inducible protein (CA064719), a protein homologous to B12D domain, described as a pathogen inducible protein whose function remains unknown, (CA097100), a chitinase class III (CA095811), a β -glucuronidase (sGUS) precursor (CA064600) that is involved in the initiation of H_2O_2 metabolism, and a hypothetical protein with no conserved domain (CA101236). These sugarcane clones respectively showed 85.8% (accession CO520609), 81.2% (BM499107), 41.6% (BE510590), and 52.0% (CF024252) identity with maize EST sequences from the Maize Genetics and Genomics Database (MaizeGDB, <http://www.maizegdb.org/blast.php>) except CA101236 that produced no significant hit. Fig. 3B compares the gene expression profiles obtained from macroarrays and RNA-gel blotting. The absolute-fold induction or repression values from the RNA blots were close to those on the DNA-array, indicating a high consistency between the two data sets, although the relative intensity scale was different between the methods.

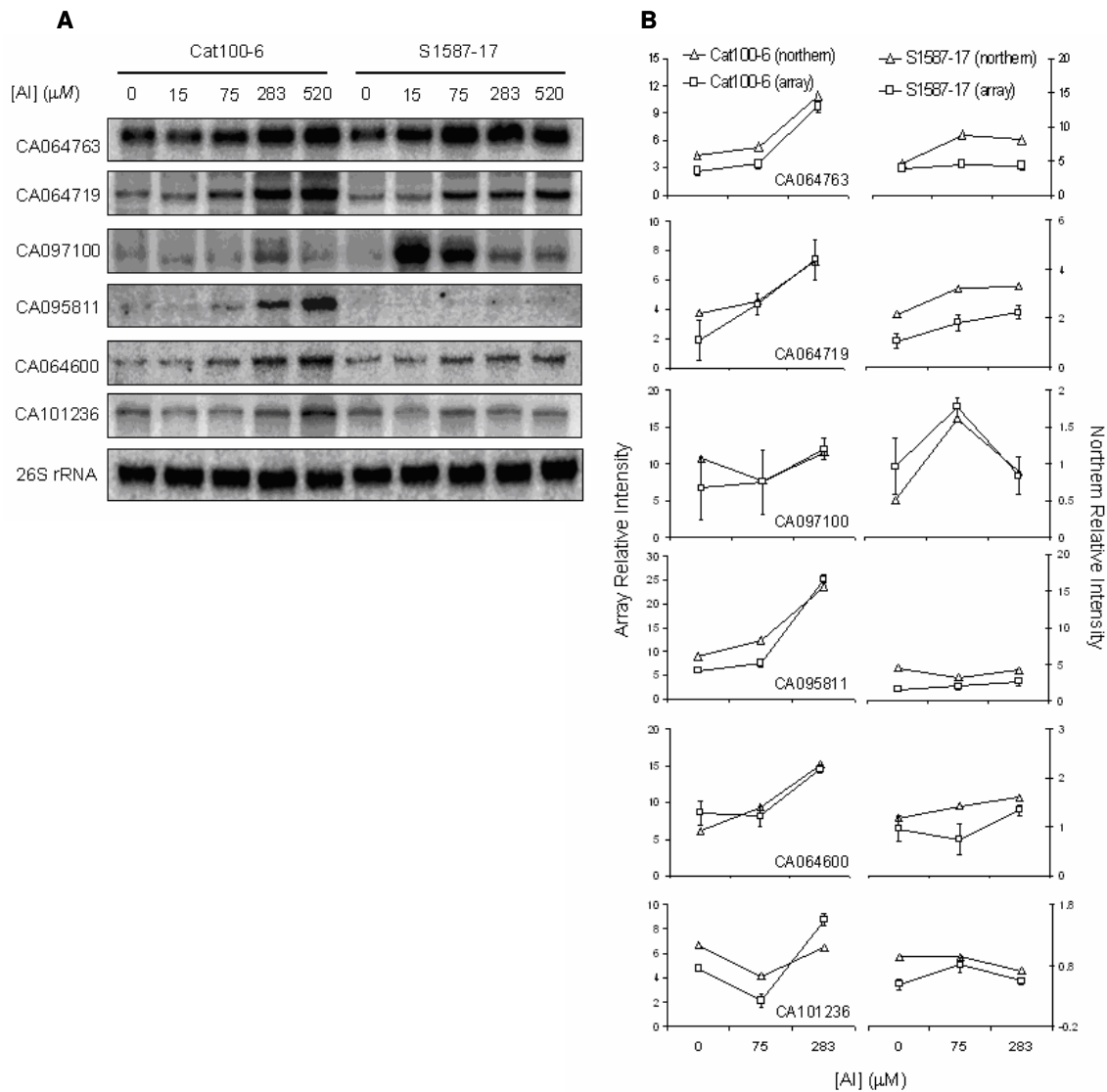


Figure 3: Comparison between EST macroarray and RNA-blot analysis for Al-altered maize cDNAs. (A) In RNA gel blots, each lane was loaded with 10 μ g of total RNA isolated from root tips exposed to increasing Al levels. The RNA loading was monitored using a 26S rRNA as probe. (B) The graphs show the induction kinetics observed in the macroarrays (square) and RNA blots (triangle).

Data Clustering of AI-Responsive Genes

Clustering of the apparently arbitrary distribution of gene expression patterns might reveal groups of genes with strong correlations in their expression profiles, therefore exposing a more comprehensible structure that may provide meaningful information about the function of unknown genes.

The algorithm HaiNet was used to cluster the 85 genes previously selected by the SAM analysis. HaiNet automatically determined 22 clusters of genes based on their density distribution. The cluster where each gene was positioned is indicated on the ninth column of Table 1. The algorithm also produced a hierarchical topology of the similarity relation among clusters based on the expression profile of each cluster (Fig. 4). The more stable clusters were identified in the earliest level of the hierarchy and were positioned close to the base of the tree. Clusters 1, 2, and 3 are examples of this behavior. Clusters 4, 5, 7, and 8 efficiently grouped genes involved in closely-related functions, such as plant defense. It was found that many of the genes with unknown function or genes with no hit in the GenBank database were uniformly distributed among the 22 clusters, indicating no tendency of specific grouping. On the other hand, some of these unknown and no hit genes were located in clusters with genes involved in similar pathways or metabolic routes, thus suggesting the involvement of them in known biochemical processes. This was the case for CA102633 and CA101236 cDNAs of unknown proteins that were located in the same cluster of chitinase III and phenylalanine ammonia-lyase (PAL), enzymes involved in plant defense mechanisms.

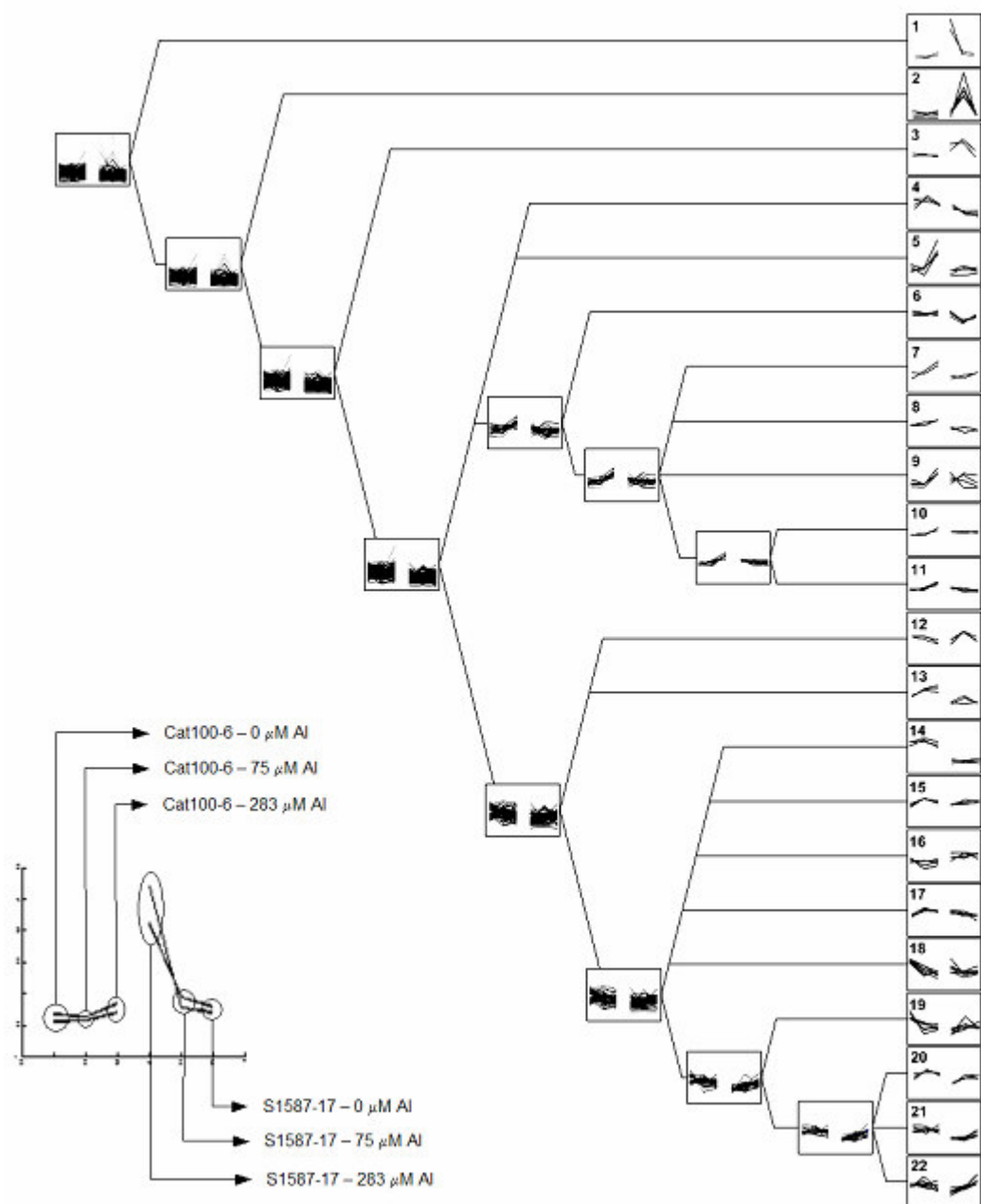


Figure 4: Hierarchical clustering of 85 ESTs selected in the macroarray analysis.

This clustering of data based in the absolute levels of transcripts in both maize lines also gave insights on which genes that were induced in Cat100-6 might be related to the higher tolerance observed in this maize line (Table 1 and Fig. 4). This is because a higher induction of a gene in the tolerant line in response to AI is not a good parameter to explain AI tolerance, since the unchanged levels of the transcripts from the same gene in the AI-sensitive line may be higher. From those genes identified in the macroarrays experiments as induced by AI in both maize lines, five had at least three times more transcripts in the AI-tolerant line than in the sensitive one in the same AI treatment: CA095811 (encoding a chitinase III), CA095754 (xylanase inhibitor protein), CA098848 (ERD protein) and CA102633 and CA101236 (both unknown proteins). The most promising genes among those induced only in Cat100-6 were CA064742 (encoding an alkaline α -galactosidase), CA064608 (glycerol 3-phosphate permease), CA097462 (LUC7-Like protein), CA096797 (unknown protein), CA064787 (papain-like cysteine peptidase), CA064602 (cinnamyl alcohol dehydrogenase), CA097041 (no hit), CA064605 (Small GTP-binding protein) and CA064763 (phenylalanine ammonia-lyase).

Discussion

Aluminum-induced root growth inhibition in maize

As a result of Al infiltration into the root apex, the primary effect observed when plants are growing in phytotoxic levels is the inhibition of root growth (Foy *et al.* 1978; Ryan *et al.* 1993). The growth inhibition observed in roots of both Cat100-6 and S1587-17 was dependent of Al concentration. Although the toxic effect of Al was almost twice as severe to the root growth of S1587-17 than to Cat100-6 (Fig. 1). This result shows that high levels of Al might surpass the tolerance observed even for Al-tolerant genotypes. The data on root growth under Al-stress obtained in this work agrees with results of previous works developed with the same maize genotypes in similar conditions (Boscolo *et al.*, 2003; Jorge *et al.*, 2001; Moon *et al.*, 1997). Twenty-four hours of Al-stress is shown to be enough time to promote visible alterations in the root phenotype of both maize lines, mainly in the highest concentrations of 283 and 520 μM of Al (Fig. 1).

Cross-species hybridization of macroarrays

Maize and sugarcane are grasses that belong to the same family Poaceae and to the same tribe Andropogoneae (Clayton & Renvoize, 1986) and recent comparative DNA sequence information confirms that all grasses examined to date diverge from one common ancestral population of “grass alleles” (Freeling 2001). Comparative mapping within maize, sorghum and sugarcane has previously revealed the existence of high synteny (regions with homologous chromosomal

segments) among these crops (Dufour *et al.*, 1997). Comparative maps among different grass species demonstrate that the information from one species can be extrapolated onto other ones, for breeding, ecology, evolution and molecular biology purposes (Guimarães *et al.*, 1997).

Using large-scale gene expression profiling in which the tester (cDNA obtained from maize used as probe) and the target (cDNA from sugarcane library spotted on the nylon filters) were originated from different but related plant species, we identified 85 genes with expression significantly altered by AI (Table 1). The confirmation of the expression profiling using RNA-gel blots and the high quality signal produced by the heterologous hybridization of the sugarcane filters with maize probes (data not shown) for most of the 2,304 clones spotted on the filters are an indication that there is enough conservation of the nucleotide sequence between these two species to utilize comparative techniques. Thus, we have demonstrated that sugarcane EST macroarrays are a useful tool to analyse gene expression profiles in related grasses species, such as maize.

This result is in agreement with those from Van Zil *et al.* (2002), that also used a heterologous approach to study the expression profile of 384 cDNAs from a conifer of genus *Pinus*, observed high correlations of gene expression when using labelled cDNA from another conifer of genus *Picea* and even from *Nicotiana tabacum*, at a reduced scale. Moreover, the results obtained from Renn *et al.* (2004) working with different fish species suggest that with sound statistical analysis and additional replicates even subtly regulated genes can be identified in the distantly related species by a cDNA macroarray heterologous hybridization. These authors also show that expression-profiling results have consistent

hybridization not only for closely related taxonomic groups (less than 10 million years divergence) but also for other very distantly related taxonomic groups (more than 200 million years divergence). Therefore, the results obtained in this work using a comparative approach permit us to conclude that biological information from one monocotyledonous species might be effectively utilized to study and improve other related monocotyledonous species.

Effect of Al-stress in the expression profiling of maize root tips

Because complex traits are polygenic and gene interactions are fundamental properties of these traits, taking advantage of a genomic approach to the study of gene expression improves the power to reveal the complex network of interrelated functional modules that are involved (Aubin-Horth *et al.*, 2005). Although the functional categorization of genes is not a simple task, since many genes seem to be involved in more than one biochemical process, we have tried to assembly the genes identified in this work into 6 main groups based on their putative function (Table 1). The first group is comprised of genes involved with transcriptional and post-transcriptional regulation. Genes encoding proteins and enzymes that participate in sugar metabolism and other sugar-dependent processes constitute the second group. The third group contains genes previously cited as stress-responsive, including genes characterized as regulated by Al-stress. The fourth group includes genes involved with signal transduction pathways. The fifth group is composed of proteins whose known function does not match with any of the groups described above and the last group is formed of unknown proteins and genes with no significant hit in open access genetic

sequence database. Certainly, the central purpose of this classification was not to produce an exact categorization of all genes identified in this work, since many genes can be involved with more than one biological process, but to coordinate the discussion about these genes into groups with related function.

Several genes involved with the control of gene expression were up- or down-regulated by AI stress in both maize lines. The gene encoding the ribosomal protein S27 (accession CA095852) was up-regulated only in Cat100-6 plants growing in presence of AI. Interestingly, this protein displays DNA binding properties and is involved in response to DNA injury and degradation of damaged RNAs (Revenkova *et al.*, 1999). Another interesting observation was the up-regulation of a gene encoding an acetyltransferase (CA096135) in Cat100-6. This enzyme is responsible for acetylation of nucleosomal histones and, in general, there is a positive correlation between the histone acetylation state and transcriptional activity of genes (Bordoli *et al.*, 2001). Another gene involved in genetic regulation, up-regulated in Cat100-6 by AI, encoded a H2B histone (CA095663). H2B is a target for posttranslational modification of chromatin topology, altering the regulation of gene expression (Henry *et al.*, 2003). This finding suggests that primary or secondary effects promoted by AI stress in root cells might be responsible for profound and complex alterations in mechanisms of genetic expression.

Fluctuations of other transcripts involved in gene regulation reinforce the conclusion that AI promotes important imbalances in the intricate network involved with gene expression. The genes encoding the proteins CBP80 (Cap-Binding Protein 80Kda, CA096094) and LUC7 (lethal unless CBC, CA097462), both up-

regulated in Cat100-6, are involved in mRNA processing. While CBP proteins are involved in the formation of the cap-binding complex (CBC) (Kmieciak *et al.*, 2002), LUC7 is required for the complex formation of pre-mRNA splicing machinery (Fortes *et al.*, 1999). A gene that was up-regulated in S1587-17 (CA095885) belongs to the family of NAC (NAM, ATAF1, and CUC2) domain proteins which are members of putative transcript factors found exclusively in plants and involved in development and recognition of environmental stimuli such as cold stress (Nogueira *et al.*, 2005), pathogen infection and wounding (Collinge and Boller, 2001).

Other important proteins that participate in the control of gene expression can be observed in Table 1. The ASR (ABA [abscisic acid]-, stress-, and ripening-induced, CA064719) is involved in a common transduction pathway of sugar and ABA signaling (Çakir *et al.*, 2003). Additional AI responsive-genes encoding proteins involved in transcriptional regulation included AIB3 (ABA-insensitive, CA102674), which is known to be an activator of the expression of a large number of genes during embryo maturation (Kurup *et al.*, 2000) and WD-40 repeat proteins (CA096776), which are components of complexes involved in assembling and modification of chromatin (Kenzior and Folk, 1998).

The EST clone CA096797, up-regulated in Cat100-6 and down-regulated in S1587-17, encodes an unknown protein that has a bromo-adjacent homology (BAH) conserved domain containing a protein-protein interaction module specialized in gene silencing. Therefore, it might play important role in DNA methylation, replication and transcriptional regulation (Callebaut *et al.*, 1999).

The expression profiles of genes involved with sugar metabolism also were broadly affected by AI-stress. Although these genes are thought to have a main role in structural, storage, and metabolic functions (Ruan *et al.*, 1997), currently it is well known that many of these genes have important roles in signal transduction processes. In plants, sugars not only function as metabolic resources and structural constituents of cells, but also act as important regulators of various processes associated with plant growth and development (Rolland *et al.*, 2002). In this work we can cite for example those genes encoding enolase (CA096567), sucrose synthase (CA102731 and CA097212), hexokinase (CA095925), alcohol dehydrogenase (CA102690), glycerol 3-phosphate permease (G3PP, CA064608), and UDP-glycosyltransferase (CA102772). The last gene was strongly up-regulated by AI in S1587-17 and the enzyme encoded for it is not involved directly with energy metabolism but with the glycosylation of several molecules such as proteins and lipids, activating or inactivating these compounds (Ciereszko *et al.*, 2001). Probably, rapid alterations in the transcript level of this gene will promptly affect the cellular homeostasis.

Interestingly, some genes involved with sugar metabolism preferentially expressed in conditions of O₂ deprivation, such as enolase and alcohol dehydrogenase, were affected by the AI-stress. These enzymes are mainly expressed when plants switch its aerobic metabolism to a fermentative metabolism during situations of hypoxia or anoxia. The increased expression of these genes by AI-stress might be an indication that AI affects the O₂ availability into the cell. In fact, Sugimoto *et al.* (2004) suggests that the increase of enolase expression in an

Al-tolerant microorganism (*Penicillium chrysogenum*) might be involved with the sufficient supply of energy required by cells under stressing situation.

A gene encoding for the important enzyme 3-deoxy-D-arabino-heptulosonate-7-phosphate (DAHP, CA120042), which is involved with the synthesis of aromatic amino acids, was induced in Cat100-6 up to four times, while in S1587-17 it remained constant (Table 1). The DAHP synthase catalyzes the first committed step in aromatic amino acid biosynthesis which are precursors to a diverse array of plant secondary metabolites, including lignin, anthocyanic pigments, auxin, and antimicrobial phytoalexins (Keith *et al.*, 1991).

Monodehydroascorbate reductase (CA101237) was another enzyme whose gene was up-regulated by Al in the roots of Cat100-6. Monodehydroascorbate radicals are generated in plant cells enzymatically by the hydrogen peroxide scavenging enzyme, ascorbate peroxidase, and non-enzymatically via the univalent oxidation of ascorbate superoxide, hydroxyl, and various organic radicals. Regeneration of ascorbate, an important antioxidant (Murthy and Zilinskas, 1994), is achieved by monodehydroascorbate reductase (MDAR) using NAP(P)H as an electron donor. Moreover, Lukaszewski and Blevins (1996) reported that Al-stress resulted in a reduction of ascorbate concentration in root apices of squash (*Cucurbita pepo*), indicating that an increase on the transcript levels of MDAR in the root tips of Cat100-6 might be a response to the high levels of ascorbate oxidized by Al.

The Cdl19 (CA097428) protein functions as a metallochaperone localized at plasma membrane, which directly binds Cd and increases upon heavy metal stress (Suzuki *et al.*, 2002). Although its expression remained constant in Cat100-6, this

gene was up-regulated in S1587-17. Suzuki *et al.* (2002) showed that the overexpression of the Cdl19 cDNA conferred Cd tolerance in transgenic *Arabidopsis*, suggesting that this protein plays an important role in the maintenance of heavy metal homeostasis and/or in detoxification by endowing the plasma membrane with the capacity to serve as an initial barrier against inflow of free heavy metal ions into cells. Perhaps a similar mechanism is present in S1587-17 at the level of the plasma membrane in an attempt to alleviate Al-toxicity, although conclusive experiments need to be done in order to confirm this relation.

The gene encoding the β -tubulin protein (CA119990) was also up-regulated in S1587-17, while remaining unaltered in Cat100-6. This result might be explained by the fact that the root tip tissue was severely wounded in S1587-17, but not in Cat100-6 (not shown). Perhaps the cytoskeleton in Cat100-6 was less affected than in S1587-17. It was previously observed that microtubules play an important role in many cellular processes, such as cell division and elongation. β -tubulins, which are the basic components of microtubules, are encoded by a multigene family in eukaryotes and their nucleotide sequence are highly conserved in protein coding regions. The cytoskeleton in the root cells of maize is especially sensitive to Al (Sivaguru *et al.*, 2000; Sivaguru *et al.*, 1999, Elison *et al.*, 1998). MacDonald *et al.* (1987) reported that Al directly influenced tubulin polymerization *in vitro*, and Blancaflor *et al.* (1998) related Al-inducing growth inhibition in root apices of maize as a function of the stabilization of microtubules in the central elongation zone. According to Schwarzerová *et al.* (2002) the microtubular cytoskeleton is one of the early targets of Al cations.

In spite of the fact that the secretion of organic acid from roots has been cited as a probable mechanism involved with Al tolerance in Cat100-6 (Jorge and Arruda, 1997), we have not identified genes involved with biosynthesis of organic acids whose expression was significantly altered by Al stress. For instance, ESTs corresponding to genes encoding four enzymes from the TCA cycle (aconitase, isocitrate dehydrogenase, fumarase and malate dehydrogenase) were present in the arrays but their transcript levels were unchanged (data not shown). This result is in agreement with the observations from Xiao *et al.* (2005) that were not able to find Al-induced genes involved with organic acids biosynthesis in wheat. In fact, internal organic acid concentration has no clear correlation with Al tolerance or organic acids release in maize and other species (Mariano *et al.*, 2005).

Gene expression associated to Al tolerance

Based upon the identification of several genes whose expression was modulated by Al-stress and whose functions are so diverse, it is clear that Al stress induces a very complex biochemical response in maize roots, triggering processes involved in defense, signalling and metabolic pathways. The clustering of the expression levels in each Al treatment allowed us to identify another set of genes that might have a closer relation to the tolerance observed in Cat100-6. Genes that have the same expression profile and the same transcript level in both maize lines certainly cannot explain the differences in root growth we observed. Clusters 4, 7, 10, 11 and 14 grouped the genes with the most interesting expression patterns.

Among the genes identified in the macroarrays experiments as induced by Al in both maize lines, five had at least three times more transcripts in the Al-

tolerant line than in the sensitive one in the same Al treatment: CA095811 (encoding a chitinase III), CA095754 (xylanase inhibitor protein), CA098848 (ERD protein) and CA102633 and CA101236 (both unknown proteins).

The role of chitinases in plants is considered to be part of their defense mechanism against fungal pathogens that contain chitin as a structural component (Watanabe *et al.*, 1992). In addition, the increase in the transcription of chitinase genes or in chitinase activity may be induced by other external stimuli such as wounding, drought, cold, ozone, heavy metals, excessive salinity and UV light (revised by Kasprzewska, 2003). This enzyme was also recently appointed as involved in Al-stress alleviation in plants of Norway Spruce (*Picea abies*) (Nagy *et al.*, 2004). Chitinases also thought to play a role in growth and development by catalyzing the hydrolytic cleavage of arabinogalactan proteins (AGPs), thus affecting cell-to-cell communication (Wiweger *et al.*, 2003). An increase in the level of chitinase transcripts and consequently in the level of the enzymatic activity might be involved with the developmental response of Cat100-6 roots when exposed to Al.

Two xylanase inhibitor genes originated from different assembled sequences (CA095754 and CA095754) were up-regulated by Al in Cat100-6. Xylanases are enzymes that promote the degradation of xylan, the most abundant natural polysaccharide after cellulose (Fierens *et al.*, 2003). Therefore, inhibitors of xylanases play an important role in plant defense by preventing enzymatic hydrolysis by microorganisms and predators (Juge *et al.*, 2004). As observed in lignin metabolism, inhibitors of xylanase genes might contribute to maintaining cell wall strength as a response to Al-stress.

Abscisic acid (ABA) is an important plant hormone that helps plants cope with several environmental stimulus (Zeevaart, 1999). ABA was demonstrated to increase the expression level of several genes including genes from the group ERD (early-responsive to dehydration, CA098848). These genes share similarity with HSP (heat shock proteins) from a variety of organisms and they are early-responsive to dehydration stress (Kiyosue *et al.*, 1994). The higher expression of the gene encoding an ERD in the Al-tolerant line suggests that Al causes dehydration and that its mitigation can improve plant responses to Al stress.

The most interesting genes among those induced only in Cat100-6 were CA064742 (encoding an alkaline α -galactosidase), CA064608 (glycerol 3-phosphate permease), CA097462 (LUC7-Like protein), CA096797 (unknown protein), CA064787 (papain-like cysteine peptidase), CA064602 (cinnamyl alcohol dehydrogenase), CA097041 (no hit), CA064605 (Small GTP-binding protein) and CA064763 (phenylalanine ammonia-lyase).

The gene encoding an alkaline α -galactosidase might be involved in the signalization by the hormone methyl-jasmonate (MeJA). A homologue from rice has been implicated in leaf senescence and is supposed to act in the release of linolenic acid, which is the primary substrate for MeJA synthesis (Lee *et al.*, 2004). This enzyme is also involved in the initial metabolism of sugars, mainly from raffinose family of oligosaccharides (Carmi *et al.*, 2003). The balance and metabolism of these sugars have been associated with acquisition of cold and desiccation tolerance (Keller and Pharr, 1996). An efficient balance on the level of

these oligosaccharides would help to reduce the dehydration caused by Al stress in Cat100-6.

Glycerol 3-phosphate permease is a protein that remains uncharacterized in plants but that seems to be important in transport of glycerol 3-phosphate (G3P), a molecule that plays a major role in glycolysis and phospholipid biosynthesis (Lemieux *et al.*, 2004). Glycerol 3-phosphate permeases can provide osmotic protection in yeast due to uptake of glycerol (Holst *et al.*, 2000). Interestingly, the molecule G3P is also involved in the maintainment of membrane lipids integrity during chilling stress, that is also related to dehydration stress (Murata *et al.*, 1992; Wolter *et al.*, 1992). Together with the expression of other genes potentially related to dehydration, these data reinforce our view that Al induces an osmotic stress, which seems to be alleviated in a more efficient way in Cat100-6 roots.

The higher expression of LUC7 (involved in RNA splicing) and the papain-like cysteine peptidases (CA064787) suggest that specific RNA and protein targets could be modulated by these proteins to increase Cat100-6 responses to Al stress. In fact, papain-like cysteine peptidases have been involved in apoptosis and wounding stress (Funk *et al.*, 2002). The putative RNA targets of LUC7 would certainly add an extra component on the complex of plant responses to Al.

A remarkable feature in plant development is the ability to exhibit a number of adaptative and protective responses to environmental stresses. Two of these genes encoding to phenylalanine ammonia-lyase (PAL, CA064763) and cinnamyl alcohol dehydrogenase (CAD, CA064602), which are important enzymes involved with lignification processes, were strongly up-regulated in the presence of Al. Lignin is a heterogenous aromatic polymer produced in the phenylpropanoid

pathway which also supplies intermediates for synthesis of phytoalexins, flavonoids and tannins (Whetten and Sederoff, 1995) and most of these compounds perform important roles in the mechanisms of plant defense. CAD is an enzyme involved with catalysis reduction of various phenylpropenyl aldehyde derivatives, that are converted to monolignols, the main precursors of lignins and lignans (Kim *et al.*, 2004). The activation of the phenylpropanoid metabolism has been reported for various biotic or abiotic stresses such as wounding, pathogen attack, UV irradiation, heavy metals, and drought (Dixon and Paiva, 1995). Even for Al-stress, the enzyme PAL has frequently been suggested as a constituent of the Al-stress alleviation process in plants (Hamel *et al.*, 1998; Snowden *et al.*, 1995; Snowden and Gardner, 1993). The enzyme PAL is supposed to play a beneficial role in detoxifying Al that has entered the symplasm by sequestration (Hamel *et al.*, 1998).

The small GTP-binding proteins (CA064605) are encoded by a multigene family and play a central role in modulation of cell growth, division, differentiation and morphogenesis, cytoskeletal organization, and vesicle trafficking (Yang and Watson, 1993). Since plants are unable to change their location, they have evolved mechanisms that allow quickly response to environmental stimulus and GTP-binding proteins play critical roles as signal transducers (Sano and Ohashi, 1995). The gene encoding a small GTP-binding protein strongly up-regulated in Cat100-6 in response to Al indicates that differences in the perception of Al might favour these plants in the activation of stress responses.

It is worth to mention that genetic analysis in Cat100-6 and S1587-17 have shown that Al tolerance is controlled by two distinct *loci* named *Alm1* and *Alm2*

located on the short arm of chromosomes 10 and 6, respectively. Together, these two *loci* are responsible to explain most of the AI tolerance trait in Cat100-6 (Sibov *et al.*, 1999). Since the distance between molecular markers and these *loci* remains large, until now none gene was linked to *Alm1* or *Alm2*. Therefore, a fine mapping of the *locus* Alm is underway and will give interesting insights on the identity of *Alm1* and *Alm2*.

In summary, in this work we have improved our understanding on the mechanisms triggered by AI stress in maize. Furthermore, the cross-species hybridization is shown to be an effective approach to study large-scale gene expression profiles between related grasses species. A major challenge that remains is the evaluation of the real contribution of each gene to the phenotype of AI tolerance in Cat100-6. Approaches using individual gene overexpression and gene knockout are underway for this propose.

Acknowledgments

We would like to thank Dr. Nicole Pence for her critical reading of this manuscript. This work was supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, grants 04/05131-7 and 04/09536-9), European Commission (project INCO II RDT ICA4-CT-2000-30017), and PADCT/CNPq (project 62.0472/98.7). MM received a Research Fellowship from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). GMAC received fellowships from Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). FTSN, RGG, and JMF were recipients of fellowships from FAPESP.

References

- Altschul, S.F., Madden, T.L., Schaffer, 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: 3389-402.
- Aubin-Horth, N., Letcher, B.H. and Hofmann, A. (2005). Interaction of rearing environment and reproductive tactic on gene expression profiles in atlantic salmon. *J. Hered.* 96: 261-278.
- Bennetzen, J.L. and Freeling, M. (1997). The unified grass genome: synergy in synteny. *Genome Res.* 7: 301-306.
- Bezerra, G.B., Cançado, G.M.A., Menossi, M., Castro, L.N. and Von Zuben, F.J. (2005). Recent advances in gene expression data clustering: a case study with comparative results. *Genet. Mol. Res.* 4: 514-524.
- Blancaflor, E.B., Jones, D.L. and Gilroy, S. (1998). Alterations in the cytoskeleton accompany aluminum-induced growth inhibition and morphological changes in primary roots of maize. *Plant Physiol.* 118: 159-172.
- Bordoli, L., Netsch, M., Lüthi, U., Lutz, W. and Eckner, R. (2001). Plant orthologs of p300/CBP: conservation of a core domain in metazoan p300/CBP acetyltransferase-related proteins. *Nucleic Acids Res.* 29: 589-597.
- Boscolo, P.R.S., Menossi, M. and Jorge, R.A. (2003). Aluminum-induced oxidative stress in maize. *Phytochemistry* 62: 181-189.
- Çakir, B., Agasse, A., Gaillard, C., Saumonneau, A., Delrot, S. and Atanassova, R. (2003). A grape ASR protein involved in sugar and abscisic acid signaling. *Plant Cell* 15: 2165–2180.
- Callebaut, I., Courvalin, J.C. and Mornon, J.P. (1999). The BAH (bromo adjacent homology) domain: a link between methylation, replication and transcriptional regulation. *FEBS Lett.* 446: 189-193.
- Cançado, G.M.A., Rosa Jr., V.E.D., Fernandez, J.H., Maron, L.G., Jorge, R.A. and Menossi, M. (2005). Glutathione S-transferase and aluminum toxicity in maize. *Funct. Plant Biol.* 32: 1045-1055.

- Carmi, N., Zhang, G., Petreikov, M., Gao, Z., Eyal, Y., Granot, D. and Schaffer, A.A. (2003). Cloning and functional expression of alkaline α -galactosidase from melon fruit: similarity to plant SIP proteins uncovers a novel family of plant glycosyl hydrolases. *Plant J.* 33: 97-106.
- Chabouté, M.E., Clément, B., Sekine, M., Philipps, G. and Chaubet-Gigot, N. (2000). Cell Cycle regulation of the tobacco ribonucleotide reductase small subunit gene is mediated by E2F-like elements. *Plant Cell* 12: 1987-2000.
- Ciereszko, I., Johansson, H., Hurry, V. and Kleczkowski, L.A. (2001). Phosphate status affects the gene expression, protein content and enzymatic activity of UDP-glucose pyrophosphorylase in wild-type and pho mutants of *Arabidopsis*. *Planta* 212: 598–605.
- Clark, L.G., Zhang, W. and Wendel, J.F. (1995) A phylogeny of the grass family (*Poaceae*) based on *ndhF* sequence data. *Syst. Bot.* 20: 436-460.
- Clayton, W.D. and Renvoize, S.A. (1986). Genera Graminum. *Grasses of the World*. Kew Bull. Add. Series XII. (H.M. Stationery Office: London.)
- Collinge, M. and Boller, T. (2001). Differential induction of two potato genes, *Stprx2* and *StNAC*, in response to infection by *Phytophthora infestans* and to wounding. *Plant Mol. Biol.* 46: 521-529.
- Cruz-Ortega, R., Cushman, J. C. and Ownby, J.D. (1997). cDNA clones encoding 1,3-beta-glucanase and a fimbrin-like cytoskeletal protein are induced by Al toxicity in wheat roots. *Plant Physiol.* 114: 1453-1460.
- Cruz-Ortega, R. and Ownby, J.D. (1993). A protein similar to PR (pathogenesis-related) proteins is elicited by metal toxicity in wheat roots. *Physiol. Plant.* 89: 211-219.
- Dixon, R.A. and Paiva, N.L. (1995). Stress-induced phenylpropanoid metabolism. *Plant Cell* 7: 1085-1097.
- Draye, X., Lin, Y.R., Qian, X.Y., Bowers, J.E., Burow, G.B., Morrell, P.L., Peterson, D.G., Presting, G.G., Ren, S.X., Wing, R.A. and Paterson, A.H. (2001). Toward integration of comparative genetic, physical, diversity, and cytomic maps for grasses and grains, using the sorghum genome as a foundation. *Plant Physiol.* 125: 1325-1341.

- Dufour, P., Deu, M., Grivet, L., D'Hont, A., Paulet, F. and Glaszmann, J.C. (1997). Construction of a composite sorghum genome map and comparison with sugarcane, a related complex polyploid. *Theor. Appl. Genet.* 94: 409-418.
- Elison, B., Blancaflor, E.B., Jones, D.L. and Gilroy, S. (1998). Alterations in the cytoskeleton accompany aluminum-induced growth inhibition and morphological changes in primary roots of maize. *Plant Physiol.* 118: 159-172.
- Ezaki, B., Gardner, R.C., Ezaki, Y. and Matsumoto, H. (2000). Expression of aluminum-induced genes in transgenic *Arabidopsis* plants can ameliorate aluminum stress and/or oxidative stress. *Plant Physiol.* 122: 657-655.
- Fierens, K., Brijs, K., Courtin, C.M., Gebruers, K., Goesaert, H., Raedschelders, G., Robben, J., Van-Campenhout, S., Volckaert, G. and Delcour, J.A. (2003). Molecular identification of wheat endoxylanase inhibitor TAXI-11, member of a new class of plant proteins. *FEBS Lett.* 540: 259-63.
- Fortes, P., Bilbao-Cortes, D., Fornerod, M., Rigaut, G., Raymond, W., Seraphin, B. and Mattaj, I.W. (1999). Luc7p, a novel yeast U1 snRNP protein with a role in 5' splice site recognition. *Genes Dev.* 13: 2425-2438.
- Foy, C.D., Chaney, R.L. and White, M.C. (1978). Physiology of metal toxicity in plants. *Annu. Rev. Plant Physiol.* 29: 511-566.
- Freeling, M. (2001). Grasses as a single genetic system. Reassessment 2001. *Plant Physiol.* 125: 1191-1197.
- Funk, V., Kositsup, B., Zhao, C. and Beers, E.P. (2002). The *Arabidopsis* xylem peptidase XCP1 is a tracheary element vacuolar protein that may be a papain ortholog. *Plant Physiol.* 128: 84-94.
- Guimarães, C.T., Sills, G.R. and Sobral, B.W.S. (1997). Comparative mapping of Andropogoneae: *Saccharum* L. (sugarcane) in relation to sorghum and maize. *Proc. Natl. Acad. Sci. USA.* 94: 14261-14266.
- Hamel, F., Breton, C. and Houde, M. (1998). Isolation and characterization of wheat aluminum-regulated genes: possible involvement of aluminum as a pathogenesis response elicitor. *Planta* 205: 531-538.

- Henry, K.W., Wyce, A., Lo, W.S., Duggan, L.J., Emre, N.C.T., Kao, C-F., Pillus, L., Shilatifard, A., Osley, M.A. and Berger, S.L. (2003). Transcriptional activation via sequential histone H2B ubiquitylation and deubiquitylation, mediated by SAGA-associated Ubp8. *Genes Dev.* 17: 2648–2663.
- Herrera-Estrella, L. (1999). Transgenic plants for tropical regions: some considerations about their development and their transfer to the small farmer. *Proc. Natl. Acad. Sci. USA.* 96: 5978-5981.
- Holst, B., Lunde, C., Lages, F., Oliveira, R., Lucas, C. and Kielland-Brandt, M.C. (2000). GUP1 and its close homologue GUP2, encoding multimembrane-spanning proteins involved in active glycerol uptake in *Saccharomyces cerevisiae*. *Mol. Microbiol.* 37: 108-124.
- Horst, W.J., Püschel, A.K. and Schmohl, N. (1997). Induction of callose formation is a sensitive marker for genotypic aluminium sensitivity in maize. *Plant Soil* 192: 23-30.
- Jones, D.L. and Kochian, L.V. (1995). Aluminum inhibition 1,4,5-triphosphate signal transduction pathway in wheat roots: a role in Al toxicity? *Plant Cell* 7: 1913-1922.
- Jorge, R.A. and Arruda, P. (1997). Aluminum-induced organic acid exudation by roots of an aluminum-tolerant tropical maize. *Phytochemistry* 45: 675-681.
- Jorge, R.A., Menossi, M. and Arruda, P. (2001). Probing the role of calmodulin in Al toxicity in maize. *Phytochemistry* 28: 415-422.
- Juge, N., Payan, F. and Williamson, G. (2004). XIP-I, a xylanase inhibitor protein from wheat: a novel protein function. *Biochim. Biophys. Acta* 1696: 203-211.
- Kasprzewska, A. (2003). Plant chitinases - regulation and function. *Cell Mol. Biol. Lett.* 8: 809-824.
- Keith, B., Dong, X., Ausubel, F.M. and Fink, G.R. (1991). Differential induction of 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase genes in *Arabidopsis thaliana* by wounding and pathogenic attack. *Proc. Natl. Acad. Sci. USA.* 88: 8821 8825.

- Keller, F. and Pharr, M. (1996). Metabolism of carbohydrates in sinks and sources: galactosyl-sucrose oligosaccharides. In: Zamski, E. and Schaffer, A.A., eds. *Photoassimilate Distribution in Plants and Crops*. Marcell Dekker, 157-184.
- Kellogg, E.A. (2001) Evolutionary history of the grasses. *Plant Physiol.* 125: 1198-1205.
- Kenzior, A.L. and Folk, W.R. (1998). AtMSI4 and RbAp48 WD-40 repeat proteins bind metal ions. *FEBS Lett.* 440: 425-429.
- Kim, S.J., Kim, M.R., Bedgar, D.L., Moinuddin, S.G.A., Cardenas, C.L., Davin, L.B., Kang, C.H. and Lewis, N.G. (2004). Functional reclassification of the putative cinnamyl alcohol dehydrogenase multigene family in *Arabidopsis*. *Proc. Natl. Acad. Sci. USA.* 101: 1455-1460.
- Kiyosue, T., Yamaguchi-Shinozaki, K. and Shinozaki, K. (1994). Cloning of cDNAs for genes that are early-responsive to dehydration stress (ERDs) in *Arabidopsis thaliana* L.: identification of three ERDs as HSP cognate genes. *Plant Mol. Biol.* 25: 791-798.
- Kmiecik, M., Simpson, C.G., Lewandowska, D., Brown, J.W.S. and Jarmolowski, A. (2002). Cloning and characterization of two subunits of *Arabidopsis thaliana* nuclear cap-binding complex. *Gene* 283: 171-183.
- Kochian, L.V., Hoekenga, O.A. and Pineros, M.A. (2004). How do crop plants tolerate acid soils? Mechanisms of aluminum tolerance and phosphorous efficiency. *Annu. Rev. Plant. Biol.* 55: 459-493.
- Kurup, S., Jones, H.D. and Holdsworth, M.J. (2000). Interactions of the developmental regulator ABI3 with proteins identified from developing *Arabidopsis* seeds. *Plant J.* 21: 143-155.
- Lee, R.H., Lin, M.C. and Chen, S.C. (2004). A novel alkaline alpha-galactosidase gene is involved in rice leaf senescence. *Plant Mol. Biol.* 55: 281-95.
- Lemieux, M.J., Huang, Y. and Wang, D.N. (2004). Glycerol-3-phosphate transporter of *Escherichia coli*: Structure, function and regulation. *Res. Microbiol.* 155: 623-629.
- Logemann, J., Schell, J. and Willmitzer, L. (1987). Improved method for the isolation of RNA from plant-tissues. *Anal. Biochem.* 163: 16-20.

- Lukaszewski, K.M. and Blevins, D.C. (1996). Root growth inhibition in boron-deficient or aluminum-stressed squash may be a result of impaired ascorbate metabolism. *Plant Physiol.* 112: 1135-1140.
- Ma, J.F., Ryan, P.R. and Delhaize, E. (2001). Aluminum tolerance in plants and complexing role of organic acids. *Trends Plant Sci.* 6: 273-278.
- MacDonald, T.L., Humphreys, W.G. and Martin, R.B. (1987). Promotion of tubulin assembly by aluminum ion in vitro. *Science* 236: 183-186.
- Mariano, E.D., Jorge, R.A., Keltjens, W.G. and Menossi, M. (2005). Metabolism and root exudation of organic acid anions under aluminium stress. *Braz. J. Plant Physiol.* 17: 157-172.
- Moon, D.H., Ottoboni, L.M.M., Souza, A.P., Sibov, S.T., Gaspar, M. and Arruda, P. (1997). Somaclonal-variation-induced aluminium sensitive mutant from an aluminum inbred maize tolerant line. *Plant Cell Rep.* 16: 686–691.
- Murata, N., Ishizaki-Nishizawa, Q., Higashi, S., Hayashi, H., Tasaka, Y. and Nishida, I. (1992). Genetically engineered alterations in the chilling sensitivity of plants. *Nature* 356: 710-713.
- Murthy, S.S. and Zilinskas, B.A. (1994) Molecular cloning and characterization of a cDNA encoding pea monodehydroascorbate reductase. *J. Biol. Chem.* 269: 31129-31133.
- Nagy, N.E., Dalen, L.S., Jones, D.L., Swensen, B., Fossdal, C.G. and Eldhuset, T.D. (2004). Cytological and enzymatic responses to aluminium stress in root tips of Norway spruce seedlings. *New Phytol.* 163: 595–607.
- Newman, T., de Bruijn, F.J., Green, P., Keegstra, K., Kende, H., McIntosh, L., Ohlrogge, J., Raikhel, N., Sommerville, S. and Thomashow, M. (1994). Genes galore: A summary of methods for accessing results from large-scale partial sequencing of anonymous *Arabidopsis* cDNA clones. *Plant Physiol.* 106: 1241-1255.
- Nogueira, F.T.S., 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.

- Nogueira, F.T.S., Schlögl, P.S., Camargo, S.R., Fernandez, J.H., De Rosa Jr., V.E., Pompermayer, P. and Arruda, P. (2005). SsNAC23, a member of the NAC domain protein family, is associated with cold, herbivory and water stress in sugarcane. *Plant Sci.* 169: 93-106.
- Renn, S.C.P., Aubin-Horth, N. and Hofmann, H.A. (2004) Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray. *BMC Genomics* 5: 42-55.
- Revenkova, E., Masson, J., Koncz, C., Afsar, K., Jakovleva, L. and Paszkowski, J. (1999). Involvement of *Arabidopsis thaliana* ribosomal protein S27 in mRNA degradation triggered by genotoxic stress. *EMBO J.* 18: 490-499.
- Richards, K.D., Schott, E.J., Sharma, Y.K., Davis, K.R. and Gardner, R.C. (1998). Aluminum induces oxidative stress genes in *Arabidopsis thaliana*. *Plant Physiol.* 116: 409-418.
- Rolland, F., Moore, B. and Sheen, J. (2002). Sugar sensing and signaling in plants. *Plant Cell* 14: 185-205.
- Ruan, Y.L., Chourey, P.S., Delmer, D.P. and Perez-Grau, L. (1997). The differential expression of sucrose synthase in relation to diverse patterns of carbon partitioning in developing cotton seed. *Plant Physiol.* 115: 375-385.
- Ryan, P.R., DiTomaso, J.M. and Kochian, L.V. (1993). Aluminum toxicity in roots: an investigation of spatial sensitivity and the role of the root cap. *J. Exp. Bot.* 44: 437-446.
- Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989). *Molecular cloning: a laboratory manual*. Cold Spring Harbor, New York, USA.
- Sano, H. and Ohashis, Y. (1995). Involvement of small GTP-binding proteins in defense signal-transduction pathways of higher plants. *Proc. Natl. Acad. Sci. USA.* 92: 4138-4144.
- Schwarzerová, K., Zelenkova, S., Nick, P. and Opatrny, Z. (2002). Aluminum-induced rapid changes in the microtubular cytoskeleton of tobacco cell lines. *Plant Cell Physiol.* 43: 207-216.
- Schummer, M., Ng, V.L.V., Baumgarner, R.E., Nelson, P.S., Schummer, B., Bednarski, D.W., Hassell, L., Baldwin, R.L., Karlan, B.Y. and Hood, L.

- (1999). Comparative hybridization of an array of 21,500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. *Gene* 238: 375-385.
- Schummer, M., Ng, W.L., Nelson, P.S., Bumgarner, R.E. and Hood, L. (1997) An inexpensive and-held device for the construction of high density nucleic acid arrays. *Biotechniques* 23: 1087-1092.
- Sibov, S.T., Gaspar, M., Silva, M.J., Ottoboni, L.M.M., Arruda, P. and Souza, A.P. (1999). Two genes control aluminum tolerance in maize: genetic and molecular mapping analyses. *Genome* 42: 475-482.
- Sivaguru, M., Baluška, F., Volkmann, D., Felle, H. and Horst, W.J. (1999). Impacts of aluminum on the cytoskeleton of the maize root apex: short-term effects on the distal part of the transition zone. *Plant Physiol.* 119: 1073-1082.
- Sivaguru, M., Fujiwara, T., Šamaj, J., Baluška-Yang, Z., Osawa, H., Maeda, T., Mori, T., Volkmann, D. and Matsumoto, H. (2000). Aluminum-induced 1–3-β-d-glucan inhibits cell-to-cell trafficking of molecules through plasmodesmata: A new mechanism of Al toxicity in plants. *Plant Physiol.* 124: 991-1018.
- Snowden, K.C. and Gardner, R.C. (1993). Five genes induced by aluminum in wheat (*Triticum aestivum* L) roots. *Plant Physiol.* 103: 855-861.
- Snowden, K.C., Richards, K.D. and Gardner, R.C. (1995). Aluminum-induced genes: induction by toxic metals, low calcium, and wounding and pattern of expression in root tips. *Plant Physiol.* 107: 341-348.
- Sugimoto, M., Saiki, Y., Zhang, D. and Kawai, F. (2004). Cloning and characterization of preferentially expressed genes in an aluminum-tolerant mutant derived from *Penicillium chrysogenum*. *FEMS Microbiol. Lett.* 15: 137-142.
- Suzuki, N., Yamaguchi, Y., Koizumi, N. and Sano, H. (2002). Functional characterization of a heavy metal binding protein Cdl19 from *Arabidopsis*. *Plant J.* 32: 165-173.
- Tusher, V.G., Tibshirani, R. and Chu, G. (2001). Significance Analysis of Microarrays. *Proc. Natl. Acad. Sci. USA.* 98: 5116-5121.

- Van Zyl, L., Von Arnold, S., Bozhkov, P., Chen, Y., Egertsdotter, U., MacKay, J., Sederoff, R.R., Shen, J., Zelena, L. and Clapham, D.H. (2002). Heterologous array analysis in Pinaceae: hybridization of *Pinus taeda* cDNA arrays with cDNA from needles and embryogenic cultures of *P. taeda*, *P. sylvestris* or *Picea abies*. *Comp. Funct. Genomics* 3: 306-318.
- Vettore, A.L., Silva, F.R., Kemper, E.L., Souza, G.M., *et al.* (2003). Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. *Genome Res.* 13: 2725-2735.
- Vettore, A.L., Silva, F.R., Kemper, E.L. and Arruda, P. (2001). The libraries that made SUCEST. *Genet. Mol. Biol.* 24: 1-7.
- Vincentz, M., Cara, F.A.A., Okura, V.K., Silva, F.R., *et al.* (2004). Evaluation of monocot and eudicot divergence using the sugarcane transcriptome. *Plant Physiol.* 134: 951-959.
- Watanabe, T., Oyanagi, W., Suzuki, K., Ohnishi, K. and Tanaka, H. (1992). Structure of the gene encoding chitinase D of *Bacillus circulans* WL-12 and possible homology of the enzyme to other prokaryotic chitinases and class III plant chitinases. *J. Bacteriol.* 174: 408-414.
- Watt, D.A. (2003). Aluminum-responsive genes in sugarcane: identification and analysis of expression under oxidative stress. *J. Exp. Bot.* 54: 1163-1174.
- Whetten, R. and Sederoff, R. (1995). Lignin Biosynthesis. *Plant Cell* 7: 1001-1013.
- Wiweger, M., Farbos, I., Ingouff, M., Lagercrantz, U. and Von-Arnold, S. (2003). Expression of Chia4-Pa chitinase genes during somatic and zygotic embryo development in Norway spruce (*Picea abies*): similarities and differences between gymnosperm and angiosperm class IV chitinases. *J. Exp. Bot.* 54: 2691-2699.
- Wolter, F.P., Schmidt, R. and Heinz, E. (1992). Chilling sensitivity of *Arabidopsis thaliana* with genetically engineered membrane lipids. *EMBO J.* 11: 4685-4692.
- Yamamoto, Y., Kobayashi, Y. and Matsumoto, H. (2001). Lipid peroxidation is an early symptom triggered by aluminum, but not the primary cause of elongation inhibition in pea roots. *Plant Physiol.* 125: 199-208.

- Yang, Z. and Watson, J.C. (1993). Molecular cloning and characterization of rho, a ras-related small GTP-binding protein from the garden pea. *Proc. Natl. Acad. Sci. USA*. 90: 8732-8736.
- Zhang, W.H. and Rengel, Z. (1999). Aluminium induces an increase in cytoplasmic Ca^{2+} in intact wheat roots. *Aust. J. Plant. Physiol.* 26: 401-409.
- Zeevaart, J. (1999). Absciscic acid metabolism and its regulation. In: Hooykaas P, Hall M, Libbenga K, eds. *Biochemistry and molecular biology of plant hormones*. Elsevier Science, 189-207.
- Xiao, K., Bai, G.H. and Carver, B.F. (2005). Nylon filter arrays reveal differential expression of expressed sequence tags in wheat roots under aluminium stress. *J. Integr. Plant. Biol.* 47: 839-848.

Capítulo 5: O algoritmo ISER - seleção de genes diferencialmente expressos por transformações de dados não lineares e ajustes locais.

O usual número reduzido de réplicas experimentais presentes nos estudos utilizando arranjos de DNA costuma limitar a aplicação de testes estatísticos aos dados. Assim, nosso grupo de pesquisa desenvolveu um algoritmo de análise de dados voltado para otimizar a seleção de genes diferencialmente expressos quando poucas réplicas dos dados estão disponíveis. O algoritmo, chamado de ISER (*Intensity-dependent **S**election of **E**xpression **R**atios*), mostrou-se eficiente em testes realizados tanto em dados simulados como num conjunto real de dados de arranjos de DNA, provando ser uma ferramenta útil à comunidade científica. A descrição do algoritmo encontra-se em artigo publicado na revista *Bioinformatics* e em seu material suplementar, ambos a seguir.

Gene expression

ISER: selection of differentially expressed genes from DNA array data by non-linear data transformations and local fitting

R. D. Drummond^{1,3}, A. Pinheiro², C. S. Rocha¹ and M. Menossi^{1,3,*}

¹Laboratório de Genoma Funcional—Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, Campinas 13083-875, PO Box 6010, Brazil, ²Departamento de Estatística, Instituto de Matemática, Estatística e Computação Científica, Universidade Estadual de Campinas, Campinas 13083-859, Brazil and ³Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas, Campinas 13083-971, Brazil

Received on May 25, 2005; revised on and accepted on October 17, 2005

Advance Access publication October 25, 2005

ABSTRACT

Summary: This report describes an algorithm (intensity-dependent selection of expression ratios or ISER) developed to analyse DNA array data by optimizing the selection of genes with the most significant variations in expression amongst two RNA samples. The algorithm is designed for use when little or no replication of array hybridizations is available.

Availability: ISER is written in R language, and its code and on-line version are freely available at <https://ipe.cbmeg.unicamp.br/pub/PMmA>

Contact: menossi@unicamp.br

Supplementary information: <https://ipe.cbmeg.unicamp.br/pub/ISER>

DNA array technology is the main strategy for assessing gene expression profiles on a large scale. An important application of DNA arrays is the identification of genes that show significant changes in their expression in different RNA samples. However, because of the characteristic noise of array data and the usually limited number of experimental replicates, a statistical framework to select differentially expressed genes is not easily determined (Lee *et al.*, 2000). This is especially the case in macroarray experiments (Freeman *et al.*, 2000), where the requirement for larger amounts of RNA for each hybridization can be a limiting factor in the number of replicates, and each labelled cDNA sample is hybridized to a different nylon membrane, thereby increasing the data variability (in contrast to microarrays, in which two samples labelled with different dyes are hybridized on the same slide; Schena *et al.*, 1996).

Some studies have developed strategies to select differentially expressed genes (Tusher *et al.*, 2001; Kerr *et al.*, 2000), but they usually require three or more experimental replicates to evaluate the data variability inherent to the method. However, in array experiments, it is often necessary to compare two samples without replicates, and in this case a strategy frequently adopted is to select those genes whose signals from both samples show a fold change of two or more (or any other arbitrarily fixed value; Schena *et al.*,

1996). Other methods have also been developed to select differentially expressed genes when little experimental replication is available, e.g. the methods described by Mutch *et al.* (2002) and Loguinov *et al.* (2004). In this report, we describe an algorithm (intensity-dependent selection of expression ratios or ISER) designed to optimize the selection of genes that show the most significant variations in expression. One of the novelties of this method is the non-linear transformation to normality via a sliding window of variable size that allows the use of a normal distribution instead of heavier tails distributions and ensures accuracy of the *P*-values.

The algorithm is based on the commonly adopted assumption that most of the genes probed in arrays have a constant expression in the samples studied (Quackenbush, 2002) so that differentially expressed genes are identified as those that are outliers in the global distribution of expression ratios. Another (empirically supported) aspect that ISER deals successfully with is that the variance of the expression ratios depends on the mean log intensity of the gene (Mutch *et al.*, 2002). The initial motivation for ISER is to find a data transformation that results in a normal distribution of the expression ratios (Box and Cox, 1964), which are generally considered to have a lognormal distribution (Friddle *et al.*, 2000). In this case, ISER will return a log transformation. However, if small or large variations from this transformation are present, the sliding window will adapt it locally to the data.

The input for ISER is a text file containing the signal intensities (raw or normalized, if desired) for both RNA samples of each gene. The algorithm follows the steps below.

- (1) To avoid working with zero values, a small constant (10^{-4} times the minimum expression value different from zero) is added to all expression values. For each gene, the expression ratio and the geometric mean of the signals from both samples are calculated.
- (2) A data transformation is found that provides the transformed ratio values closest to the normal distribution. This is done by

*To whom correspondence should be addressed.

choosing the best exponent lambda in the Box–Cox family of transformations:

$$BC(x) = \begin{cases} \log(x) & \text{if } \lambda = 0, \\ \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0. \end{cases}$$

The most suitable transformation is then applied to the expression ratios and to the geometric means of the expression values, to yield what are referred to, respectively, as the transformed ratios and transformed intensities.

- (3) The genes are sorted according to their transformed intensities, in ascending order. A sliding window over this order is used to locally estimate the mean and standard deviation of the transformed expression ratios. The size of this window is determined locally so as to optimize the value of the Kolmogorov–Smirnov test for normality (Lehman, 1994), applied to the transformed ratios of the genes inside the window. Within each sliding window, the algorithm computes the trimmed mean and standard deviation (with respect to the trimmed mean) of the transformed ratios, both based on their 60% central values. These estimates and the *P*-value defined by the user are employed to determine the local upper and lower limits of the transformed ratios, outside of which the genes are considered as differentially expressed.
- (4) A linear spline is used to approximate the relationship between these limits and the mean of the transformed intensities in each sliding window, thereby establishing the intensity-dependent limits for the transformed ratios to be selected as differentially expressed.

The performance of ISER was tested on simulated datasets generated based on the ANOVA model applied to DNA arrays described by Kerr *et al.* (2000). These datasets mimicked the results of array experiments with 3000 genes in which different numbers of genes were previously assumed as differentially expressed (induced/repressed). The model used also simulated an intensity-dependent bias; further details of the model are given in the Supplementary Information. These datasets were also analysed with the algorithm described by Mutch *et al.* (2002), as well as with the fixed fold criterion for the selection of differentially expressed genes. Local fitting performance for selecting differentially expressed genes should be superior in cases with few replications. Thus both ISER and the method described by Mutch *et al.* (2002) achieved better results than the fixed fold criterion (Fig. 1). Moreover, the ISER algorithm is more sensitive to intensity-dependent bias in the data, thus showing a performance superior to both of the other methods, as can be seen in Figure 1. This superiority was true for each false positive rate and for each of the dataset specifications analysed. We also tested ISER for its comparative performance when applied to raw or loess-normalized values (see Supplementary Information). The results showed no significant difference.

ISER was also tested with real datasets from Nogueira *et al.* (2003), studying the genetic response of sugarcane to low temperatures. It confirmed the differential expression of 46 of the 59 genes originally selected by those authors and identified an additional set of 147 genes as differentially expressed (see Supplementary Information). This experiment was performed with two replicates, and the original data analysis consisted of calculating, for each

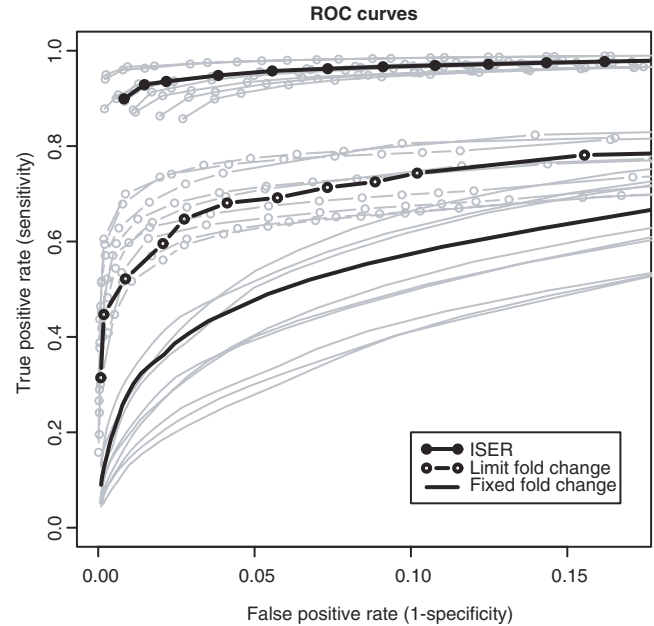


Fig. 1. Receiver operating characteristic (ROC) curves for the three methods used to analyse simulated datasets. Each grey line is the mean curve for a subset of datasets with the same specifications (see Supplementary Information). For each method of analysis, the black line is the median of the grey lines.

replicate and each time point (3, 6, 12, 24 and 48 h), the mean and standard deviation of the logarithms of the expression ratios— $\log(\text{treated/control})$ and selecting those genes that, in both replicates, showed a logarithm of the expression more than 1.65 SD distant from the mean, and a fold-change >2 . Since this analysis is not sensitive to intensity-dependent bias in the data, it may select more false-positive genes than ISER, and this might be the case of the 13 genes originally selected, which were rejected by the algorithm. In fact, the majority (9) of these genes were selected as induced at the time point of 48 h, although in one of the experimental replicates for this time point, their intensities were contained in a region where ISER detected a greater variability in the data, thus not selecting them. On the other hand, ISER additionally selected several genes of biological relevance, including transcription factors, genes related to cellular communication (transmembrane kinases), stress response (heat shock protein and others) and protein metabolism. A complete table of the differentially expressed genes identified by the algorithm is available in the Supplementary Information.

In several studies using DNA arrays, a subset of the arrayed genes is considered as housekeeping, i.e. genes that are supposed to have a constant expression in the samples being studied. ISER can use this assumption instead of supposing that most of the genes probed in the arrays have a constant expression. In this case, steps 2, 3 and 4 of the algorithm are carried out using only the housekeeping genes, thus establishing the intensity-dependent criterion of selection of the differentially expressed genes, based solely on those genes. This criterion is then applied to all the genes probed in the arrays.

Although ISER is designed to deal with data from single replicated experiments, comparing two RNA samples, it can be used when more samples are compared or replicates are present. In this case,

the algorithm is applied separately to the comparison of each pair of samples and to each replicate, thus selecting genes that, for any pair of samples, are identified as differentially expressed on all replicates or on a large (pre-defined) proportion of them. The proposed algorithm avoids lognormal assumptions and shows similar performances for raw and normalized data. ISER is reasonably fast and does not have cumbersome memory requirements, thus providing researchers with a very useful tool for analysing array data.

ACKNOWLEDGEMENTS

We would like to thank the two anonymous reviewers for their helpful comments and suggestions. R.D.D. was supported by a fellowship from a federal funding agency (CAPES—Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazil), C.S.R. was supported by a fellowship from the UNIEMP Institute (Brazil) and M.M. received a research fellowship from another federal funding agency (CNPq—Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil). This work was partially supported by grants 02/01167-1 and 03/07244-0 from the São Paulo State funding agency (FAPESP—Fundação de Amparo à Pesquisa do Estado de São Paulo, Brazil).

Conflict of Interest: none declared

REFERENCES

- Box, G.E.P. and Cox, D.R. (1964) An analysis of transformations. *J. R. Stat. Soc. Ser. B*, **26**, 211–252.
- Freeman, W.M. *et al.* (2000) Fundamentals of DNA hybridization arrays for gene expression analysis. *BioTechniques*, **29**, 1042–1055.
- Friddle, C.J. *et al.* (2000) Expression profiling reveals distinct sets of genes altered during induction and regression of cardiac hypertrophy. *Proc. Natl Acad. Sci. USA*, **97**, 6745–6750.
- Kerr, M.K. *et al.* (2000) Analysis of variance for gene expression microarray data. *J. Comp. Biol.*, **7**, 819–837.
- Lee, M.L. *et al.* (2000) Importance of replication in microarrays gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl Acad. Sci. USA*, **97**, 9834–9839.
- Lehman, E. (1994) *Testing Statistical Hypotheses*, 2nd Edn. Chapman Hall, New York.
- Loguinov, A.V. *et al.* (2004) Exploratory differential gene expression analysis in microarray experiments with no or limited replication. *Genome Biol.*, **5**, R18.
- Mutch, D.M. *et al.* (2002) The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data. *BMC Bioinformatics*, **3**, 17.
- Nogueira, F.T. *et al.* (2003) RNA expression profiles and data mining of sugarcane response to low temperatures. *Plant Physiol.*, **132**, 1811–1824.
- Quackenbush, J. (2002) Microarray data normalization and transformation. *Nat. Genet.*, **32**, 496–501.
- Schena, M. *et al.* (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl Acad. Sci. USA*, **93**, 10614–10619.
- Tusher, V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

ISER: Application to simulated data sets

Simulation

The performance of ISER was tested on simulated data sets generated based on the ANOVA model applied to DNA arrays described by Kerr et al. (2000). These data sets mimicked the results of array experiments with 3000 genes in which different numbers of genes were previously assumed to be differentially expressed (induced/repressed). Each expression value, S_{ij} , of gene i on array/die j (respectively in case of macro and microarrays), was generated as

$$\log_2(S_{ij}) = M + G_i + D_j + F_{ij} + e_{ij},$$

where:

- M = mean signal intensity, with the same value being used for all genes. This value was generated, for each data set, as a sum of a general mean (U , generated as $N(5, 2^2)$) and a replicate mean (A , which varies between replicates of the experiment, when more than one replicate is present in the model, and is generated as $N(6, 3^2)$).
- G_i = mean gene intensity; which is generated from a uniform distribution independently for each gene.
- D_j = dye effect, in case of microarrays. In macroarrays it can be seen as the effect of the different arrays used for the hybridization of each sample. For each column of data (treated and control), it is generated from a distribution $N(0, 0.5^2)$.
- F_j = log fold change, which differs from zero only for differentially expressed genes, where it is generated from a gamma distribution (Lehman 1994) with different values of mean and standard deviation (see below). In half of the genes assumed as differentially expressed, these values were added to the "treated" column of the data set, simulating induced genes, and in the other half they were added to the "control" column, simulating repressed genes.
- e_{ij} = measurement error; generated, for each expression value in the simulated data sets, from a normal distribution with a mean of zero and a standard deviation proportional to the inverse of the gene intensity.

After that, an intensity-dependent bias was added to the data sets. The signal intensity of each gene (mean of $\log(S_{i1})$ and $\log(S_{i2})$) was calculated, and the minimum, maximum, 0.3-quantil and 0.7-quantil of the intensities were obtained. Then, for each of these four values, a fluctuation is randomly generated (the fluctuation at the minimum, y_1 , is taken from a normal distribution $N(0, 0.5^2)$, the fluctuation at the 0.3-quantil, y_2 , is taken from a normal distribution $N(y_1, 0.25^2)$, the fluctuation at the 0.7-quantil, y_3 , is taken from a normal distribution $N(y_2, 0.25^2)$ and the fluctuation at the maximum, y_4 , is taken from a normal distribution $N(y_3, 0.25^2)$, and after that we ensure that y_2 , y_3 and y_4 have the same signal as y_1 , changing signals whenever needed) and a cubic spline interpolation is used to generate a fluctuation function for all intensity values. This fluctuation is added to data by summing half of it to $\log(S_{i1})$ and subtracting half of it from $\log(S_{i2})$, thus altering the gene expression ratio without changing its intensity.

After all, the signal generated could be written as

$$\log_2(S_{ij}) = U + A + G_i + D_j + F_{ij} + e_{ij} + \text{fluctuation}(\text{intensity})$$

Table 1: Groups of parameters of the gamma distribution used to generate fold change logs.

Group	Mean	Std. Deviation
1	1.0	0.1
2	1.0	0.5
3	1.0	1.0
4	1.5	0.1
5	1.5	0.5
6	1.5	1.0
7	2.0	0.1
8	2.0	0.5
9	2.0	1.0

Each data set represented the results of two macroarray hybridizations (or one hybridization of a microarray with two samples marked with different dyes) containing 3000 genes to be compared. The number of genes with a fold change different from zero and the mean and standard deviation adopted in the gamma distribution varied among simulations. The number of differentially expressed genes was either 100, 200 or 300, half of them induced and half repressed. The distribution means considered were 1, 1.5 or 2 for induced genes and -1 , -1.5 or -2 for repressed ones; the distribution standard deviations were fixed at 0.1, 0.5 or 1. Altogether, these parameters led to 27 different simulation setups, and 300 data sets of each type were simulated. The setup parameters were organized as shown in table 1.

Comparison with other methods of analysis

Simulated data sets were analyzed with ISER as well as with the algorithm described by Mutch *et al.* (2002) and the fixed fold criterion for the selection of differentially expressed genes. The simulation results are shown in tables 2-4 for 100, 200 and 300 differentially expressed genes, respectively. In each simulation, the induced genes were defined as those with a log fold change greater than 1, corresponding to a fold change greater than 2. The repressed genes were defined as those with a log fold change lower than -1 , corresponding to a fold change lower than $1/2$. The percentages of truly induced and truly repressed genes were defined as the percentages of these genes correctly selected by the algorithm.

The following notation is used in tables 2-13.

- **Group** - subset of simulations, as in Table 1.
- **ISER** - results obtained with the algorithm.
- **LFC** - results obtained with the algorithm Limit Fold Change.
- **FCx** - results obtained with the fixed fold criterion of selection.
- **p** - p-value used in ISER and LFC tests.
- **FC** - Fold change cut-off used in FCx criterion.
- **FP** - False positive rate (1-specificity) - mean fraction of genes without differential expression erroneously selected as differentially expressed genes.
- **TP** - True positive rate (sensitivity) - mean fraction of differentially expressed genes correctly selected by each method

Comparison of ISER performance for raw and loess-normalized data

To compare the performance of ISER when dealing with raw or loess-normalized data, 100 simulated data sets were generated for each one of the 27 simulation setups described above, and analyzed directly or after being loess-normalized. The results showed no significant difference, as can be seen in Tables 11 to 13 and in Figure 2.

Table 2: ISER results with 100 differentially expressed genes.

p	Group 1		Group 2		Group 3		Group 4		Group 5		Group 6		Group 7		Group 8		Group 9	
	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP
0.10%	0.015	0.786	0.012	0.871	0.008	0.919	0.002	0.9	0.006	0.899	0.008	0.919	0.002	0.964	0.002	0.95	0.005	0.939
0.50%	0.023	0.841	0.021	0.906	0.017	0.942	0.009	0.931	0.013	0.929	0.016	0.942	0.009	0.977	0.009	0.966	0.012	0.956
1%	0.032	0.859	0.03	0.916	0.026	0.947	0.018	0.937	0.021	0.936	0.025	0.948	0.018	0.978	0.018	0.968	0.021	0.96
2%	0.049	0.884	0.048	0.934	0.045	0.957	0.035	0.95	0.038	0.948	0.043	0.957	0.035	0.984	0.035	0.975	0.038	0.968
3%	0.066	0.899	0.065	0.944	0.063	0.964	0.052	0.957	0.056	0.956	0.06	0.963	0.052	0.987	0.052	0.979	0.056	0.973
4%	0.083	0.911	0.083	0.949	0.081	0.968	0.069	0.962	0.073	0.961	0.078	0.968	0.069	0.989	0.07	0.981	0.074	0.977
5%	0.101	0.919	0.101	0.954	0.1	0.971	0.087	0.966	0.091	0.966	0.096	0.971	0.087	0.991	0.088	0.984	0.092	0.979
6%	0.119	0.926	0.12	0.958	0.118	0.974	0.105	0.969	0.109	0.969	0.114	0.973	0.105	0.992	0.106	0.985	0.11	0.982
7%	0.137	0.932	0.138	0.961	0.137	0.976	0.123	0.972	0.127	0.972	0.133	0.975	0.124	0.992	0.125	0.987	0.128	0.983
8%	0.155	0.938	0.156	0.965	0.156	0.978	0.142	0.974	0.146	0.974	0.151	0.977	0.142	0.993	0.143	0.988	0.147	0.984
9%	0.174	0.942	0.175	0.966	0.174	0.98	0.16	0.976	0.165	0.976	0.17	0.979	0.161	0.994	0.162	0.989	0.165	0.985
10%	0.192	0.947	0.194	0.969	0.193	0.981	0.179	0.977	0.183	0.977	0.189	0.98	0.18	0.994	0.181	0.99	0.184	0.986

Table 3: ISER results with 200 differentially expressed genes.

p	Group 1		Group 2		Group 3		Group 4		Group 5		Group 6		Group 7		Group 8		Group 9	
	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP
0.10%	0.027	0.772	0.02	0.863	0.013	0.914	0.002	0.891	0.008	0.894	0.012	0.919	0.002	0.957	0.002	0.94	0.007	0.934
0.50%	0.035	0.829	0.029	0.902	0.022	0.938	0.007	0.925	0.015	0.924	0.02	0.943	0.007	0.972	0.008	0.96	0.013	0.955
1%	0.042	0.851	0.037	0.913	0.031	0.943	0.014	0.933	0.022	0.932	0.029	0.948	0.014	0.972	0.015	0.963	0.021	0.959
2%	0.057	0.879	0.054	0.931	0.048	0.954	0.028	0.946	0.036	0.946	0.044	0.958	0.028	0.978	0.029	0.972	0.036	0.968
3%	0.072	0.895	0.07	0.941	0.066	0.961	0.044	0.955	0.052	0.954	0.06	0.964	0.044	0.982	0.044	0.977	0.052	0.973
4%	0.088	0.905	0.086	0.948	0.083	0.966	0.059	0.96	0.068	0.959	0.077	0.968	0.059	0.984	0.06	0.98	0.067	0.976
5%	0.104	0.914	0.103	0.953	0.1	0.97	0.076	0.965	0.084	0.963	0.093	0.972	0.076	0.986	0.077	0.982	0.084	0.978
6%	0.121	0.921	0.12	0.957	0.118	0.972	0.093	0.968	0.101	0.967	0.11	0.974	0.093	0.988	0.094	0.984	0.101	0.98
7%	0.137	0.927	0.137	0.961	0.135	0.974	0.11	0.971	0.118	0.969	0.128	0.976	0.109	0.989	0.111	0.986	0.118	0.982
8%	0.154	0.932	0.155	0.963	0.154	0.976	0.127	0.973	0.135	0.971	0.145	0.978	0.127	0.99	0.128	0.987	0.135	0.984
9%	0.172	0.937	0.172	0.966	0.172	0.978	0.145	0.975	0.153	0.973	0.163	0.98	0.144	0.991	0.146	0.988	0.153	0.985
10%	0.189	0.941	0.19	0.968	0.19	0.979	0.163	0.977	0.171	0.975	0.181	0.982	0.162	0.992	0.164	0.989	0.171	0.986

Table 4: ISER results with 300 differentially expressed genes.

p	Group 1		Group 2		Group 3		Group 4		Group 5		Group 6		Group 7		Group 8		Group 9	
	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP
0.10%	0.038	0.757	0.027	0.857	0.017	0.906	0.001	0.862	0.011	0.876	0.017	0.908	0.001	0.878	0.002	0.875	0.009	0.89
0.50%	0.047	0.821	0.037	0.899	0.027	0.934	0.006	0.905	0.017	0.913	0.025	0.936	0.006	0.904	0.006	0.907	0.015	0.923
1%	0.054	0.843	0.046	0.911	0.035	0.942	0.011	0.916	0.023	0.922	0.032	0.944	0.012	0.914	0.012	0.918	0.021	0.934
2%	0.067	0.873	0.061	0.928	0.052	0.954	0.023	0.935	0.036	0.939	0.047	0.956	0.024	0.929	0.024	0.937	0.034	0.948
3%	0.081	0.89	0.076	0.938	0.068	0.961	0.037	0.947	0.05	0.948	0.062	0.962	0.037	0.939	0.038	0.948	0.048	0.957
4%	0.095	0.902	0.091	0.945	0.084	0.965	0.051	0.954	0.064	0.954	0.077	0.967	0.052	0.946	0.052	0.956	0.063	0.964
5%	0.11	0.912	0.106	0.95	0.101	0.969	0.065	0.959	0.079	0.959	0.092	0.971	0.066	0.952	0.067	0.962	0.078	0.968
6%	0.125	0.919	0.122	0.955	0.117	0.972	0.08	0.963	0.094	0.963	0.108	0.973	0.082	0.958	0.082	0.967	0.093	0.972
7%	0.14	0.926	0.138	0.958	0.134	0.974	0.096	0.966	0.11	0.965	0.124	0.975	0.098	0.964	0.098	0.971	0.109	0.974
8%	0.156	0.931	0.154	0.962	0.152	0.976	0.112	0.969	0.127	0.968	0.14	0.977	0.114	0.969	0.114	0.975	0.126	0.977
9%	0.172	0.936	0.171	0.965	0.169	0.978	0.129	0.972	0.143	0.97	0.157	0.979	0.131	0.973	0.131	0.977	0.142	0.979
10%	0.189	0.941	0.188	0.967	0.187	0.979	0.147	0.974	0.16	0.972	0.174	0.98	0.147	0.977	0.148	0.98	0.159	0.981

Table 5: LFC results with 100 differentially expressed genes.

p	Group 1		Group 2		Group 3		Group 4		Group 5		Group 6		Group 7		Group 8		Group 9	
	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP
1%	0.008	0.432	0.005	0.535	0.008	0.596	0.001	0.450	0.001	0.452	0.004	0.503	0.000	0.480	0.001	0.437	0.001	0.419
2%	0.016	0.498	0.015	0.607	0.023	0.689	0.002	0.553	0.005	0.571	0.010	0.667	0.001	0.605	0.002	0.602	0.007	0.614
3%	0.035	0.536	0.033	0.635	0.043	0.714	0.012	0.620	0.016	0.630	0.023	0.724	0.007	0.700	0.009	0.695	0.016	0.701
4%	0.054	0.557	0.052	0.650	0.061	0.726	0.030	0.655	0.036	0.658	0.041	0.747	0.023	0.743	0.023	0.741	0.033	0.739
5%	0.071	0.571	0.068	0.661	0.076	0.733	0.049	0.673	0.054	0.674	0.060	0.759	0.042	0.762	0.041	0.763	0.051	0.756
6%	0.085	0.582	0.081	0.669	0.088	0.740	0.066	0.684	0.071	0.685	0.075	0.766	0.059	0.772	0.059	0.773	0.067	0.765
7%	0.098	0.591	0.093	0.674	0.099	0.745	0.082	0.692	0.084	0.690	0.088	0.771	0.076	0.779	0.076	0.779	0.083	0.770
8%	0.110	0.598	0.105	0.679	0.110	0.750	0.095	0.698	0.097	0.695	0.100	0.776	0.091	0.783	0.089	0.783	0.096	0.775
9%	0.121	0.605	0.117	0.684	0.121	0.753	0.108	0.702	0.109	0.700	0.112	0.779	0.104	0.787	0.101	0.787	0.108	0.779
10%	0.132	0.610	0.127	0.688	0.131	0.756	0.119	0.706	0.120	0.704	0.123	0.782	0.116	0.790	0.113	0.791	0.120	0.782
15%	0.181	0.649	0.179	0.718	0.181	0.781	0.169	0.735	0.171	0.734	0.172	0.809	0.164	0.826	0.161	0.813	0.165	0.814
20%	0.231	0.665	0.229	0.730	0.231	0.789	0.219	0.747	0.222	0.744	0.223	0.818	0.216	0.834	0.213	0.821	0.216	0.822

Table 6: LFC results with 200 differentially expressed genes.

p	Group 1		Group 2		Group 3		Group 4		Group 5		Group 6		Group 7		Group 8		Group 9	
	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP
1%	0.008	0.332	0.003	0.408	0.003	0.432	0.000	0.290	0.000	0.269	0.004	0.294	0.000	0.301	0.000	0.241	0.001	0.227
2%	0.013	0.418	0.008	0.515	0.007	0.619	0.001	0.409	0.001	0.418	0.003	0.478	0.000	0.431	0.000	0.382	0.001	0.388
3%	0.017	0.476	0.015	0.570	0.017	0.687	0.001	0.482	0.003	0.522	0.007	0.600	0.001	0.514	0.001	0.492	0.003	0.521
4%	0.024	0.522	0.028	0.605	0.033	0.719	0.003	0.542	0.006	0.596	0.014	0.670	0.001	0.594	0.002	0.578	0.008	0.623
5%	0.037	0.557	0.045	0.627	0.050	0.735	0.009	0.598	0.013	0.647	0.025	0.710	0.004	0.679	0.008	0.645	0.015	0.691
6%	0.053	0.581	0.061	0.641	0.067	0.745	0.023	0.634	0.024	0.681	0.041	0.731	0.011	0.735	0.020	0.693	0.027	0.735
7%	0.069	0.598	0.076	0.652	0.080	0.752	0.039	0.658	0.038	0.702	0.057	0.743	0.023	0.766	0.034	0.721	0.040	0.760
8%	0.083	0.610	0.090	0.658	0.093	0.757	0.054	0.672	0.053	0.716	0.073	0.751	0.035	0.783	0.049	0.737	0.056	0.774
9%	0.097	0.619	0.103	0.664	0.105	0.761	0.067	0.680	0.069	0.725	0.089	0.757	0.047	0.793	0.064	0.747	0.072	0.783
10%	0.110	0.627	0.115	0.670	0.117	0.765	0.082	0.687	0.083	0.731	0.102	0.762	0.062	0.800	0.079	0.755	0.088	0.789
15%	0.167	0.660	0.169	0.696	0.167	0.787	0.144	0.721	0.143	0.763	0.155	0.797	0.129	0.838	0.140	0.787	0.145	0.824
20%	0.219	0.677	0.221	0.709	0.219	0.795	0.198	0.733	0.197	0.775	0.208	0.807	0.186	0.846	0.194	0.795	0.199	0.832

Table 7: LFC results with 300 differentially expressed genes.

p	Group 1		Group 2		Group 3		Group 4		Group 5		Group 6		Group 7		Group 8		Group 9	
	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP
1%	0.008	0.257	0.001	0.315	0.001	0.316	0.000	0.195	0.000	0.183	0.001	0.197	0.000	0.201	0.000	0.158	0.001	0.155
2%	0.013	0.341	0.005	0.447	0.005	0.509	0.001	0.300	0.001	0.297	0.001	0.345	0.000	0.310	0.000	0.263	0.001	0.266
3%	0.018	0.402	0.011	0.516	0.010	0.614	0.001	0.385	0.001	0.400	0.003	0.475	0.001	0.401	0.000	0.360	0.001	0.377
4%	0.023	0.449	0.018	0.561	0.021	0.671	0.002	0.446	0.003	0.482	0.007	0.568	0.001	0.464	0.001	0.445	0.002	0.476
5%	0.028	0.488	0.027	0.593	0.033	0.701	0.002	0.495	0.006	0.544	0.011	0.632	0.001	0.517	0.001	0.510	0.004	0.560
6%	0.036	0.522	0.040	0.616	0.048	0.719	0.004	0.538	0.010	0.598	0.019	0.674	0.002	0.571	0.002	0.570	0.009	0.625
7%	0.048	0.550	0.054	0.632	0.064	0.731	0.009	0.582	0.017	0.640	0.030	0.703	0.004	0.628	0.007	0.623	0.016	0.676
8%	0.064	0.571	0.069	0.644	0.079	0.739	0.021	0.618	0.028	0.672	0.044	0.721	0.014	0.683	0.016	0.666	0.025	0.713
9%	0.079	0.586	0.084	0.653	0.093	0.744	0.035	0.645	0.041	0.694	0.058	0.735	0.026	0.721	0.028	0.700	0.037	0.741
10%	0.093	0.598	0.098	0.660	0.107	0.749	0.051	0.661	0.053	0.710	0.073	0.744	0.039	0.743	0.040	0.723	0.050	0.758
15%	0.153	0.637	0.155	0.692	0.160	0.776	0.110	0.714	0.116	0.759	0.133	0.785	0.092	0.806	0.097	0.788	0.110	0.811
20%	0.209	0.655	0.209	0.706	0.212	0.786	0.172	0.729	0.177	0.774	0.190	0.796	0.161	0.818	0.165	0.800	0.172	0.823

Table 8: FCx results with 100 differentially expressed genes.

FC	Group 1		Group 2		Group 3		Group 4		Group 5		Group 6		Group 7		Group 8		Group 9	
	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP
2	0.298	0.538	0.304	0.661	0.320	0.763	0.302	0.688	0.326	0.708	0.284	0.788	0.299	0.846	0.310	0.818	0.316	0.814
2.25	0.226	0.463	0.237	0.599	0.245	0.716	0.229	0.623	0.244	0.653	0.217	0.739	0.225	0.802	0.242	0.776	0.245	0.774
2.5	0.174	0.398	0.186	0.541	0.190	0.673	0.176	0.566	0.184	0.602	0.168	0.690	0.171	0.755	0.193	0.732	0.195	0.735
2.75	0.134	0.340	0.149	0.486	0.150	0.627	0.135	0.515	0.140	0.554	0.130	0.646	0.133	0.707	0.156	0.690	0.158	0.697
3	0.103	0.296	0.122	0.442	0.119	0.589	0.106	0.468	0.108	0.511	0.100	0.602	0.105	0.660	0.124	0.648	0.130	0.662
3.25	0.079	0.256	0.101	0.404	0.095	0.551	0.083	0.421	0.083	0.470	0.077	0.565	0.085	0.617	0.098	0.607	0.108	0.627
3.5	0.061	0.222	0.085	0.361	0.076	0.517	0.067	0.382	0.064	0.434	0.061	0.529	0.070	0.576	0.078	0.567	0.090	0.594
3.75	0.046	0.193	0.071	0.327	0.061	0.484	0.054	0.345	0.049	0.400	0.050	0.496	0.058	0.538	0.062	0.531	0.074	0.564
4	0.034	0.169	0.059	0.297	0.048	0.453	0.044	0.310	0.038	0.368	0.041	0.466	0.050	0.503	0.049	0.500	0.061	0.534
4.25	0.026	0.150	0.050	0.270	0.039	0.428	0.036	0.278	0.029	0.339	0.034	0.437	0.043	0.469	0.040	0.469	0.050	0.507
4.5	0.019	0.133	0.042	0.246	0.032	0.404	0.029	0.251	0.023	0.313	0.029	0.413	0.037	0.437	0.032	0.440	0.041	0.483
4.75	0.015	0.115	0.035	0.225	0.026	0.382	0.024	0.226	0.018	0.287	0.024	0.389	0.033	0.406	0.026	0.412	0.034	0.459
5	0.011	0.102	0.030	0.206	0.021	0.362	0.020	0.204	0.014	0.263	0.021	0.368	0.028	0.378	0.022	0.386	0.028	0.439
5.25	0.009	0.089	0.025	0.188	0.018	0.345	0.016	0.183	0.012	0.243	0.018	0.349	0.024	0.350	0.018	0.363	0.023	0.418
5.5	0.007	0.078	0.021	0.173	0.015	0.328	0.013	0.166	0.010	0.226	0.016	0.331	0.021	0.324	0.014	0.340	0.019	0.399
5.75	0.006	0.069	0.018	0.159	0.012	0.312	0.011	0.150	0.008	0.206	0.014	0.315	0.018	0.301	0.012	0.319	0.016	0.382
6	0.005	0.062	0.016	0.146	0.010	0.298	0.009	0.137	0.007	0.191	0.012	0.298	0.016	0.278	0.010	0.302	0.014	0.366
6.25	0.004	0.054	0.013	0.134	0.008	0.285	0.008	0.125	0.006	0.177	0.010	0.284	0.014	0.257	0.008	0.284	0.012	0.351
6.5	0.004	0.048	0.012	0.124	0.007	0.271	0.006	0.113	0.005	0.165	0.009	0.271	0.012	0.238	0.007	0.268	0.010	0.336
6.75	0.003	0.043	0.010	0.115	0.006	0.260	0.005	0.103	0.004	0.152	0.008	0.257	0.011	0.219	0.006	0.251	0.009	0.322
7	0.003	0.038	0.009	0.107	0.005	0.249	0.004	0.094	0.004	0.140	0.007	0.245	0.010	0.203	0.005	0.237	0.008	0.309
7.25	0.003	0.034	0.008	0.098	0.004	0.239	0.003	0.085	0.003	0.130	0.006	0.234	0.009	0.188	0.004	0.223	0.007	0.298
7.5	0.002	0.031	0.007	0.092	0.004	0.229	0.003	0.078	0.002	0.121	0.005	0.224	0.008	0.175	0.004	0.210	0.006	0.286
7.75	0.002	0.027	0.006	0.087	0.003	0.221	0.002	0.071	0.002	0.112	0.004	0.215	0.007	0.163	0.003	0.199	0.006	0.274
8	0.002	0.024	0.005	0.081	0.002	0.212	0.002	0.065	0.002	0.104	0.003	0.206	0.006	0.152	0.003	0.188	0.005	0.263
8.25	0.002	0.020	0.004	0.076	0.002	0.203	0.002	0.059	0.001	0.097	0.003	0.198	0.005	0.142	0.002	0.177	0.005	0.253
8.5	0.001	0.018	0.003	0.071	0.002	0.196	0.001	0.054	0.001	0.091	0.002	0.190	0.004	0.133	0.002	0.168	0.005	0.244
8.75	0.001	0.015	0.003	0.066	0.001	0.189	0.001	0.050	0.001	0.084	0.002	0.182	0.004	0.125	0.002	0.158	0.004	0.234
9	0.001	0.013	0.002	0.063	0.001	0.182	0.001	0.046	0.001	0.078	0.002	0.175	0.003	0.116	0.002	0.150	0.004	0.226
9.25	0.001	0.012	0.002	0.060	0.001	0.175	0.001	0.042	0.001	0.072	0.002	0.168	0.003	0.109	0.001	0.144	0.004	0.218
9.5	0.001	0.011	0.002	0.056	0.001	0.169	0.001	0.039	0.001	0.068	0.001	0.162	0.002	0.102	0.001	0.136	0.004	0.211
9.75	0.001	0.010	0.001	0.053	0.001	0.163	0.001	0.035	0.001	0.064	0.001	0.157	0.002	0.096	0.001	0.129	0.003	0.203
10	0.001	0.009	0.001	0.050	0.001	0.156	0.001	0.033	0.001	0.060	0.001	0.151	0.001	0.090	0.001	0.123	0.003	0.196

Table 9: FCx results with 200 differentially expressed genes.

FC	Group 1		Group 2		Group 3		Group 4		Group 5		Group 6		Group 7		Group 8		Group 9	
	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP
2	0.293	0.539	0.343	0.648	0.307	0.765	0.326	0.686	0.270	0.728	0.288	0.782	0.275	0.860	0.352	0.804	0.291	0.822
2.25	0.217	0.460	0.261	0.589	0.236	0.717	0.250	0.624	0.202	0.669	0.214	0.734	0.206	0.817	0.278	0.762	0.217	0.783
2.5	0.161	0.392	0.199	0.537	0.183	0.673	0.192	0.566	0.154	0.610	0.160	0.689	0.156	0.769	0.220	0.720	0.166	0.744
2.75	0.123	0.332	0.156	0.486	0.144	0.631	0.146	0.515	0.119	0.558	0.120	0.648	0.118	0.721	0.175	0.680	0.129	0.706
3	0.097	0.283	0.125	0.443	0.116	0.590	0.112	0.469	0.092	0.508	0.090	0.608	0.091	0.673	0.138	0.640	0.103	0.669
3.25	0.079	0.242	0.102	0.403	0.095	0.553	0.087	0.425	0.070	0.462	0.067	0.570	0.073	0.626	0.107	0.602	0.081	0.634
3.5	0.065	0.209	0.084	0.365	0.080	0.520	0.068	0.385	0.054	0.419	0.049	0.534	0.059	0.582	0.085	0.566	0.064	0.601
3.75	0.054	0.184	0.070	0.332	0.068	0.490	0.054	0.348	0.043	0.380	0.036	0.501	0.049	0.539	0.067	0.532	0.051	0.570
4	0.044	0.161	0.059	0.302	0.058	0.460	0.043	0.313	0.035	0.347	0.027	0.471	0.041	0.499	0.055	0.499	0.040	0.540
4.25	0.037	0.140	0.049	0.276	0.049	0.433	0.034	0.283	0.029	0.317	0.020	0.444	0.035	0.461	0.046	0.469	0.032	0.512
4.5	0.030	0.120	0.041	0.254	0.042	0.408	0.026	0.256	0.024	0.289	0.016	0.418	0.030	0.426	0.039	0.441	0.026	0.486
4.75	0.025	0.105	0.034	0.232	0.035	0.386	0.021	0.232	0.020	0.264	0.012	0.395	0.026	0.392	0.033	0.415	0.022	0.463
5	0.021	0.091	0.028	0.214	0.030	0.365	0.017	0.210	0.017	0.241	0.010	0.374	0.023	0.363	0.029	0.390	0.018	0.441
5.25	0.018	0.079	0.024	0.196	0.026	0.346	0.014	0.191	0.014	0.220	0.008	0.354	0.020	0.335	0.026	0.368	0.015	0.419
5.5	0.016	0.068	0.020	0.182	0.023	0.328	0.012	0.173	0.012	0.202	0.007	0.335	0.017	0.308	0.023	0.347	0.012	0.400
5.75	0.014	0.058	0.017	0.167	0.020	0.312	0.010	0.158	0.010	0.185	0.005	0.318	0.015	0.286	0.020	0.327	0.010	0.382
6	0.012	0.052	0.014	0.155	0.018	0.297	0.009	0.144	0.009	0.170	0.005	0.301	0.013	0.264	0.018	0.309	0.009	0.364
6.25	0.011	0.046	0.012	0.142	0.016	0.284	0.008	0.131	0.008	0.156	0.004	0.286	0.012	0.244	0.016	0.293	0.007	0.349
6.5	0.010	0.042	0.010	0.132	0.015	0.272	0.007	0.120	0.007	0.144	0.003	0.272	0.010	0.225	0.013	0.277	0.006	0.334
6.75	0.009	0.037	0.008	0.122	0.014	0.260	0.006	0.110	0.006	0.133	0.003	0.259	0.009	0.207	0.011	0.262	0.005	0.319
7	0.008	0.034	0.007	0.113	0.012	0.249	0.005	0.100	0.005	0.123	0.002	0.247	0.007	0.191	0.010	0.248	0.005	0.306
7.25	0.007	0.031	0.006	0.105	0.011	0.238	0.004	0.092	0.005	0.114	0.002	0.235	0.006	0.176	0.008	0.235	0.004	0.292
7.5	0.006	0.027	0.005	0.097	0.010	0.229	0.003	0.083	0.004	0.106	0.002	0.225	0.005	0.163	0.006	0.223	0.004	0.280
7.75	0.005	0.025	0.004	0.091	0.009	0.221	0.003	0.076	0.004	0.098	0.001	0.215	0.005	0.151	0.005	0.212	0.003	0.268
8	0.005	0.023	0.004	0.085	0.008	0.212	0.002	0.069	0.004	0.090	0.001	0.206	0.004	0.140	0.004	0.201	0.003	0.258
8.25	0.004	0.021	0.003	0.079	0.007	0.205	0.002	0.063	0.003	0.084	0.001	0.197	0.003	0.130	0.003	0.191	0.003	0.248
8.5	0.004	0.019	0.003	0.074	0.006	0.197	0.002	0.057	0.003	0.078	0.001	0.189	0.003	0.121	0.003	0.181	0.002	0.239
8.75	0.003	0.018	0.002	0.069	0.005	0.190	0.001	0.053	0.002	0.073	0.001	0.181	0.002	0.113	0.002	0.171	0.002	0.230
9	0.003	0.016	0.002	0.064	0.004	0.183	0.001	0.048	0.002	0.068	0.001	0.173	0.002	0.105	0.002	0.163	0.002	0.222
9.25	0.002	0.015	0.002	0.060	0.003	0.176	0.001	0.044	0.002	0.063	0.001	0.166	0.002	0.098	0.002	0.156	0.001	0.213
9.5	0.002	0.014	0.002	0.057	0.003	0.170	0.001	0.040	0.001	0.059	0.001	0.159	0.002	0.091	0.001	0.148	0.001	0.206
9.75	0.002	0.013	0.002	0.053	0.002	0.164	0.001	0.037	0.001	0.055	0.001	0.153	0.001	0.085	0.001	0.141	0.001	0.199
10	0.002	0.011	0.001	0.050	0.002	0.158	0.001	0.034	0.001	0.052	0.000	0.148	0.001	0.080	0.001	0.134	0.001	0.193

Table 10: FCx results with 300 differentially expressed genes.

FC	Group 1		Group 2		Group 3		Group 4		Group 5		Group 6		Group 7		Group 8		Group 9	
	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP
2	0.313	0.537	0.319	0.655	0.329	0.759	0.313	0.691	0.276	0.731	0.311	0.776	0.318	0.839	0.335	0.812	0.303	0.820
2.25	0.238	0.462	0.243	0.596	0.254	0.713	0.235	0.631	0.208	0.671	0.236	0.732	0.246	0.794	0.260	0.768	0.231	0.779
2.5	0.183	0.398	0.187	0.545	0.198	0.669	0.180	0.572	0.157	0.613	0.180	0.689	0.192	0.747	0.202	0.725	0.176	0.740
2.75	0.142	0.344	0.144	0.497	0.155	0.626	0.141	0.517	0.118	0.558	0.138	0.647	0.151	0.702	0.155	0.684	0.133	0.702
3	0.110	0.300	0.111	0.453	0.121	0.589	0.112	0.465	0.089	0.506	0.105	0.608	0.120	0.658	0.119	0.643	0.102	0.665
3.25	0.086	0.259	0.086	0.413	0.094	0.554	0.089	0.418	0.067	0.460	0.080	0.571	0.096	0.617	0.092	0.604	0.078	0.630
3.5	0.067	0.225	0.070	0.374	0.073	0.522	0.072	0.377	0.051	0.418	0.060	0.538	0.078	0.577	0.073	0.568	0.060	0.598
3.75	0.053	0.197	0.058	0.338	0.058	0.490	0.060	0.339	0.039	0.380	0.045	0.506	0.064	0.537	0.057	0.533	0.047	0.567
4	0.042	0.173	0.049	0.305	0.045	0.464	0.050	0.307	0.031	0.346	0.033	0.476	0.053	0.500	0.045	0.501	0.038	0.538
4.25	0.033	0.151	0.042	0.276	0.034	0.437	0.041	0.278	0.025	0.315	0.024	0.449	0.043	0.467	0.034	0.471	0.031	0.511
4.5	0.027	0.132	0.036	0.252	0.026	0.412	0.034	0.253	0.020	0.288	0.018	0.424	0.035	0.434	0.026	0.442	0.026	0.486
4.75	0.022	0.116	0.030	0.228	0.020	0.390	0.027	0.229	0.017	0.263	0.013	0.400	0.028	0.405	0.020	0.415	0.021	0.462
5	0.018	0.101	0.026	0.208	0.016	0.369	0.022	0.208	0.014	0.240	0.010	0.378	0.023	0.377	0.015	0.390	0.018	0.439
5.25	0.015	0.089	0.022	0.190	0.012	0.350	0.017	0.188	0.011	0.220	0.008	0.358	0.019	0.351	0.011	0.366	0.015	0.418
5.5	0.012	0.079	0.019	0.175	0.010	0.334	0.014	0.170	0.009	0.201	0.007	0.339	0.016	0.328	0.009	0.345	0.012	0.399
5.75	0.010	0.071	0.016	0.160	0.008	0.318	0.011	0.152	0.008	0.185	0.005	0.321	0.014	0.306	0.007	0.325	0.010	0.380
6	0.009	0.062	0.014	0.147	0.007	0.304	0.009	0.137	0.006	0.171	0.005	0.305	0.012	0.286	0.005	0.307	0.009	0.363
6.25	0.007	0.056	0.011	0.134	0.005	0.290	0.007	0.124	0.005	0.157	0.004	0.290	0.010	0.266	0.004	0.289	0.007	0.348
6.5	0.006	0.049	0.009	0.123	0.005	0.276	0.005	0.114	0.004	0.145	0.003	0.276	0.009	0.248	0.004	0.272	0.006	0.333
6.75	0.005	0.044	0.008	0.113	0.004	0.263	0.004	0.104	0.003	0.133	0.003	0.263	0.008	0.231	0.003	0.257	0.005	0.319
7	0.004	0.039	0.006	0.104	0.003	0.252	0.003	0.095	0.003	0.123	0.002	0.252	0.007	0.215	0.003	0.243	0.004	0.305
7.25	0.004	0.035	0.005	0.096	0.003	0.242	0.003	0.087	0.002	0.113	0.002	0.240	0.006	0.200	0.002	0.230	0.004	0.293
7.5	0.003	0.031	0.004	0.088	0.003	0.232	0.002	0.079	0.002	0.105	0.002	0.229	0.006	0.187	0.002	0.218	0.003	0.281
7.75	0.003	0.027	0.004	0.082	0.002	0.222	0.002	0.073	0.002	0.097	0.002	0.219	0.005	0.175	0.002	0.206	0.003	0.270
8	0.002	0.024	0.003	0.076	0.002	0.213	0.002	0.068	0.001	0.090	0.001	0.210	0.005	0.163	0.002	0.195	0.002	0.260
8.25	0.002	0.021	0.003	0.071	0.002	0.205	0.002	0.063	0.001	0.083	0.001	0.201	0.005	0.153	0.001	0.185	0.002	0.250
8.5	0.002	0.019	0.002	0.066	0.001	0.197	0.001	0.058	0.001	0.077	0.001	0.193	0.004	0.144	0.001	0.175	0.002	0.240
8.75	0.002	0.017	0.002	0.061	0.001	0.189	0.001	0.054	0.001	0.073	0.001	0.185	0.004	0.135	0.001	0.167	0.002	0.231
9	0.001	0.015	0.002	0.057	0.001	0.182	0.001	0.050	0.001	0.067	0.001	0.178	0.004	0.127	0.001	0.158	0.001	0.223
9.25	0.001	0.013	0.002	0.053	0.001	0.175	0.001	0.046	0.001	0.063	0.001	0.171	0.004	0.119	0.001	0.150	0.001	0.215
9.5	0.001	0.012	0.001	0.050	0.001	0.169	0.001	0.043	0.001	0.059	0.001	0.164	0.004	0.112	0.001	0.142	0.001	0.208
9.75	0.001	0.011	0.001	0.047	0.001	0.163	0.001	0.040	0.001	0.055	0.001	0.158	0.004	0.105	0.001	0.135	0.001	0.201
10	0.001	0.010	0.001	0.044	0.001	0.158	0.001	0.037	0.001	0.052	0.001	0.153	0.003	0.099	0.001	0.128	0.001	0.193

Table 11: ISER results with 100 differentially expressed genes, raw x loess-normalized data.

	p	Group 1		Group 2		Group 3		Group 4		Group 5		Group 6		Group 7		Group 8		Group 9	
		FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP
R a w d a t a	1%	0.032	0.873	0.029	0.930	0.026	0.945	0.018	0.940	0.022	0.943	0.025	0.955	0.018	0.980	0.018	0.969	0.021	0.966
	2%	0.049	0.894	0.047	0.946	0.044	0.956	0.035	0.952	0.039	0.955	0.043	0.961	0.035	0.986	0.035	0.976	0.039	0.972
	3%	0.066	0.906	0.065	0.954	0.062	0.963	0.052	0.960	0.056	0.961	0.061	0.967	0.052	0.988	0.052	0.979	0.056	0.975
	4%	0.084	0.915	0.083	0.960	0.081	0.967	0.070	0.965	0.074	0.966	0.079	0.971	0.070	0.989	0.070	0.982	0.074	0.979
	5%	0.101	0.923	0.101	0.964	0.099	0.969	0.088	0.968	0.092	0.969	0.097	0.974	0.088	0.991	0.087	0.984	0.092	0.981
	6%	0.119	0.928	0.119	0.968	0.118	0.971	0.106	0.971	0.110	0.971	0.116	0.977	0.106	0.992	0.106	0.985	0.110	0.982
	7%	0.137	0.933	0.138	0.971	0.136	0.973	0.124	0.973	0.129	0.974	0.134	0.980	0.124	0.993	0.124	0.986	0.128	0.983
	8%	0.155	0.939	0.156	0.974	0.155	0.976	0.143	0.975	0.147	0.976	0.153	0.981	0.143	0.993	0.142	0.987	0.147	0.984
	9%	0.174	0.941	0.174	0.975	0.174	0.978	0.162	0.977	0.165	0.979	0.171	0.982	0.161	0.993	0.161	0.987	0.165	0.986
	10%	0.193	0.945	0.193	0.975	0.193	0.981	0.180	0.979	0.184	0.980	0.190	0.983	0.180	0.994	0.180	0.988	0.184	0.987
	11%	0.211	0.949	0.212	0.977	0.212	0.982	0.200	0.980	0.203	0.982	0.209	0.984	0.200	0.994	0.199	0.989	0.203	0.988
	12%	0.230	0.953	0.231	0.979	0.231	0.983	0.219	0.982	0.222	0.983	0.228	0.985	0.219	0.995	0.218	0.989	0.222	0.988
	13%	0.249	0.956	0.250	0.980	0.251	0.984	0.239	0.983	0.241	0.984	0.247	0.986	0.238	0.995	0.237	0.989	0.242	0.989
	14%	0.269	0.959	0.270	0.981	0.270	0.984	0.258	0.983	0.261	0.984	0.266	0.987	0.257	0.996	0.257	0.990	0.261	0.989
	15%	0.288	0.961	0.289	0.982	0.289	0.985	0.276	0.984	0.281	0.986	0.285	0.988	0.277	0.996	0.276	0.990	0.280	0.990
	16%	0.307	0.962	0.308	0.983	0.309	0.986	0.296	0.985	0.300	0.987	0.305	0.988	0.296	0.996	0.296	0.991	0.300	0.990
	17%	0.327	0.964	0.328	0.983	0.328	0.986	0.316	0.985	0.319	0.987	0.325	0.988	0.316	0.996	0.316	0.991	0.319	0.991
	18%	0.346	0.964	0.348	0.983	0.348	0.987	0.335	0.986	0.339	0.988	0.344	0.988	0.336	0.997	0.336	0.991	0.339	0.991
	19%	0.366	0.967	0.367	0.984	0.368	0.988	0.355	0.986	0.359	0.988	0.364	0.989	0.356	0.997	0.356	0.992	0.359	0.992
	20%	0.387	0.968	0.387	0.986	0.388	0.990	0.375	0.988	0.379	0.989	0.383	0.990	0.376	0.997	0.376	0.992	0.379	0.992
N o r m a l i z e d d a t a	1%	0.032	0.872	0.029	0.931	0.026	0.945	0.018	0.941	0.022	0.943	0.026	0.954	0.018	0.981	0.018	0.969	0.021	0.966
	2%	0.049	0.897	0.047	0.947	0.045	0.957	0.035	0.953	0.039	0.955	0.043	0.962	0.035	0.986	0.035	0.976	0.039	0.973
	3%	0.066	0.908	0.065	0.955	0.063	0.963	0.053	0.960	0.057	0.961	0.061	0.968	0.052	0.988	0.052	0.979	0.057	0.976
	4%	0.084	0.916	0.083	0.961	0.081	0.967	0.070	0.965	0.075	0.966	0.080	0.971	0.070	0.990	0.070	0.982	0.074	0.979
	5%	0.102	0.922	0.101	0.964	0.099	0.969	0.089	0.969	0.092	0.968	0.098	0.974	0.088	0.991	0.087	0.984	0.093	0.981
	6%	0.120	0.930	0.120	0.968	0.118	0.970	0.107	0.971	0.110	0.972	0.116	0.977	0.106	0.992	0.106	0.985	0.111	0.982
	7%	0.138	0.934	0.138	0.971	0.137	0.974	0.125	0.974	0.129	0.974	0.135	0.979	0.125	0.992	0.124	0.986	0.129	0.983
	8%	0.157	0.939	0.156	0.973	0.155	0.977	0.144	0.975	0.148	0.976	0.153	0.981	0.143	0.993	0.142	0.987	0.147	0.984
	9%	0.175	0.943	0.175	0.975	0.174	0.979	0.163	0.977	0.166	0.978	0.172	0.982	0.162	0.993	0.161	0.988	0.166	0.986
	10%	0.194	0.946	0.194	0.976	0.194	0.980	0.182	0.979	0.185	0.980	0.191	0.983	0.181	0.994	0.180	0.988	0.185	0.987
	11%	0.213	0.950	0.213	0.977	0.213	0.982	0.201	0.981	0.204	0.981	0.210	0.983	0.200	0.995	0.199	0.989	0.204	0.988
	12%	0.231	0.953	0.232	0.978	0.232	0.983	0.220	0.982	0.223	0.983	0.228	0.985	0.219	0.995	0.218	0.989	0.223	0.988
	13%	0.250	0.956	0.251	0.979	0.251	0.984	0.240	0.983	0.242	0.984	0.247	0.985	0.239	0.995	0.238	0.989	0.242	0.989
	14%	0.270	0.958	0.270	0.980	0.270	0.985	0.259	0.984	0.262	0.984	0.267	0.988	0.258	0.996	0.258	0.990	0.262	0.989
	15%	0.289	0.960	0.290	0.981	0.290	0.986	0.278	0.984	0.282	0.985	0.286	0.988	0.277	0.996	0.277	0.990	0.281	0.990
	16%	0.308	0.962	0.309	0.982	0.309	0.986	0.297	0.985	0.301	0.986	0.305	0.988	0.297	0.996	0.297	0.991	0.301	0.990
	17%	0.328	0.964	0.329	0.983	0.329	0.986	0.317	0.986	0.320	0.987	0.325	0.988	0.317	0.997	0.317	0.991	0.320	0.990
	18%	0.347	0.965	0.348	0.983	0.349	0.987	0.336	0.986	0.340	0.988	0.345	0.988	0.337	0.997	0.337	0.991	0.340	0.991
	19%	0.367	0.966	0.368	0.985	0.368	0.988	0.356	0.987	0.360	0.988	0.365	0.989	0.357	0.997	0.357	0.992	0.359	0.992
	20%	0.387	0.968	0.388	0.986	0.388	0.989	0.376	0.987	0.380	0.989	0.385	0.990	0.377	0.997	0.377	0.992	0.380	0.992

Table 12: ISER results with 200 differentially expressed genes, raw x loess-normalized data.

	p	Group 1		Group 2		Group 3		Group 4		Group 5		Group 6		Group 7		Group 8		Group 9	
		FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP
R a w d a t a	1%	0.042	0.855	0.038	0.919	0.032	0.949	0.014	0.941	0.022	0.942	0.029	0.950	0.014	0.976	0.015	0.968	0.021	0.965
	2%	0.058	0.882	0.054	0.933	0.049	0.959	0.028	0.954	0.037	0.954	0.044	0.961	0.028	0.982	0.030	0.974	0.036	0.973
	3%	0.073	0.898	0.070	0.940	0.066	0.964	0.043	0.961	0.053	0.961	0.060	0.966	0.043	0.985	0.045	0.979	0.052	0.977
	4%	0.089	0.911	0.086	0.946	0.083	0.968	0.059	0.966	0.069	0.965	0.077	0.971	0.059	0.987	0.060	0.981	0.067	0.979
	5%	0.105	0.917	0.103	0.952	0.101	0.971	0.075	0.969	0.085	0.968	0.093	0.973	0.075	0.989	0.077	0.984	0.084	0.981
	6%	0.122	0.924	0.120	0.956	0.118	0.973	0.092	0.972	0.101	0.970	0.110	0.975	0.092	0.990	0.094	0.986	0.100	0.983
	7%	0.139	0.929	0.137	0.960	0.136	0.975	0.109	0.974	0.119	0.973	0.127	0.978	0.109	0.991	0.111	0.987	0.117	0.985
	8%	0.156	0.935	0.154	0.962	0.154	0.977	0.126	0.976	0.136	0.975	0.145	0.979	0.127	0.992	0.129	0.988	0.135	0.985
	9%	0.174	0.939	0.172	0.965	0.172	0.978	0.144	0.977	0.154	0.977	0.163	0.980	0.144	0.993	0.147	0.988	0.153	0.986
	10%	0.192	0.944	0.190	0.967	0.190	0.979	0.162	0.978	0.172	0.979	0.181	0.981	0.163	0.994	0.165	0.989	0.170	0.987
	11%	0.210	0.946	0.208	0.970	0.209	0.981	0.180	0.979	0.190	0.980	0.199	0.982	0.181	0.994	0.183	0.990	0.189	0.988
	12%	0.228	0.950	0.227	0.972	0.227	0.982	0.199	0.980	0.208	0.981	0.218	0.983	0.199	0.995	0.202	0.991	0.207	0.989
	13%	0.246	0.953	0.245	0.973	0.246	0.984	0.217	0.982	0.227	0.982	0.236	0.984	0.219	0.995	0.221	0.991	0.226	0.990
	14%	0.264	0.955	0.264	0.974	0.265	0.984	0.237	0.982	0.246	0.983	0.256	0.985	0.237	0.995	0.239	0.992	0.245	0.991
	15%	0.283	0.957	0.282	0.976	0.284	0.985	0.256	0.983	0.265	0.984	0.275	0.986	0.257	0.996	0.258	0.992	0.265	0.991
	16%	0.301	0.959	0.301	0.977	0.303	0.986	0.276	0.984	0.284	0.985	0.294	0.987	0.277	0.996	0.278	0.993	0.284	0.992
	17%	0.321	0.960	0.320	0.978	0.323	0.987	0.295	0.985	0.304	0.986	0.313	0.988	0.296	0.996	0.298	0.993	0.303	0.992
	18%	0.340	0.962	0.339	0.980	0.343	0.987	0.315	0.985	0.324	0.986	0.333	0.988	0.316	0.996	0.317	0.994	0.323	0.992
	19%	0.359	0.964	0.359	0.981	0.362	0.988	0.335	0.986	0.344	0.987	0.352	0.989	0.336	0.996	0.337	0.994	0.342	0.993
	20%	0.379	0.964	0.379	0.981	0.382	0.988	0.356	0.986	0.364	0.988	0.372	0.989	0.357	0.996	0.357	0.994	0.362	0.993
N o r m a l i z e d d a t a	1%	0.042	0.856	0.038	0.917	0.032	0.949	0.014	0.941	0.022	0.942	0.029	0.950	0.014	0.975	0.015	0.967	0.021	0.965
	2%	0.058	0.883	0.054	0.934	0.049	0.958	0.028	0.954	0.037	0.954	0.045	0.960	0.028	0.981	0.030	0.974	0.036	0.973
	3%	0.073	0.900	0.070	0.941	0.066	0.964	0.043	0.961	0.053	0.961	0.060	0.967	0.044	0.985	0.045	0.978	0.052	0.977
	4%	0.089	0.911	0.087	0.947	0.083	0.969	0.059	0.966	0.069	0.965	0.077	0.971	0.059	0.988	0.061	0.981	0.068	0.979
	5%	0.106	0.918	0.104	0.953	0.101	0.971	0.076	0.970	0.085	0.968	0.093	0.973	0.076	0.989	0.077	0.983	0.084	0.981
	6%	0.123	0.924	0.120	0.957	0.119	0.973	0.092	0.972	0.102	0.970	0.110	0.975	0.093	0.991	0.094	0.985	0.100	0.983
	7%	0.140	0.931	0.138	0.960	0.136	0.975	0.110	0.974	0.119	0.973	0.127	0.978	0.110	0.991	0.111	0.987	0.118	0.984
	8%	0.157	0.936	0.155	0.963	0.154	0.976	0.127	0.976	0.136	0.976	0.145	0.979	0.127	0.992	0.129	0.987	0.135	0.985
	9%	0.175	0.940	0.173	0.965	0.172	0.978	0.145	0.977	0.154	0.977	0.163	0.980	0.145	0.993	0.147	0.988	0.153	0.986
	10%	0.192	0.944	0.191	0.968	0.191	0.979	0.162	0.978	0.172	0.978	0.181	0.981	0.163	0.994	0.165	0.989	0.171	0.987
	11%	0.211	0.947	0.209	0.970	0.209	0.981	0.181	0.979	0.190	0.980	0.199	0.982	0.181	0.994	0.184	0.990	0.189	0.988
	12%	0.228	0.950	0.227	0.972	0.228	0.982	0.199	0.981	0.209	0.981	0.218	0.983	0.200	0.995	0.203	0.991	0.208	0.989
	13%	0.247	0.953	0.246	0.973	0.247	0.984	0.218	0.982	0.227	0.982	0.237	0.984	0.219	0.995	0.221	0.991	0.227	0.990
	14%	0.265	0.956	0.264	0.975	0.266	0.985	0.237	0.983	0.246	0.983	0.256	0.985	0.238	0.995	0.240	0.992	0.246	0.990
	15%	0.283	0.958	0.283	0.976	0.285	0.985	0.257	0.983	0.265	0.984	0.275	0.986	0.258	0.996	0.259	0.992	0.265	0.991
	16%	0.302	0.959	0.302	0.978	0.304	0.986	0.276	0.984	0.285	0.985	0.294	0.987	0.278	0.996	0.279	0.993	0.284	0.992
	17%	0.322	0.961	0.321	0.979	0.323	0.987	0.296	0.985	0.304	0.986	0.313	0.988	0.297	0.996	0.299	0.993	0.304	0.992
	18%	0.341	0.962	0.340	0.980	0.343	0.987	0.316	0.985	0.324	0.987	0.333	0.988	0.317	0.996	0.318	0.993	0.323	0.992
	19%	0.360	0.963	0.360	0.981	0.363	0.987	0.336	0.986	0.344	0.987	0.352	0.989	0.337	0.996	0.338	0.994	0.343	0.993
	20%	0.379	0.965	0.380	0.981	0.383	0.988	0.356	0.987	0.364	0.988	0.372	0.989	0.358	0.996	0.358	0.994	0.363	0.993

Table 13: ISER results with 300 differentially expressed genes, raw x loess-normalized data.

	p	Group 1		Group 2		Group 3		Group 4		Group 5		Group 6		Group 7		Group 8		Group 9	
		FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP
Random	1%	0.053	0.853	0.046	0.918	0.035	0.947	0.011	0.923	0.023	0.930	0.032	0.950	0.012	0.921	0.013	0.915	0.021	0.944
	2%	0.067	0.884	0.061	0.934	0.052	0.956	0.024	0.941	0.036	0.944	0.047	0.960	0.024	0.935	0.025	0.934	0.035	0.957
	3%	0.081	0.898	0.076	0.942	0.068	0.963	0.037	0.951	0.050	0.953	0.061	0.967	0.037	0.945	0.039	0.945	0.048	0.965
	4%	0.095	0.910	0.091	0.948	0.084	0.967	0.051	0.959	0.064	0.959	0.076	0.970	0.051	0.953	0.053	0.953	0.062	0.970
	5%	0.110	0.919	0.107	0.954	0.101	0.970	0.066	0.964	0.079	0.963	0.091	0.973	0.066	0.959	0.067	0.960	0.077	0.975
	6%	0.125	0.926	0.122	0.958	0.117	0.972	0.081	0.967	0.094	0.966	0.107	0.976	0.081	0.966	0.083	0.964	0.093	0.978
	7%	0.140	0.932	0.138	0.962	0.134	0.974	0.097	0.970	0.110	0.969	0.124	0.978	0.097	0.970	0.099	0.969	0.109	0.980
	8%	0.156	0.936	0.155	0.964	0.151	0.976	0.114	0.973	0.126	0.972	0.140	0.980	0.114	0.975	0.115	0.973	0.126	0.982
	9%	0.172	0.941	0.171	0.967	0.169	0.978	0.130	0.975	0.143	0.973	0.157	0.981	0.131	0.979	0.132	0.976	0.142	0.983
	10%	0.189	0.945	0.188	0.969	0.187	0.980	0.147	0.976	0.160	0.975	0.174	0.982	0.148	0.981	0.149	0.979	0.159	0.985
	11%	0.206	0.948	0.206	0.971	0.205	0.980	0.165	0.978	0.178	0.977	0.192	0.983	0.165	0.984	0.167	0.982	0.176	0.986
	12%	0.223	0.951	0.223	0.972	0.223	0.982	0.182	0.979	0.195	0.978	0.210	0.984	0.183	0.986	0.185	0.984	0.193	0.987
	13%	0.241	0.954	0.241	0.974	0.242	0.983	0.201	0.981	0.214	0.980	0.228	0.985	0.201	0.989	0.202	0.986	0.212	0.988
	14%	0.258	0.957	0.259	0.975	0.261	0.983	0.219	0.982	0.232	0.981	0.246	0.986	0.220	0.991	0.221	0.987	0.230	0.989
	15%	0.276	0.958	0.277	0.977	0.279	0.984	0.238	0.983	0.251	0.981	0.265	0.987	0.239	0.993	0.240	0.989	0.249	0.990
	16%	0.294	0.960	0.296	0.978	0.298	0.985	0.257	0.984	0.270	0.982	0.284	0.988	0.258	0.994	0.259	0.990	0.267	0.990
	17%	0.312	0.962	0.314	0.979	0.318	0.986	0.276	0.985	0.289	0.983	0.303	0.989	0.277	0.995	0.278	0.991	0.287	0.991
	18%	0.331	0.964	0.333	0.981	0.337	0.986	0.296	0.985	0.308	0.984	0.322	0.989	0.296	0.996	0.298	0.992	0.306	0.992
	19%	0.351	0.965	0.353	0.983	0.357	0.987	0.317	0.986	0.328	0.985	0.341	0.990	0.316	0.996	0.318	0.993	0.326	0.992
	20%	0.370	0.966	0.372	0.983	0.376	0.987	0.337	0.987	0.348	0.986	0.361	0.990	0.336	0.997	0.338	0.993	0.346	0.992
Normal	1%	0.053	0.854	0.046	0.918	0.036	0.947	0.011	0.923	0.023	0.929	0.032	0.950	0.012	0.922	0.013	0.912	0.021	0.943
	2%	0.067	0.883	0.061	0.934	0.052	0.956	0.024	0.941	0.036	0.944	0.047	0.959	0.024	0.937	0.025	0.931	0.035	0.957
	3%	0.082	0.899	0.076	0.942	0.068	0.963	0.037	0.952	0.050	0.953	0.062	0.966	0.037	0.947	0.039	0.942	0.048	0.965
	4%	0.096	0.911	0.091	0.948	0.085	0.966	0.052	0.959	0.065	0.959	0.077	0.970	0.051	0.955	0.053	0.952	0.062	0.970
	5%	0.110	0.919	0.107	0.953	0.101	0.970	0.067	0.964	0.079	0.963	0.092	0.973	0.066	0.961	0.068	0.959	0.078	0.974
	6%	0.125	0.926	0.122	0.958	0.117	0.972	0.082	0.967	0.094	0.967	0.108	0.976	0.082	0.967	0.083	0.964	0.093	0.977
	7%	0.141	0.933	0.139	0.962	0.134	0.975	0.097	0.970	0.110	0.969	0.124	0.978	0.098	0.972	0.099	0.969	0.109	0.980
	8%	0.157	0.938	0.155	0.965	0.152	0.976	0.114	0.973	0.127	0.971	0.141	0.980	0.114	0.976	0.115	0.973	0.125	0.981
	9%	0.173	0.941	0.172	0.967	0.170	0.978	0.130	0.975	0.143	0.973	0.158	0.981	0.131	0.980	0.132	0.976	0.142	0.983
	10%	0.189	0.945	0.189	0.969	0.188	0.980	0.148	0.977	0.160	0.975	0.175	0.982	0.148	0.983	0.150	0.979	0.159	0.985
	11%	0.206	0.948	0.206	0.971	0.206	0.981	0.165	0.978	0.178	0.977	0.193	0.983	0.166	0.986	0.167	0.982	0.176	0.986
	12%	0.224	0.952	0.224	0.972	0.224	0.982	0.183	0.980	0.196	0.979	0.210	0.984	0.184	0.988	0.185	0.984	0.194	0.987
	13%	0.242	0.954	0.242	0.974	0.242	0.983	0.201	0.981	0.214	0.980	0.229	0.985	0.202	0.990	0.202	0.986	0.211	0.988
	14%	0.259	0.956	0.259	0.975	0.261	0.983	0.220	0.982	0.233	0.981	0.246	0.986	0.220	0.992	0.221	0.987	0.230	0.989
	15%	0.277	0.959	0.277	0.977	0.280	0.984	0.239	0.983	0.251	0.981	0.265	0.987	0.239	0.993	0.240	0.989	0.249	0.990
	16%	0.295	0.961	0.296	0.978	0.299	0.985	0.258	0.984	0.270	0.982	0.284	0.988	0.258	0.994	0.259	0.990	0.268	0.990
	17%	0.313	0.962	0.315	0.980	0.318	0.986	0.277	0.985	0.289	0.983	0.303	0.989	0.277	0.995	0.279	0.991	0.287	0.991
	18%	0.332	0.964	0.334	0.982	0.338	0.986	0.297	0.985	0.309	0.984	0.322	0.989	0.297	0.996	0.298	0.992	0.307	0.992
	19%	0.352	0.965	0.353	0.983	0.357	0.987	0.317	0.986	0.328	0.985	0.342	0.990	0.317	0.996	0.318	0.993	0.326	0.992
	20%	0.371	0.966	0.373	0.983	0.377	0.987	0.338	0.987	0.348	0.986	0.361	0.990	0.337	0.997	0.338	0.993	0.346	0.992

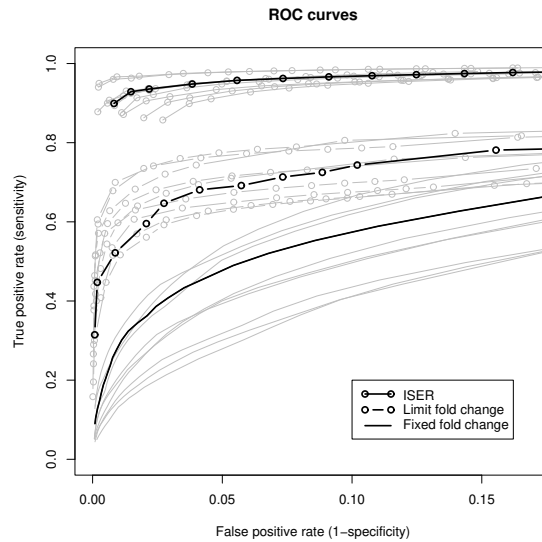


Figure 1: Receiver operating characteristic (ROC) curves for the three methods used to analyze simulated data sets. Each grey line is the mean curve for a subset of data sets with the same specifications. For each method of analysis, the black line is the median of the grey lines.

Conclusions

The algorithm showed a good performance on simulated data and correctly selected almost 90% of the differentially expressed genes. Its performance was superior to both of the other methods of analysis, and was unaffected by the normalization adopted. The simulation results proved that ISER is a valuable tool for identifying differentially expressed genes in array experiments.

References

- Kerr, M.K., Martin, M. and Churchill, G.A. (2000) Analysis of variance for gene expression microarray data. *J. Comp. Biol.* **7**, 819-837.
- Mutch, D.M., Berger, A., Mansourian, R., Rytz, A. and Roberts, M.A. (2002) The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data. *BMC Bioinformatics* **3**, 17.

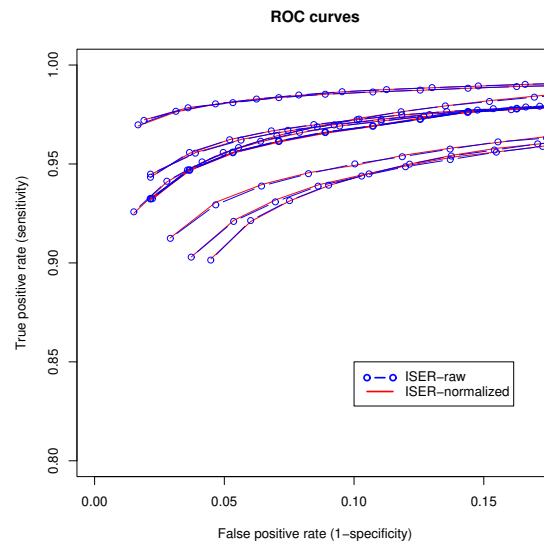


Figure 2: Receiver operating characteristic (ROC) curves for the comparison of ISER performance for raw and normalized data. Each line is the mean curve for a subset of data sets with the same specifications. The superposition of red and blue lines indicates that there is no difference on ISER's performance when dealing with raw or normalized data.

ISER: Analysis of a real macroarray data set

A macroarray data set was also used to test the performance of ISER. The experimental data of Nogueira *et al.* (2003), which contain the expression profile of sugarcane plantlet leaves exposed to cold, were downloaded from the GEO datasets of the National Center for Biotechnology Information (NCBI, <http://www.ncbi.nlm.nih.gov/>) and analyzed with the algorithm.

The experimental data consisted of two replicate sets, each one containing the relative expression profiles of 1536 sugarcane genes after five different periods (0, 3, 6, 12, 24 and 48 h) of plant exposure to low temperature (4°C). The cDNA of these genes was fixed on nylon arrays (two spots for each gene). After hybridizing each array with radioactively marked probes, the radioactive signal from each spot was quantified. The numbers entered into ISER were the averages of the signals obtained from the replicate spots corresponding to each gene. ISER was used to compare the time points (3, 6, 12, 24 and 48 h) with the control (0 h), in both experimental replicates. For each time point, genes selected by the algorithm in both replicates were accepted as differentially expressed.

The original data analysis consisted of calculating the expression ratios (treatment/control) and their logarithms for each time point and each replicate. Then, since these logratios were assumed to be normally distributed, their mean and standard deviation were calculated and the genes that showed, in both replicates, a logratio more than 1.65 standard deviations distant from the mean and a ratio ≥ 2 (respectively ≤ 0.5 for repressed genes) were considered to be differentially expressed. This approach does not consider the dependence of data variability on the average signal intensity of each gene, thus it may select more false-positive genes than ISER. Of the 59 genes originally selected by Nogueira *et al.* (2003), 46 (78%) were also selected by ISER. The other 13 genes not selected by ISER could be false-positives. In fact, the majority (9) of these genes were selected as induced at the time point of 48 hours, although in one of the experimental replicates for this time point, their intensities were contained in a region where ISER detected a greater variability in the data, thus not selecting them. However, among this 9 genes originally selected as induced at 48h and discarded by ISER, 4 showed similarity with rice polyubiquitins. Since another ubiquitin is also selected as induced by cold (both by ISER and by the original data analysis), these genes could be true positives that ISER could not select due to the greater variability of the data in one of the experimental replicates for 48h. Although, this is not a major drawback of ISER, since losses can always occur when experiments are done with few replicates.

On the other hand, ISER selected an additional set of 147 genes as differentially expressed in at least one time point (confirmed by both experiments), and the differential expression of 62 of these genes was also confirmed in two or more time points. Several of these genes are of biological relevance, including transcription factors, genes related to cellular communication (transmembrane kinases), stress response (heat shock protein and others) and protein metabolism. These genes could complement the original results of Nogueira *et al.* (2003), providing further characterization of the sugarcane response to low temperatures. For example, Nogueira *et al.* reported the repression of a heat shock protein (HSP82) in response to cold, and ISER also identified another heat shock protein (HSP90) that was also repressed by the cold treatment and a putative heat stress transcription factor that was differentially expressed during the experiment. All of the genes selected by ISER as being differentially expressed, including the 46 that were also identified by Nogueira *et al.* (2003), are shown on tables 1 and 2. The latter authors also confirmed the differential expression of five genes by RNA blots. Of these, four (SCMCCL6027D02, SCCCHR1003H01, SCMCCL6027G02 and SCCCHR1004E10) were also selected by ISER, and the fifth one (SCCCAD1002H02) showed

a moderate induction that were not confirmed by ISER on both experimental replicates.

Although ISER is designed to deal with data from single replicated experiments, comparing two RNA samples, it can be used when more samples are compared or replicates are present. In this case, the algorithm is applied separately to the comparison of each pair of samples and to each replicate, thus selecting genes that, for any pair of samples, are identified as differentially expressed on all replicates or on a large (pre-defined) proportion of them, as were done with the data set from Nogueira *et al.* (2003).

References

- Altschul,S.F., Madden,T.L., Schaffer,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**,3389-402.
- Nogueira,F.T., De Rosa,V.E.Jr., Menossi,M., Ulian,E.C. and Arruda,P. (2003) RNA expression profiles and data mining of sugarcane response to low temperatures. *Plant Physiol.* **132**, 1811-1824.

Table 1: Cold-Induced genes

CLONE-ID	GB acc	Blast hit ^a	E-value	Description ^b	Ratios	Treatment/Control (0h) ^c	3h	6h	12h	24h	48h
SCCCHR1003E12 ^d	BU103480	BAD33561.1	7.00E+66	ABI3interacting protein 2, AIP2	2.12	1.36	1.08	1.03	0.87		
SCCCAD1003B03 ^d	BU102661	P80607	0	Alpha1,4glucanprotein synthase	1.31	1.69	1.92	1.85	3.1		
SCMCCL6027C02 ^d	BU103119	XP_477093.1	0	cellulose synthase4	1.25	2.39	2.01	2.56	3.42		
SCEPCL6023E08 ^d	BU102929	XP_468806.1	0	cytosolic pyruvate orthophosphate dikinase	1.44	1.78	1.8	2.1	1.23		
SCCCAD1002F02 ^d	BU102621	BAD29526.1	5.00E+34	dehydrationresponsive family proteinlike	0.97	1.1	1.64	2.24	2.08		
SCCCHR1003A12 ^d	BU103443	XP_469762.1	4.00E+22	expressed protein	1.35	2.01	2.29	2.6	2.48		
SCCCHR1003D12 ^d	BU103471	AAM47584.1	3.00E+61	expressed protein	2.1	1.23	0.97	1.03	0.92		
SCMCCL6027G02 ^d	BU103672	AAT77085.1	1.00E+177	expressed protein	1.12	1.28	1.54	1.76	3.36		
SCCCHR1004E10 ^d	BU103552	XP_479156.1	4.00E+67	ids4 protein	0.98	1.58	2.86	2.43	3.87		
SCCCHR1003H07 ^d	BU103497	AAY42385.1	0	inorganic phosphate transporter 1	1.43	1.54	1.88	2.08	2.62		
SCMCCL6027D02 ^d	BU103128	AAW62955.1	1.00E+177	NAC23	2.12	3.1	3.97	5.86	8.26		
SCCCHR1004F11 ^d	BU103559	XP_470088.1	8.00E+62	NACdomain protein	1.17	1.67	3.44	3.16	4.26		
SCRLCL6032B05 ^d	BU103181	NP_912420.1	4.00E+24	NAM (no apical meristem) protein	0.94	1	1.98	2.13	2.58		
SCEPCL6023H04 ^d	BU102958	AAQ88396.1	0	nonphotosynthetic NADPmalic enzyme	1.57	2.04	2.13	1.2	1.09		
SCEPCL6021E11 ^d	BU103667	CAA44607.1	5.00E+56	ocsbinding factor 1	1.18	1.83	1.62	1.98	2.54		
SCEPCL6029G10 ^d	BU103091	CAA66667.1	1.00E+131	polyubiquitin	1.08	1.36	2.05	1.94	3.4		
SCCCLR1068H03 ^d	BU103658	AAC72195.1	0	pyruvate dehydrogenase E1 alpha subunit	1.97	1.52	1.14	2.12	2.03		
SCCCHR1004E11 ^d	BU103553	AAT81740.1	0	xanthine dehydrogenase,	1.09	1.63	2.13	2.23	2.76		
SCCCAD1003C07 ^d	BU102677	NP_916117.1	7.00E+19	WRKY transcription factor 13	0.83	1.44	1.73	2.17	2.44		
SCCCAD1001H05 ^d	BU103670	AJ309824.2*	0	25S rRNA	1.42	1.66	1.95	3.22	3.78		
SCCCHR1003G02 ^d	BU103669	AJ309824.2*	0	25S rRNA	1.48	1.49	2.23	3.35	3.97		
SCCCHR1003H01 ^d	BU103674	BT017638.1*	5.00E+75	EL01N0438F11.c mRNA sequence	3.82	4.03	8.51	15.29	29.77		
SCCCHR1002F05	BU103405	BAD88311.1	0	adenyl cyclaselike	1.09	1.16	1.6	1.69	1.86		
SCEPCL6028C08	BU102988	AAQ07252.2	2.00E+59	alkaline alpha galactosidase 2	1.3	1.32	1.47	2.04	2.06		
SCCCHR1002D06	BU103384	CAA44861.1	0	Alphatubulin 3	1.04	2.18	3.4	5.12	6.52		
SCCCAD1002H06	BU102644	XP_467694.1	1.00E+157	Avr9/C9 rapidly elicited protein 231	1.19	1.62	1.84	2.68	4.09		
SCCCHR1002H02	BU103426	AAK56130.1	6.00E+33	betaexpansin 7	1.69	1.59	0.97	1.18	1.11		
SCEPCL6023D03	BU102914	NP_909159.1	1.00E+24	calmodulinbinding protein like	1.01	1.58	2.42	2.16	2.78		
SCCCAD1003A06	BU102654	XP_467350.1	0	cytochrome P450	1.49	2.11	3.33	3.15	3.12		
SCCCHR1002H03	BU103427	AAP33506.2	3.00E+94	cytosolic 6phosphogluconate dehydrogenase	1.91	1.77	1.33	1.73	1.6		
SCCCAD1002A06	BU102576	NP_916649.1	1.00E+124	embryonic abundant proteinlike	1.19	1.51	1.4	1.52	1.96		
SCEPCL6028H09	BU103025	XP_466050.1	1.00E+134	heat stress transcription factor	1.45	0.91	2.09	1.35	2.28		
SCCCAD1001A02	BU102493	CAD42650.1	1.00E+127	heparanase	1.3	1.7	1.64	1.8	2.33		
SCUTCL6036G06	BU103268	AAT85280.1	0	homeobox leucinezipper protein,	0.74	1.05	1.21	1.69	1.34		
SCCCAD1001G07	BU102556	XP_475352.1	2.00E+36	hypothetical protein	1.54	1.75	2.13	2.29	2.16		
SCCCAD1002C02	BU102590	XP_467153.1	7.00E+26	hypothetical protein	1.64	0.79	0.82	0.93	0.93		
SCCCAD1002C05	BU102592	AAR91148.1	8.00E+54	hypothetical protein	1.16	1.35	3.39	1.75	1.9		
SCCCAD1003C09	BU102678	AAR91148.1	8.00E+54	hypothetical protein	0.49	1.01	1.7	0.83	0.85		
SCCCHR1004C04	BU103525	NP_174427.3	3.00E+38	leucinerich repeat transmembrane protein kinase	2	1.29	1.01	0.9	0.95		
SCCCLR1068G01	BU103646	XP_464408.1	0	leucinerich repeat transmembrane protein kinase	1.32	1.31	2.15	2.07	2.91		
SCCCHR1002H08	BU103432	Q43270	0	Phospholipase D alpha 1	1.32	1.63	1.35	3.95	4.91		
SCCCHR1004B01	BU103513	BAD38543.1	1.00E+105	PrMC3	4.17	1.5	0.98	0.85	0.96		
SCCCHR1003H02	BU103494	1.00E+125		Proteasome subunit alpha type 1	3.18	1.69	1.25	1.54	1.85		
SCCCLR1068D12	BU103623	AAK53813.1	2.00E+60	Ran binding protein1	1.61	1.21	1.16	1.07	0.74		
SCCCAD1001F04	BU102543	CAB80215.1	0	subtilisin proteinaselike	1.21	1.27	1.41	1.86	1.86		
SCCCAD1002C03	BU102591	BAD30897.1	1.00E+45	translation initiation factor 4G	2.57	2.37	1.17	1.25	1.16		
SCCCAD1002D02	BU102599	AAX95106.1	1.00E+176	Tub family,	1.71	1.3	1.59	1.62	1.58		
SCEPCL6021D02	BU102839	AAL17949.1	2.00E+89	type IIB calcium ATPase	1.5	1.82	1.08	0.95	1.19		
SCEPCL6023G06	BU102950	O82427	0	24methylsterol Cmethyltransferase 2	1.36	1.48	1.74	1.71	1.53		
SCEPCL6029D10	BU103067	XP_465282.1	0	26S proteasome regulatory particle tripleA ATPase	0.77	0.79	0.89	0.66	0.36		
SCCCHR1002H04	BU103428	NP_908322.1	4.00E+98	40S ribosomal protein S5	2.58	1.06	0.95	1.05	0.77		
SCMCCL6027H12	BU103169	AAT76426.1	0	53 exonuclease	Inf	Inf	Inf	Inf	Inf		
SCCCAD1001E04	BU102535	XP_474415.1	3.00E+29	OSJNBa0088H09.11	1.02	1.29	1.92	2.21	2.15		
SCCCAD1001H10	BU102568	XP_462684.1	2.00E+35	OSJNBa0093F12.14	1.49	1.49	1.86	1.8	2.19		
SCEPCL6021G03	BU102867	XP_474162.1	6.00E+20	OSJNBb0022F16.3	2.57	1.51	1.15	0.68	2.41		
SCCCHR1002G08	BU103420	XP_474954.1	8.00E+71	OSJNBb0072N21.11	1.11	1.38	1.29	1.53	1.89		
SCCCLR1068G12 ^d	BU103656	NP_909038.1	5.00E+25	unknown protein	1.3	1.17	2.12	1.83	1.67		
SCCCAD1002E02	BU102609	AAF01523.1	1.00E+27	unknown protein	1.06	1.09	1.32	1.79	1.53		
SCCCHR1004H04	BU103572	XP_478840.1	2.00E+34	unknown protein	1.97	1.43	1.17	1.37	1.47		
SCEPCL6021A08	BU102813	XP_467299.1	2.00E+11	unknown protein	1.28	1.73	1.86	1.94	3.26		
SCCCAD1001G06 ^d	BU103671	CAB96849.1	0	putative protein	1.09	1.18	1.98	1.86	2.47		
SCCCAD1001G12	BU102561		No hits ^e		5.76	9.39	11.74	8.53	5.79		
SCCCAD1003A02	BU102651		No hits ^e		1.01	1.88	1.67	1.79	2.13		
SCEPCL6029C11	BU103060		No hits ^e		0.83	1.03	1.52	2.29	2.84		
SCCCAD1001B12	f				1.22	1.6	1.81	3.17	3.25		
SCCCAD1001E10	f				1.03	1.17	1.64	1.51	1.31		
SCCCAD1001F03	f				1.07	1.5	1.71	2.47	2.24		
SCCCAD1001G02	f				1.02	1.15	1.37	2.05	1.66		
SCCCAD1002D01	f				0.94	0.95	1.9	3.47	2.93		
SCCCAD1002D06	f				0.83	1.01	1.82	2.31	2.09		
SCCCAD1002F10	f				1	1.37	2.15	1.95	2.31		
SCCCAD1002G06	f				1.02	1.3	2	2.06	2.34		
SCCCAD1002H09	f				1.05	1.48	1.73	2.81	2.79		
SCCCAD1003A04	f				1.02	1.7	1.84	3.41	2.89		
SCCCAD1003A12	f				1.07	1.83	1.98	3.43	3.89		
SCCCAD1003D11	f				1.18	1.29	2.14	3.21	3.01		
SCCCAD1003E05	f				0.98	1.21	1.99	2.69	2.71		
SCCCAD1003F07	f				0.91	1.74	2.32	2.77	3.03		
SCCCAD1003H04	f				1.26	1.92	2.34	2.02	2.97		
SCCCHR1001F03	f				1.18	1.21	1.99	2.24	1.84		
SCCCHR1001H02	f				1.67	1.3	1.07	0.95	0.75		
SCCCHR1001H10	f				2.04	1.27	1.53	1.53	1.68		
SCCCHR1003G12	f				1.04	1.07	1.78	1.86	1.93		
SCCCHR1004D12	f				1.1	1.35	2.13	2.6	2.86		
SCCCHR1004H05	f				1.6	1.87	1.36	1.25	1.36		
SCCCLR1068H02	f				1.82	1.06	0.99	1.58	1.25		
SCEPCL6023C12	f				1.58	1.52	1.38	1.62	1.4		
SCEPCL6029D08	f				1.02	1.05	1.77	2.65	1.92		
SCRLCL6032B07	f				1.13	0.97	1.92	2.72	2.17		
SCRLCL6032F07	f				0.88	1	1.27	2.26	1.6		
SCUTCL6036A11	f				NA	NA	Inf	NA	Inf		
SCCCHR1001F09	f				Inf	Inf	NA	Inf	NA		
SCCCHR1001F11	f				Inf	NA	Inf	Inf	Inf		

a Sequences were compared using the BLASTX algorithm (Altschul *et al.*, 1997), excepting those hits marked with an (*), which were found with BLASTN. **b** Putative function of gene products, according to its similar sequences. **c** Geometric means of expression ratios obtained from two independent experiments. **d** Genes also identified as cold-responsive by Nogueira *et al.* **e** Only Blast hits with e-value better than e-10 were considered. **f** No sequencing available. **Inf** Ratios with zero control signal from one of the replicates. **NA** Ratios with zero signals from both control and treated from one of the replicates.

Table 2: Cold-repressed genes

CLONE-ID	GB acc	Blast hit ^a	E-value	Description ^b	Ratios Treatment/Control (0h) ^c				
					3h	6h	12h	24h	48h
SCCCHR1001G10 ^d	BU103340	CAA45117.1	0	acetohydroxyacid synthase	0.88	0.77	0.97	0.47	0.38
SCAGCL6016D10 ^d	BU102762	XP_483191.1	0	heat shock protein 82	0.77	0.72	0.57	0.58	0.29
SCEPCL6023F09 ^d	BU102941	XP_483191.1	0	heat shock protein 82	0.86	0.8	0.61	0.57	0.32
SCRLCL6032B08 ^d	BU103183	AAV44023.1	0	long chain acylCoA synthetase	0.74	0.46	0.97	0.8	0.62
SCRLCL6032B11 ^d	BU103185	NP_912420.1	4.00E+24	NAM (no apical meristem) protein	0.68	0.39	0.68	0.78	0.6
SCCCHR1004G12 ^d	BU103568	AAK26758.1	1.00E+162	plasma membrane integral protein ZmPIP21	0.69	0.6	0.63	0.36	0.22
SCRLCL6032H12 ^d	BU103226	AAAX95167.1	0	receptorlike protein kinase	0.71	0.51	0.65	0.69	0.65
SCEPCL6029A12 ^d	BU103039	XP_470378.1	1.00E+134	sarcoplasmic reticulum protein (with alternative splicing)	0.45	0.33	1.17	0.53	0.38
SCRLCL6032F11 ^d	BU103210	NP_921084.1	1.00E+167	SCARECROW gene regulatorlike	0.69	0.49	0.67	0.64	0.52
SCEPCL6028A06 ^d	BU102970	XP_550348.1	0	stressinduced protein stl1	1.16	0.38	1.11	0.95	1.11
SCRLCL6032B09 ^d	BU103184	NP_919820.1	6.00E+89	thaumatinlike protein	1.1	0.36	1.03	1.44	1.04
SCRLCL6032G12	BU103219	A48892	9.00E+47	abscisic acidinduced protein HVA22	0.7	0.3	0.7	0.62	0.46
SCEPCL6023D07	BU102918	AAU93346.1	3.00E+16	actin	0.76	0.88	0.68	0.52	0.39
SCEPCL6029E01	BU103069	AAC34997.1	1.00E+119	alcohol dehydrogenase 1	0.57	0.62	0.44	0.37	0.36
SCRLCL6032C12	BU103192	CAA27681.1	0	alcohol dehydrogenase 1	0.58	0.51	0.46	0.51	0.37
SCAGCL6016A04	BU102726	XP_482917.1	0	alphaamylase isozyme 3D precursor	1.18	0.47	1.03	0.98	1.02
SCCCHR1001B02	BU103288	AAC24977.1	6.00E+11	anaerobically inducible early gene 2	0.85	0.7	0.61	0.45	0.36
SCEPCL6029F01	BU103077	AAK56131.1	1.00E+149	betaexpansin 8	0.69	0.77	0.73	0.56	0.48
SCRLCL6032E06	BU103202	XP_479316.1	1.00E+136	bZIP family transcription factor	0.55	0.54	0.88	0.71	0.49
SCMCCL6027A04	BU103100	NP_173179.1	6.00E+13	COP1interacting proteinrelated	1.14	0.37	1.05	0.98	1.1
SCRLCL6032D08	BU103198	BAD87339.1	5.00E+43	copper chaperone (CCH)related proteinlike	0.96	0.45	1.46	0.99	1.15
SCCCAD1002G02	BU102631	XP_479561.1	1.00E+116	DAZ associated protein 1	0.27	0	1.91	1.98	0.96
SCCCHR1002B10	BU103367	BAD72453.1	0	DNA helicase	0.63	0.87	0.67	0.84	0.75
SCRLCL6032H07	BU103224	AAV44069.1	0	DNAbinding protein GBP16	0.86	0.51	0.75	0.95	0.84
SCCCHR1002H11	BU103435	AAM23263.1	4.00E+32	DnaJlike protein	0.92	1.02	0.56	0.74	0.76
SCCCAD1003A01	BU102650	BAD53291.1	6.00E+57	frataxin	0.63	0.94	0.89	0.56	0.77
SCEPCL6029B05	BU103044	AAL68853.1	1.00E+170	gb protein	0.85	0.57	0.86	0.97	0.68
SCCCLR1068D10	BU103622	AAL79732.1	0	heat shock protein 90	0.64	0.64	0.93	0.62	0.35
SCEPCL6021G05	BU102869	AAL79732.1	0	heat shock protein 90	0.67	0.69	0.65	0.46	0.39
SCCCLR1068A10	BU103589	BAD61631.1	1.00E+177	HEN4	0.63	0.71	0.93	0.44	0.51
SCEPCL6021F02	BU102857	BAD35388.1	0	HGA4	0.91	0.97	0.58	0.52	0.35
SCRLCL6032H04	BU103222	XP_477826.1	2.00E+64	oxidation resistance llike protein	0.79	0.63	0.72	0.91	0.77
SCRLCL6032G08	BU103217	AAT40504.1	1.00E+101	polyprotein	0.88	0.51	0.82	0.67	0.58
SCCCHR1003H09	BU103499	XP_467209.1	1.00E+158	serine carboxypeptidase	1.43	0.92	0.49	0.41	0.41
SCRLCL6032G02	BU103212	NP_910702.1	6.00E+16	sphingosine kinase	1	0.27	0.73	0.89	0.95
SCRLCL6032H06	BU103223	XP_468503.1	0	splicing factor 3B subunit 2	2.53	0	1.58	0	1.59
SCAGCL6016C06	BU102750	AAT81711.1	1.00E+169	transcription factor, putative	0.86	0.51	0.9	0.79	0.69
SCMCCL6027A01	BU103098	Q9SWF9	5.00E+49	zinc finger protein	0	0.27	2.34	0	2.13
SCRLCL6032G05	BU103215	NP_922075.1	3.00E+29	zinc finger protein	0.9	0.43	0.89	1.2	0.83
SCMCCL6027H04	BU103162	XP_472294.1	2.00E+51	OSJNBa0053B21.14	0	0.64	1.14	1.1	0.91
SCCCHR1001F01	BU103328	XP_474592.1	1.00E+104	OSJNBb0115I21.11	0.94	0.18	0.43	0.37	0.27
SCCCHR1001B11 ^d	BU103294	AAU90205.1	7.00E+40	unknown protein	0.72	0.51	0.39	0.24	0.15
SCCCAD1004B05 ^d	BU103517	AAU90205.1	7.00E+40	unknown protein	0.74	0.47	0.42	0.26	0.14
SCEPCL6023A10 ^d	BU102893	BAD81221.1	8.00E+56	unknown protein	0.92	0.93	0.52	0.54	0.47
SCCCAD1002C12	BU102598	AAU44012.1	1.00E+167	unknown protein	0.95	0.77	0.62	0.53	0.6
SCCCHR1001E06	BU103321	XP_483821.1	1.00E+62	unknown protein	0.84	0.94	0.87	0.55	0.44
SCRLCL6032F08	BU103208	XP_476128.1	1.00E+60	unknown protein	0.92	0.44	0.89	1.12	0.68
SCRLCL6032H11	BU103225	XP_467299.1	2.00E+11	unknown protein	0.88	0.54	1.08	1.16	1.03
SCAGCL6016A10 ^d	BU102732		No hits ^e		0.42	0.91	0.91	0.69	0.62
SCCCAD1002A11 ^d	BU102579		No hits ^e		0.59	0.59	0.43	0.24	0.31
SCCCHR1001B05 ^d	BU103291		No hits ^e		0.59	0.38	0.35	0.19	0.21
SCRLCL6032D11 ^d	BU103199		No hits ^e		0.63	0.43	0.65	0.6	0.38
SCRLCL6032E11 ^d	BU103203		No hits ^e		0.58	0.41	0.69	0.8	0.46
SCRLCL6032G07 ^d	BU103216		No hits ^e		0.76	0.4	0.77	0.94	0.59
SCAGCL6016B06	BU102739		No hits ^e		0.92	0.53	0.85	0.94	0.56
SCCCHR1004A06	BU103506		No hits ^e		0.62	0.58	0.53	0.65	0.36
SCEPCL6021C01	BU102828		No hits ^e		0.91	0.88	0.63	0.71	0.53
SCEPCL6023E06	BU102927		No hits ^e		0.56	0.57	1.11	0.68	0.5
SCEPCL6029F09	BU103080		No hits ^e		0.52	0.38	1.28	0.88	0.93
SCEPCL6029F11	BU103082		No hits ^e		0.72	0.91	0.89	0.7	0.46
SCMCCL6027D07	BU103131		No hits ^e		0.58	0.91	1.08	1.08	0.83
SCUTCL6036E11	BU103257		No hits ^e		1.01	1.19	0.7	0.99	0.93
SCEPCL6023A04 ^d	f				1.26	0.33	1.42	1.23	1.25
SCRLCL6032E10 ^d	f				0.74	0.5	0.81	0.72	0.5
SCCCAD1002A12	f				0.84	1.03	0.5	0.41	0.47
SCCCAD1002C11	f				0.79	0.7	0.59	0.91	0.8
SCCCAD1002F11	f				0.66	0.55	0.46	0.41	0.29
SCCCAD1003B01	f				0.73	0.91	0.89	0.41	0.52
SCCCAD1003G01	f				0.84	0.64	0.63	0.38	0.31
SCCCHR1001A09	f				0.78	0.88	0.57	0.52	0.45
SCCCHR1001A10	f				0.69	0.68	0.58	0.5	0.36
SCCCHR1004A01	f				0.43	0.58	0.3	0.57	0
SCCCHR1004C01	f				1.18	0.88	0.29	1.14	0.35
SCEPCL6021B01	f				0.4	0.78	0.92	0.5	0.82
SCEPCL6021G11	f				0.47	0	3.06	2.08	1.46
SCEPCL6023D10	f				0.42	0.42	1.07	0.48	0.32
SCEPCL6028D02	f				1.13	0.54	1.06	0.77	1.2
SCEPCL6029F07	f				0.45	0.7	0.98	0.54	0.39
SCEPCL6029H03	f				0.72	0.78	0.63	0.54	0.57
SCEPCL6029H04	f				0.5	0.75	1.24	0.61	0.96
SCEPCL6029H06	f				0.38	0.27	2.03	0.42	0.44
SCEPCL6029H07	f				0.51	0.18	0.96	0.7	0.62
SCMCCL6027B06	f				0.5	0.44	1.15	0.47	0.66
SCRLCL6032A09	f				0.88	0.58	1.12	0.75	1.4
SCRLCL6032B10	f				0.75	0.54	0.92	1.11	0.74
SCRLCL6032B12	f				0.77	0.44	0.73	0.73	0.59
SCRLCL6032C10	f				0.88	0.59	0.75	0.66	0.48
SCRLCL6032D09	f				0.94	0.54	0.85	0.92	0.77
SCRLCL6032E04	f				0.62	0.47	1.08	1.09	0.91
SCRLCL6032E05	f				0.65	0.4	1.15	1.21	0.83
SCRLCL6032E08	f				0.61	0	1.21	1.05	0.7
SCRLCL6032E09	f				0.88	0.52	0.78	0.7	0.52
SCRLCL6032F10	f				0.91	0.54	0.84	0.69	0.57
SCRLCL6032F12	f				1.58	0.49	0.74	0.57	1.63
SCRLCL6032G06	f				0.92	0.47	0.73	0.84	0.87
SCRLCL6032G09	f				0.73	0.32	0.75	0.54	0.4
SCRLCL6032H05	f				0.84	0.39	0.74	0.78	0.88
SCRLCL6032H08	f				0.73	0.45	0.84	0.88	0.59
SCRLCL6032H09	f				0.83	0.49	0.96	0.82	0.67
SCRLCL6032H10	f				0.87	0.52	0.79	0.74	0.66

a Sequences were compared using the BLASTX algorithm (Altschul *et al.*, 1997). **b** Putative function of gene products, according to

its similar sequences. **c** Geometric means of expression ratios obtained from two independent experiments. **d** Genes also identified as cold-responsive by Nogueira *et al.* **e** Only Blast hits with e-value better than e-10 were considered. **f** No sequencing available.

Capítulo 6: O algoritmo ISER 2 - Seleção de genes diferencialmente expressos em experimentos com réplicas

Na análise de dados de experimentos com arranjos de DNA com réplicas, é preciso em primeiro lugar distinguir as réplicas técnicas das biológicas. As primeiras são provenientes de diferentes pontos de fixação (*spots*) de um mesmo arranjo ou de diferentes hidridações de arranjos com cDNAs obtidos das mesmas amostras de RNA, ou ainda de diferentes amostras de RNA extraídas dos mesmos indivíduos ou tecidos. Já as réplicas biológicas são obtidas a partir de amostras de RNA extraídas independentemente de diferentes indivíduos ou *pools* (conjuntos) de indivíduos num mesmo experimento ou, preferencialmente, de experimentos independentes. A distinção entre esses tipos de réplicas é necessária porque entre as réplicas técnicas não há variabilidade biológica nem a variabilidade causada por flutuações nas condições experimentais, de forma que toda a variação nestes dados se deve apenas às imprecisões da técnica, sejam elas oriundas da marcação da sonda, da hibridação dos arranjos ou da quantificação das imagens obtidas (Yang e Speed, 2002). As réplicas biológicas, por outro lado, são afetadas por todas essas fontes de variação e ainda pela variabilidade biológica, possibilitando assim estimar a mesma (a precisão da estimativa, porém, vai depender do número de réplicas biológicas disponíveis, Churchill, 2002).

Em experimentos com réplicas, o algoritmo ISER pode ser aplicado independentemente a cada hibridação, considerando separadamente as razões de expressão obtidas para cada *spot* nos arranjos. Selecionam-se então como

diferencialmente expressos aqueles genes para os quais todos os respectivos *spots* (ou uma grande fração pré-definida deles) se destaquem como tal. Este critério, porém, pode ser rigoroso demais quando o número de réplicas é grande, com o que é interessante reformular o algoritmo para explorar melhor os dados de experimentos com réplicas.

6.1 – Experimentos apenas com réplicas técnicas, sem réplicas biológicas

Quando o experimento não apresenta réplicas biológicas, nada se pode inferir sobre a variabilidade entre indivíduos na expressão de cada gene. Deseja-se portanto apenas inferir, na comparação de duas amostras de RNA, quais genes apresentam diferenças significativas nos seus níveis de expressão. Suponhamos então que os dados obtidos encontram-se pareados, isto é, para cada gene existem pares de valores de expressão, nos quais cada valor corresponde a uma amostra de RNA. Este é o caso, por exemplo, de dados de microarranjos com dois canais, em que ambas as amostras de RNA são hibridadas em paralelo ao arranjo, marcadas com fluoróforos diferentes. Suponhamos também que para cada gene um ou mais pares de valores são obtidos, provenientes de diferentes *spots* de uma ou mais hibridações, e ainda que a maioria dos genes apresenta níveis de expressão constantes nas amostras estudadas (Quackenbush, 2002).

Nestes casos as análises podem ser iniciadas com um procedimento semelhante ao ISER: a cada hibridação analisada, calculam-se as razões de expressão e as intensidades (médias geométricas dos dois sinais) de cada par de

dados (que doravante chamaremos de *spot*) e adota-se uma transformação dos dados que conduza as razões transformadas o mais próximo possível da distribuição normal. Essa transformação pode ser definida *a priori* como sendo a logarítmica ou ser escolhida entre a família Box-Cox de transformações. Porém, nesse último caso deve-se escolher uma mesma transformação a ser aplicada aos dados de todas as hibridações estudadas (para cada transformação da família Box-Cox calcula-se o logaritmo da função de verossimilhança para as razões transformadas a cada hibridação, sendo selecionada a transformação que minimiza o somatório desses valores). Essa transformação é aplicada tanto às razões de expressão quanto às intensidades.

Aplica-se então uma janela deslizante aos dados de cada hibridação, de maneira análoga ao que é feito no algoritmo ISER, empregando estimadores robustos para determinar, a cada posição da janela, a média e o desvio das razões transformadas. Ainda de maneira análoga ao ISER, um *spline* linear é utilizado para determinar funções de média e desvio dependentes da densidade. Com isso a cada spot pode-se obter um valor corrigido da razão transformada, dado por:

$$x_i = R_i - M(I_i) \quad (6.1)$$

em que R_i é a razão transformada obtida no spot, I_i é a respectiva intensidade transformada e $M(I_i)$ é a função de média calculada em I_i . Essa subtração é equivalente a normalizar os dados (ou renormalizar, caso alguma normalização prévia já tenha sido aplicada). Além disso, para cada x_i , obtém-se um desvio dado por $\sigma_i = D(I_i)$, onde D é a função de desvio. Vale observar que esse valor fornece

uma estimativa da variabilidade da técnica na faixa de intensidade em que se encontram os sinais de cada *spot*, possibilitando tratar de maneira diferenciada *spots* de um mesmo gene que apresentam diferentes intensidades (devido, por exemplo, a diferenças na quantidade de DNA fixada nos *spots*).

Uma vez realizados esses procedimentos em todas as hibridações analisadas, pode-se então adotar para os dados de cada gene g o seguinte modelo:

$$x_{gi} = \theta_g + e_{gi} \quad (6.2)$$

em que θ_g é o real valor da razão de expressão transformada do gene g e e_{gi} é um erro causado pela variabilidade inerente ao método, com distribuição $N(0, \sigma_{gi}^2)$, sendo o valor de σ_{gi} determinado pela função de desvio dependente da intensidade.

Com isso, a cada gene g os dados procedentes dos vários spots de cada hibridação são considerados estimativas do valor θ_g , cada uma com seu respectivo desvio. Pode-se então aplicar um modelo *Bayesiano* (com *priori* não informativo, Bernardo e Smith, 2000) ou um estimador de máxima verossimilhança (Lehman, 1983) - as duas estratégias são equivalentes, uma vez que utilizamos *priori* não informativo no modelo *Bayesiano* - para obter a melhor estimativa do real valor θ_g :

$$\bar{\theta}_g = \frac{\sum \frac{x_{gi}}{\sigma_{gi}^2}}{\sum \frac{1}{\sigma_{gi}^2}}. \quad (6.3)$$

Vale observar que esse estimador é uma média das razões transformadas e normalizadas x_{gi} , ponderada pelos inversos das variâncias σ_{gi}^2 , o que significa que, quando houver diferenças nas intensidades dos *spots*, calcularemos a média dando peso maior para as razões transformadas provenientes de *spots* cujas respectivas intensidades se situam em regiões de menor variabilidade da técnica (em geral, essa variabilidade decresce com o aumento da intensidade, por isso daremos maior peso aos *spots* que apresentam sinais mais fortes). Temos ainda que o desvio desse estimador é dado por:

$$S_g = \frac{1}{\sqrt{\sum \frac{1}{\sigma_{gi}^2}}} . \quad (6.4)$$

Podemos então supor que $\bar{\theta}_g \approx N(\theta_g, S_g^2)$. Como o objetivo da análise é seleccionar os genes onde $\theta_g \neq 0$, observamos que, se $\theta_g = 0$, teremos $\bar{\theta}_g \approx N(0, S_g^2)$, ou seja,

$$\frac{\bar{\theta}_g}{S_g} = \frac{\sum \frac{x_{gi}}{\sigma_{gi}^2}}{\sqrt{\sum \frac{1}{\sigma_{gi}^2}}} \approx N(0,1). \quad (6.5)$$

Desta forma, verificar se um gene é diferencialmente expresso ($\theta_g \neq 0$) torna-se equivalente a avaliar se o valor de $\bar{\theta}_g / S_g$ é significativamente diferente de zero. Mais que isso, podemos ainda atribuir um p-valor para cada gene, que reflete a probabilidade de, sob hipótese nula, encontrarmos dados tão ou mais expressivos que os efetivamente observados:

$$P(Dados | \theta_g = 0) = 2 \times \min \left(\Phi \left(\frac{\bar{\theta}_g}{S_g} \right), \Phi \left(\frac{-\bar{\theta}_g}{S_g} \right) \right), \quad (6.4)$$

sendo Φ a função de distribuição acumulada da normal padrão. Esses p-valores podem então ser utilizados como critério de seleção dos genes diferencialmente expressos, fixando-se para isso um valor de corte para os mesmos abaixo do qual os genes serão selecionados. A distribuição desses valores, calculados para todos os genes, pode ainda ser utilizada para determinar o FDR na seleção resultante de cada valor de corte adotado, como descrito na Seção 6.3.

6.2 – Experimentos com réplicas técnicas e biológicas

Em experimentos onde há diversas réplicas biológicas, torna-se possível avaliar a variabilidade na expressão de cada gene entre essas réplicas. Muitas vezes, porém, o número de réplicas biológicas é ainda pequeno (duas ou três) e testes estatísticos para avaliar essa variabilidade teriam o poder muito reduzido. Além disso, a variabilidade inerente à técnica continua presente, podendo inclusive apresentar-se heterogênea entre as réplicas (devido, por exemplo, a diferenças na eficiência da hibridação). O número de réplicas técnicas ou biológicas disponíveis pode ainda variar de gene para gene, caso o número de pontos de fixação (*spots*) referentes a cada gene nos arranjos não seja homogêneo ou se parte dos dados for eliminada durante a filtragem dos mesmos (Capítulo 2). Assim, é conveniente em primeiro lugar avaliarmos separadamente os dados de cada réplica biológica, reunindo as diversas réplicas técnicas para obter, para cada gene, um estimador de sua razão de expressão.

Seguindo um procedimento análogo ao descrito na seção 6.1, em primeiro lugar, devemos selecionar a transformação que será aplicada aos dados. Para evitar trabalhar com diferentes transformações, esta deve ser a mesma para todas as réplicas biológicas e técnicas. Assim, para cada transformação da família Box-Cox (Box e Cox, 1964), calculamos o logaritmo da função de verossimilhança para as razões transformadas de todas as hibridações do experimento. Selecionamos a transformação que minimizar o valor do somatório desses logaritmos, que é então aplicada tanto às razões de expressão quanto às intensidades de todas as hibridações.

Uma vez que os dados estão transformados, podemos dividi-los por réplicas biológicas. A cada uma destas, separadamente, continuamos seguindo um procedimento semelhante ao descrito na seção 6.1, aplicando aos dados de cada hibridação uma janela deslizante para estimar localmente a média e o desvio dos dados e um *spline* linear para definir funções de média e desvio dependentes da intensidade transformada. Obtemos então para cada *spot* o valor normalizado da razão transformada, $x_i = R_i - M(I_i)$ (sendo R_i a razão transformada obtida no *spot*, I_i a respectiva intensidade transformada e $M(I_i)$ a função de média calculada em I_i) e, para cada gene, reunimos os dados de todas as réplicas técnicas para obter um estimador da sua razão de expressão transformada e seu desvio, dados por:

$$\bar{\theta}_{gj} = \frac{\sum_i \frac{x_{gij}}{\sigma_{ij}^2}}{\sum_i \frac{1}{\sigma_{ij}^2}} \quad \text{e} \quad S_{gj} = \frac{1}{\sqrt{\sum_i \frac{1}{\sigma_{ij}^2}}}, \quad (6.5)$$

em que g é o índice do gene, j corresponde as diversas réplicas biológicas e i varia entre as réplicas técnicas.

Nos experimentos em que o número de réplicas biológicas é reduzido, como é o caso do experimento de deficiência de fosfato descrito no Capítulo 7, a aplicação de múltiplos testes estatísticos não é recomendável para verificar, a cada gene, se a média das razões transformadas é significativamente diferente de zero. Nesses casos, uma opção é selecionar como diferencialmente expressos os genes que se destacarem coerentemente como tal em todas as réplicas biológicas, ou numa fração pré-definida delas. Uma maneira de formalizar esse critério de seleção é através do uso de estatísticas de ordem (Puri e Sem, 1971), descrito a seguir.

Tomando-se, para cada gene g e cada réplica biológica j , o estimador $\bar{\theta}_{gj}$ de sua real razão transformada θ_{gj} , como visto na seção 6.1 podemos supor que o estimador tem distribuição $N(\theta_{gj}, S_{gj}^2)$. Entre os genes não diferencialmente expressos, temos que a real razão de expressão transformada é igual a zero ($\theta_{gj}=0$), logo, definindo-se $y_{gj} = \frac{\bar{\theta}_{gj}}{S_{gj}}$, temos que, nesses casos, cada valor y_{gj} tem distribuição $N(0,1)$. Podemos então ordenar, para um mesmo gene, esses valores provenientes das diferentes réplicas biológicas, obtendo $y_{g(1)} \leq y_{g(2)} \leq \dots \leq y_{g(n)}$, em que n é o número de réplicas biológicas. O k -ésimo desses valores pode então ser usado como critério de seleção dos genes diferencialmente expressos, uma vez que caso esse valor seja significativamente maior que zero, todos os valores $y_{g(k+1)} \dots y_{g(n)}$ também o serão (indicando que o gene é induzido), enquanto que,

caso esse valor seja significativamente menor que zero, todos os valores $y_{g(1)} \dots y_{g(k-1)}$ também o serão (indicando que o gene é reprimido). É claro que se pode tomar valores diferentes de k para testar se o gene é induzido ou reprimido: no primeiro caso toma-se k dentre as menores ordens ($k=1$ ou 2 , por exemplo), enquanto no segundo caso toma-se k dentre as ordens maiores ($k= n$ ou $n-1$, por exemplo).

Para testar se o k -ésimo valor, entre os $y_{g(i)}$, é significativamente diferente de zero, basta observar que, no caso de os $y_{g(i)}$ apresentarem distribuição normal $N(0,1)$, a distribuição acumulada do k -ésimo valor é dada por:

$$F_k(x) = \frac{n!}{(k-1)!} \times \Phi(x)^k \times \sum_{s=0}^{n-k} \frac{(-\Phi(x))^s}{(n-k-s)!s!(s+k)}. \quad (6.6)$$

em que ϕ e Φ são respectivamente a densidade e a distribuição acumulada da $N(0,1)$.

Aplicando-se essa fórmula ao k -ésimo valor $y_{g(k)}$ de cada gene, pode-se então associar um p-valor ao gene. Convém, porém, atribuir p-valores distintos às probabilidades de o gene não ser reprimido ou não ser induzido, uma vez que no primeiro caso considera-se um valor pequeno da ordem k (digamos k_1), enquanto no segundo considera-se um k próximo de n (k_2). Desta forma temos $p\text{-valor(induzido)} = F_k(y_{g(k_1)})$ e $p\text{-valor(reprimido)} = 1 - F_k(y_{g(k_2)})$ (no caso de $k_2 = n - k_2 + 1$ temos $p\text{-valor(reprimido)} = F_k(-y_{g(k_2)})$).

Como na seção 6.1, esses p-valores podem ser utilizados como critério de seleção dos genes diferencialmente expressos, fixando-se para isso um valor de corte para os mesmos. Vale observar que, nesse caso, pode-se fixar valores de corte diferentes para a seleção de genes induzidos ou reprimidos, caso haja mais

interesse em uns do que em outros. O FDR pode também ser determinado independentemente para as seleções de genes induzidos ou reprimidos, de acordo com as distribuições globais dos dois conjuntos de p-valores, como descrito na seção 6.3.

No caso de haver um número maior de réplicas biológicas no experimento, outras estratégias de análise podem ser aplicadas aos valores dos estimadores $\bar{\theta}_{gj}$ e de seus desvios S_{gj} para selecionar os genes diferencialmente expressos. Essas estratégias, porém, não compõem os objetivos deste trabalho. O enfoque principal do mesmo foi desenvolver estratégias para a análise de dados de experimentos que apresentam poucas réplicas biológicas, como é o caso da maioria dos experimentos com arranjos de DNA, como o experimento de deficiência de fosfato descrito no Capítulo 7. Algumas dessas estratégias serão mencionadas no restante dessa seção, apenas como sugestão para estudos futuros, mas não serão mais desenvolvidas nas demais seções deste capítulo ou nos capítulos seguintes.

Quando há um número maior de réplicas biológicas no experimento, um teste t de Student pode ser aplicado para cada gene, sobre os valores dos estimadores $\bar{\theta}_{gj}$, para verificar se a média desses valores é significativamente diferente de 0. Essa estratégia, porém, tem a desvantagem de não levar em consideração a variabilidade dos estimadores $\bar{\theta}_{gj}$, medida pelos desvios S_{gj} . Para contornar esse problema, podemos aplicar o teste t apenas aos genes que passarem numa seleção prévia, na qual são escolhidos aqueles cujas razões de expressão transformadas nas diversas réplicas biológicas se destacam da

distribuição global das mesmas. Para isso, consideramos o vetor de estimadores

$$\bar{\theta}_g = (\bar{\theta}_{g1}, \bar{\theta}_{g2}, \dots, \bar{\theta}_{gn}).$$

Observamos que, nos genes que não são diferencialmente expressos, esse vetor tem distribuição normal multivariada com média $\underline{0}_{n \times 1}$ e suas entradas são independentes, tendo portanto matriz de variância-covariância diagonal:

$$\Sigma = \begin{pmatrix} S_{g1}^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & S_{gn}^2 \end{pmatrix}. \quad (6.7)$$

Com isso, temos

$$\bar{\theta}_g' \Sigma^{-1} \bar{\theta}_g = \sum_{s=1}^n \frac{\bar{\theta}_{gs}^2}{S_{gs}^2} \quad (6.8)$$

e esse valor tem distribuição χ_n^2 (Rao, 2002), sendo n o número de réplicas biológicas e, desta forma, podemos atribuir um p-valor ao gene g e selecionar os genes que se destacam da distribuição global das razões de expressão, aplicando então o teste t apenas nesses casos.

Uma outra alternativa para a análise de experimentos com um número maior de réplicas biológicas é aplicar aos dados de cada gene uma análise bayesiana (Bernardo e Smith, 2000). Nesse caso, o p-valor de cada gene seria dado por:

$$2 \times \min \left(\int_{-\infty}^0 \int_{-\infty}^0 f_n \left(\frac{\theta_{gj} - \bar{\theta}_{gj}}{S_{gj}} \right) \times f_t \left(\frac{\sum (\theta_{gj} - \mu_g)}{\sqrt{n} \times S_{\theta}} \right) d\mu_g d\theta_{gj}, \int_0^{\infty} \int_0^{\infty} f_n \left(\frac{\theta_{gj} - \bar{\theta}_{gj}}{S_{gj}} \right) \times f_t \left(\frac{\sum (\theta_{gj} - \mu_g)}{\sqrt{n} \times S_{\theta}} \right) d\mu_g d\theta_{gj} \right) \quad (6.9)$$

em que f_n e f_t são respectivamente as densidades da distribuição normal multivariada e do teste t com n-1 graus de liberdade, θ_{gj} representam os reais

valores das razões transformadas do gene g a cada réplica biológica j , S_θ é o desvio padrão desses valores e μ_g representa o real valor da média das razões transformadas em toda a população (esse valor é aproximado pela média dos θ_{gj} , e a integral múltipla acima na verdade calcula a probabilidade de $\mu_g \neq 0$). Apesar de essa estratégia ser bastante completa, considerando tanto a variabilidade técnica quanto a biológica no cálculo dos p-valores, ela tem a grande desvantagem de ser extremamente dispendiosa computacionalmente. Essas integrais múltiplas precisam ser calculadas numericamente para cada gene, o que demanda grande esforço computacional e pode gerar imprecisões.

As estratégias descritas acima, voltadas para a análise de experimentos com um número maior de réplicas biológicas, podem ser aplicadas apenas quando esse número é maior ou igual a três. Porém, essas estratégias só são eficientes quando esse número é realmente grande (como oito ou dez, por exemplo), bem maior que o número de réplicas biológicas disponíveis na grande maioria dos experimentos com arranjos de DNA, e somente nesses casos devem ser aplicadas. Desta forma, nas próximas seções deste capítulo e nos capítulos seguintes nos concentraremos em testes e aplicações da estratégia de análise baseada em estatísticas de ordem, que é indicada para experimentos com poucas réplicas biológicas e ainda se mostra uma eficiente opção em experimentos onde há um número maior delas.

6.3 - O controle do FDR no algoritmo ISER 2

Uma vez atribuídos p-valores para todos os genes testados no estudo, como descrito nas seções 6.1 e 6.2, pretende-se selecionar como diferencialmente expressos aqueles cujo p-valor está abaixo de um determinado limite, como por exemplo 5% ou 1%. Porém, em toda seleção de genes diferencialmente expressos, há sempre uma margem de erro e, devido ao grande número de testes, uma quantidade deles será erroneamente selecionada. Uma maneira de monitorar a frequência de erros cometidos é estimar o FDR (*False Discovery Rate*, Benjamini e Hochberg, 1995), já mencionado no Capítulo 4. Esta razão é definida como a proporção esperada de genes não diferencialmente expressos incluídos ao acaso na seleção adotada, e também pode ser considerada como a probabilidade de cada gene na seleção ser um falso positivo. Matematicamente, o FDR é definido como:

$$FDR = \frac{\text{número de genes selecionados por engano}}{\text{número total de genes selecionados}}. \quad (6.10)$$

Para estimar esse valor, observamos que, entre os genes que não são diferencialmente expressos, os p-valores terão distribuição $U(0,1)$ (Teorema de Inversão, Gentle, 2003). Chamemos então de π_0 a proporção de genes cuja expressão não é diferenciada (ou seja, que estão sob a hipótese nula), entre todos os genes estudados. Se pudéssemos estimar π_0 , teríamos que o número de genes não diferenciados seria aproximadamente $N \times \pi_0$, onde N é o total de genes analisados. Como esses genes teriam p-valores distribuídos uniformemente no

intervalo (0,1), tomando-se um p-valor de corte igual a p_1 , esperaríamos selecionar $p_1 \times N \times \pi_0$ destes, o que nos levaria a estimar o FDR como:

$$FDR \approx \frac{p_1 \times N \times \pi_0}{\text{número total de genes com } p\text{-valor} \leq p_1} \quad (6.11)$$

Com isso, resta apenas estimar a proporção π_0 . No entanto, esta estimativa precisa ser feita a partir da distribuição global dos p-valores, e não é trivial. Como visto acima, os p-valores dos genes não diferencialmente expressos têm distribuição $U(0,1)$. Assim, a distribuição global dos p-valores será uma mistura de distribuições (Allison *et al.*, 2002), com a forma:

$$F = \pi_0 \times U(0,1) + (1 - \pi_0) \times F_1, \quad (6.12)$$

em que F_1 é a distribuição dos p-valores dentre os genes diferencialmente expressos. Esta última é desconhecida, porém em geral (em testes bicaudais) sua densidade é concentrada próximo de 0, uma vez que os genes diferencialmente expressos apresentam p-valores pequenos. Além disso, usualmente essa densidade é decrescente, tendendo a 0 para p-valores próximos de 1.

Uma fórmula habitualmente utilizada para estimar a proporção π_0 (Storey, 2002), baseada nas características de F_1 descritas acima, é:

$$\hat{\pi}_0 = \frac{\text{Número de } p\text{-valores} > \lambda}{N \times (1 - \lambda)}, \quad (6.13)$$

sendo N o número total de genes em estudo e λ um valor arbitrário entre 0 e 1, próximo de 1. Esta estratégia não paramétrica para estimar π_0 justifica-se pela suposição de que a densidade da distribuição F_1 tende a zero para p-valores próximos de 1. Assim, se λ é bastante grande, os p-valores maiores que λ são

todos provenientes de genes não diferencialmente expressos, tendo portanto distribuição $U(0,1)$ e, com isso, seu número esperado é aproximadamente $N \times \pi_0 \times (1 - \lambda)$. Escolhas usuais para o valor de λ são 0,75 ou 0,9, porém a eficiência desta estratégia depende muito da escolha de λ e da suposição de que a densidade de F_1 tende a zero para p-valores próximos de 1. Frequentemente, esta estratégia superestima o valor de π_0 , gerando assim uma tendência no cálculo do FDR (Pawitan et al., 2005). Além disso, esta estratégia não se aplica a situações em que a densidade de F_1 não é decrescente, como é o caso dos p-valores gerados por testes monocaudais.

Uma forma mais eficiente de aproximar F_1 é através de modelos paramétricos. Isto pode ser feito de duas formas, dependendo do teste estatístico empregado para gerar os p-valores: em testes bicaudais, onde tanto os genes induzidos como os reprimidos terão p-valores próximos de zero, podemos supor que F_1 é aproximadamente uma distribuição $Beta(\alpha, 1)$ com $0 < \alpha < 1$ (Pounds e Morris, 2003); por outro lado, em teste monocaudais, onde os p-valores dos genes diferencialmente expressos irão se concentrar nas duas extremidades do intervalo $[0, 1]$ (como é o caso dos testes baseados nas estatísticas de ordem, descritos na seção 6.2), podemos supor que F_1 é aproximadamente uma mistura de duas distribuições beta:

$$F_1 = \pi_1 \times B(\alpha, 1) + (1 - \pi_1) \times B(1, \beta), \quad (6.14)$$

em que $0 < \alpha < 1$ e $0 < \beta < 1$ (Figura 6.1).

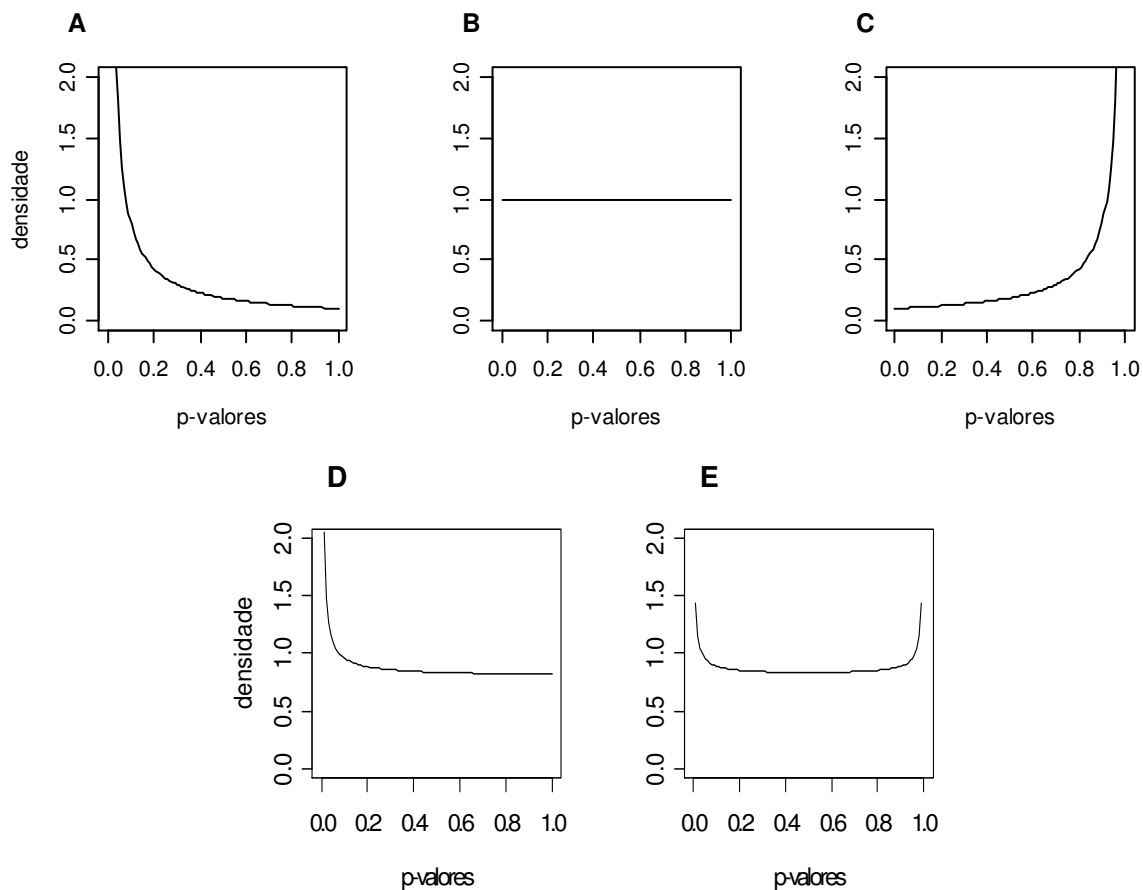


Figura 6.1: Densidades das distribuições: A) distribuição Beta(0.1,1), B) distribuição uniforme, C) distribuição Beta(1, 0.1), D) mistura beta-uniforme, E) mistura beta-uniforme-beta.

Sendo F_1 igual a uma distribuição beta ou a uma mistura de duas distribuições beta, a distribuição geral dos p-valores (equação 6.12) será uma mistura desta aproximação com uma distribuição uniforme. Desta forma, estimar a proporção π_0 equivale a estimar os parâmetros da mistura de distribuições que melhor aproxima a real distribuição dos p-valores. Desenvolvemos estratégias distintas para estimar esses parâmetros em cada caso. Designaremos por “mistura beta-uniforme” a situação em que F_1 é aproximada por uma distribuição

beta e por mistura “beta-uniforme-beta” a situação em que F_1 é aproximada pela mistura de duas distribuições beta. Vale a pena ressaltar que as estratégias desenvolvidas não se aplicam apenas aos p-valores gerados pelo algoritmo ISER2, mas também a conjuntos de p-valores obtidos com qualquer teste estatístico aplicado, desde que se possa supor que suas distribuições se adequam a um dos modelos (o que ocorre na grande maioria dos casos).

6.3.1 – Mistura beta-uniforme

Essa é a distribuição usual dos p-valores de testes bicaudais, como é o caso da maioria dos testes estatísticos usualmente aplicados a dados de arranjos de DNA. Por exemplo, a versão do algoritmo ISER2 para experimentos em que há apenas uma réplica biológica, descrita na seção 6.1, gera p-valores com essa distribuição, assim como o tradicional teste t de Student. Nesse caso a distribuição dos p-valores será modelada por:

$$F = \pi_0 \times U(0,1) + (1 - \pi_0) \times B(\alpha,1), \quad (6.15)$$

sendo $0 < \pi_0 < 1$ e $0 < \alpha < 1$. Para determinar os valores de α e π_0 correspondentes à mistura que melhor aproxima a distribuição dos p-valores, recorreremos às técnicas de simulações *Monte Carlo* e reamostras *bootstrap*, descritas a seguir.

Simulações Monte Carlo (Gamerman, 1997): Para diferentes valores de π_0 e α (o parâmetro da distribuição beta), geramos conjuntos simulados de p-valores com tantos elementos quanto genes no conjunto de dados analisado, tendo como distribuição a mistura beta-uniforme descrita acima. Para cada par de parâmetros

π_0 e α um grande número de conjuntos simulados é gerado (1000, por exemplo). Cada conjunto é ordenado e as médias a cada quantil (isto é, a média dos menores valores em cada conjunto, a média dos segundos menores, dos terceiros menores e assim sucessivamente até a média dos maiores valores) são calculadas. Esses conjuntos de médias dos valores simulados são armazenados, assim como seus respectivos pares de parâmetros, para mais tarde serem comparados com os conjuntos de p-valores provenientes do experimento.

A grande quantidade de cálculos realizados nessas simulações constitui um procedimento computacionalmente dispendioso, demandando bastante tempo para seu processamento. No entanto esse procedimento é feito apenas uma vez para todos os conjuntos de dados analisados em um experimento, ou ainda em vários experimentos com o mesmo número de genes, e portanto o dispêndio de tempo não constitui um empecilho para a aplicação dessa estratégia.

Reamostras bootstrap (Efron e Tibshirani, 1993): A aplicação desta técnica tem o objetivo de obter aproximações mais suaves da distribuição amostral dos p-valores. Para isso, conjuntos de reamostras do conjunto de p-valores com o mesmo número de elementos que este são tomados com repetições, de forma que cada reamostra tem aproximadamente a mesma distribuição do conjunto original. Um grande número de conjuntos de reamostras (1000, por exemplo) é tomado e cada conjunto é ordenado. As médias a cada quantil (isto é, a médias dos menores valores em cada re-amostra, a média dos segundos menores e assim sucessivamente) são calculadas, obtendo assim um conjunto de médias

cuja distribuição é semelhante à distribuição dos p-valores originais, porém tende a ser mais “suave” (Figura 6.2). Como a amostra de p-valores pode apresentar distribuição empírica com diferentes graus de suavidade (devido ao número de réplicas e à inerente complexidade da estrutura de variância dos dados e do problema em estudo), esta técnica suaviza o seu histograma para homogeneizar os valores que serão utilizados para estimar os parâmetros α e π_0 . Com isso, os resultados obtidos tornam-se menos dependentes de flutuações intrínsecas dos dados (ruídos).

Uma vez obtido o conjunto das médias das reamostras *bootstraps*, podemos compará-lo com os conjuntos de dados gerados pelas simulações Monte Carlo. A cada comparação, as diferenças dos valores a cada quantil são calculadas, e seus quadrados são obtidos e somados, ou seja, chamando de b_i o conjunto das médias *bootstraps* e de mc_i um conjunto de dados das simulações Monte Carlo, obtemos o somatório $\sum_{i=1}^N (b_i - mc_i)^2$, em que N é o número total de genes no experimento. Repetindo esse procedimento para todos os conjuntos de dados das simulações Monte Carlo, selecionamos então o conjunto para o qual o somatório acima é menor. Tomamos então os respectivos valores de π_0 e α como os reais parâmetros da distribuição original dos p-valores. Para aumentar a confiabilidade do método, todo esse procedimento pode ser ainda repetido diversas vezes, armazenando-se os valores dos parâmetros π_0 e α encontrados a cada repetição. Devido ao custo computacional elevado, o número de repetições adotado não pode ser muito grande: 100, por exemplo, mostrou-se satisfatório na prática, ainda que um número bem menor que a quantidade de reamostras

bootstrap utilizadas para suavizar a distribuição dos p-valores (1000, que nesse caso seria também um valor razoável). Por fim, toma-se a mediana dos valores de π_0 como sendo a real proporção de genes não diferencialmente expressos.

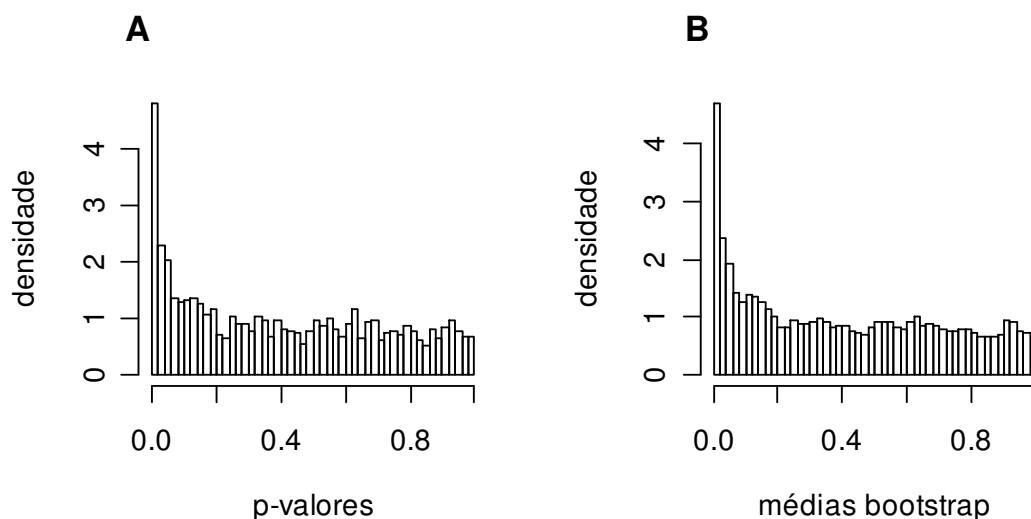


Figura 6.2: Distribuições de conjuntos *bootstraps*: A) histograma de p-valores obtidos com a aplicação do algoritmo ISER2 nos dados de uma hibridação de arranjo de DNA do experimento de deficiência de fosfato (Capítulo 7, amostra 12h-P), B) histograma do conjunto de médias de reamostras *bootstrap* do mesmo conjunto de p-valores, evidenciando a suavização obtida.

6.3.2 – Mistura beta-uniforme-beta

Essa é a distribuição usual dos p-valores de testes monocaudais, nos quais os p-valores dos genes diferencialmente expressos irão se concentrar nas duas extremidades do intervalo $[0,1]$. Esse é o caso, por exemplo, dos testes baseados em estatísticas de ordem, descritos na seção 6.2. Nesses casos não podemos supor que a distribuição dos p-valores é uma mistura beta-uniforme, e precisamos

acrescentar no modelo um termo responsável pela concentração de p-valores próximos de 1. Assumimos então uma mistura beta-uniforme-beta, na forma

$$F = \pi_0 \times U(0,1) + (1 - \pi_0) \times \pi_1 \times B(\alpha, 1) + (1 - \pi_0) \times (1 - \pi_1) \times B(1, \beta), \quad (6.16)$$

em que as densidades das distribuições beta são semelhantes às mostradas na figura 6.1.

Para determinar a proporção π_0 nesses casos, precisaríamos também determinar os parâmetros π_1 , α e β . Com isso o uso de simulações *Monte Carlo* torna-se proibitivo, já que o maior número de parâmetros a serem estimados aumenta muito as possibilidades de suas combinações, assim como o número de simulações a serem realizadas, o que demandaria vários dias de intenso uso computacional. O emprego de reamostras *bootstrap* ainda é uma estratégia eficiente para obter aproximações mais suaves da distribuição dos p-valores; porém a cada conjunto de médias de reamostras obtido é preciso adotar um procedimento diferente para estimar esses parâmetros.

Um método usualmente empregado para estimar parâmetros de um modelo de mistura de densidades é o algoritmo EM (Lange, 1997), que procura determinar os mesmos através da maximização da verossimilhança dos dados observados. Porém, devido ao grande número de genes usual em experimentos com arranjos de DNA, o algoritmo EM se torna muito lento, inviabilizando seu uso. Pawitan et al. (2005) propôs uma estratégia mais eficiente para estimar a proporção π_0 , baseada na distribuição dos valores da estatística utilizada nos dados, como por exemplo o teste t de Student. Remodelamos essa estratégia para aplicação aos p-valores gerados pelo algoritmo ISER2, como descrito a seguir.

Dividimos o intervalo $(0,1)$ em n subintervalos $[T_{k-1}, T_k]$, todos de mesmo tamanho, com $T_0=0$ e $T_n=1$ ($n=100$, por exemplo). Computamos então o vetor $y=(y_1, \dots, y_n)$, onde y_k é o número de p-valores, entre o conjunto cuja distribuição queremos aproximar, contidos no intervalo $[T_{k-1}, T_k]$. Temos então que y terá distribuição multinomial com probabilidades $(\alpha_1, \dots, \alpha_n)$, em que $\alpha_k = F(T_k) - F(T_{k-1})$, sendo F a distribuição dos p-valores que desejamos aproximar. Supondo então que F é uma mistura beta-uniforme-beta, para cada conjunto de parâmetros π_0, π_1, α e β , podemos calcular o valor das probabilidades $(\alpha_1, \dots, \alpha_n)$:

$$\alpha_k = \pi_0 \times (T_k - T_{k-1}) + (1 - \pi_0) \times \pi_1 \times (T_k^\alpha - T_{k-1}^\alpha) + (1 - \pi_0) \times (1 - \pi_1) \times ((1 - T_{k-1})^\beta - (1 - T_k)^\beta). \quad (6.17)$$

Uma vez calculadas essas probabilidades, o logaritmo da função de verossimilhança do vetor y é dado por:

$$\log L(\pi_0, \pi_1, \alpha, \beta) = \sum_{k=1}^n y_k \log(\alpha_k). \quad (6.18)$$

Desta forma, uma vez que os valores y_k são fixos, torna-se simples aplicar um algoritmo de otimização (Everitt, 1987) ao logaritmo da função de verossimilhança para encontrar os valores dos parâmetros π_0, π_1, α e β que a maximizam. Esses valores são então aceitos como os parâmetros da distribuição do conjunto de p-valores.

Por fim, combinamos esta estratégia de determinação da proporção π_0 com a técnica de *bootstrap* descrita na seção 6.3.1: tomamos reamostras *bootstrap* dos p-valores, ordenamo-las e calculamos as médias a cada quantil. Aplicamos então

a estratégia descrita acima a esse conjunto de médias para estimar o valor de π_0 . Repetimos ainda esse procedimento diversas vezes (novamente o número de repetições é limitado pelo custo computacional das mesmas: 100, por exemplo, mostrou-se satisfatório na prática), armazenando os valores de π_0 , e tomamos a mediana desses valores como sendo a real proporção de genes não diferencialmente expressos entre todos os genes analisados.

6.3.3 – FDR e q-valores

Nas seções anteriores (6.3.1 e 6.3.2) desenvolvemos estratégias para calcular o FDR na seleção de genes obtida em decorrência da definição de um valor de corte para os p-valores. Podemos agora fazer o raciocínio inverso, fixando uma frequência de falsos positivos tolerável na seleção de genes e determinando o valor de corte correspondente para os p-valores, ou seja, podemos determinar a seleção que será adotada a partir do FDR que estamos dispostos a tolerar na mesma. Uma maneira de fazer isso é utilizando o conceito de q-valor (Storey, 2002) que, atribuído a cada gene, reflete o FDR que seria obtido caso fizéssemos a seleção com um valor de corte para os p-valores grande o suficiente para o gene em questão ser incluído na mesma, isto é, se tomássemos o valor de corte igual ao p-valor de gene em questão.

Considerando então $FDR(p_1)$ como sendo o FDR esperado se tomarmos p_1 como o valor de corte dos p-valores, dado um gene g temos a princípio

$$q - valor_g = FDR(p - valor_g) = \frac{N \times \pi_0 \times p - valor_g}{\text{número de } p - \text{valores} \leq p - valor_g}. \quad (6.19)$$

Porém, uma pequena correção deve ser feita na definição acima: espera-se que o FDR seja crescente em função do valor de corte adotado, logo espera-se também que os q-valores sejam crescentes em função dos p-valores. Isto pode não ocorrer devido a pequenas flutuações na distribuição dos p-valores; assim, adotamos uma estratégia para garantir a monotonicidade dos q-valores, definindo

$$q - valor_g = \max_{p_1 \leq p - valor_g} [FDR(p_1)] = \max_{p_1 \leq p - valor_g} \left[\frac{N \times \pi_0 \times p_1}{\text{número de } p - \text{valores} \leq p_1} \right]. \quad (6.20)$$

Com essa definição, os q-valores podem passar a ser usados como critério de seleção dos genes diferencialmente expressos, no lugar dos p-valores. Desta forma, selecionar os genes definindo um valor de corte para os q-valores torna-se mais intuitivo, uma vez que este pode ser encarado como a frequência de falsos positivos tolerável na seleção.

6.4 – Testes com dados simulados

De maneira análoga ao que foi feito com o algoritmo ISER, o ISER2 foi testado em dados simulados antes de ser aplicado aos dados reais do experimento de deficiência de fosfato (Capítulo 7). Neste caso, porém, foram simulados dados de microarranjos de DNA com dois “canais”, isto é, hibridados com duas amostras de cDNA marcadas com diferentes fluoróforos. As simulações foram baseadas no modelo de dados de arranjos de DNA descrito por Kerr *et al.* (2000), com termos representativos das diversas fontes de variações nos logaritmos dos sinais obtidos nos dois canais dos arranjos sendo gerados aleatoriamente e adicionados separadamente a cada canal, para então converter-

se os valores desses logaritmos nos sinais supostamente obtidos. Os dados gerados representavam resultados de experimentos com arranjos de 2000 genes, cada um representado em quatro *spots*, hibridados com amostras de cDNA provenientes de dois grupos de indivíduos (ou tecidos) a serem comparados, que chamamos de “tratado” e “controle”. Simulavam hibridações de duas réplicas biológicas, cada qual com duas réplicas técnicas, de forma que cada conjunto analisado consistia de dados de hibridações de quatro arranjos (procurou-se com isso gerar estruturas de dados semelhantes à dos dados do experimento de deficiência de fosfato, descrito no Capítulo 7). Foram testadas situações com diferentes números de genes diferencialmente expressos: 400 (20% do total de genes), 200 (10%) e 100 (5%), sempre com metade destes induzidos e metade reprimidos. Para cada cenário, 100 conjuntos de dados foram gerados e analisados, sendo 50 deles com um desvio na intensidade de cada spot igual a 0,7 e 50 com esse desvio igual a 3 (ver detalhes abaixo).

6.4.1 – Modelo adotado nas simulações de dados

Em cada conjunto de dados, o sinal S_{gsdctb} do gene g , no spot s , no grupo i , (tratado ou controle), representado pelo canal (fluoróforo) d , na hibridação da réplica técnica t , da réplica biológica b , foi dado por

$$\log(S_{gsidtb}) = U + B_b + A_t + G_g + E_s + D_d + GB_{gib} + e_{gsidtb}, \quad (6.21)$$

em que cada um dos termos foi definido como descrito a seguir.

U era a média geral das intensidades dos spots, cujo valor é único para todas as hibridações de cada conjunto de dados simulados e tem distribuição $N(10, 0.2^2)$.

B_b representava a variação na média geral de cada réplica biológica, com distribuição $N(0, 0.5^2)$.

A_t representava a variação na média geral de cada réplica técnica, com distribuição $N(0, 0.1^2)$. Com isso, a cada hibridação a soma $U + B_b + A_t$ representava a média dos sinais de cada hibridação, em ambos os canais.

G_g representava a variação na intensidade dos genes, enquanto E_s era a variação na intensidade do sinal de cada *spot*. As distribuições desses dois conjuntos de valores foram definidas de forma que a distribuição global das intensidades de todos os *spots* fosse semelhante em todos os conjuntos de dados simulados. Em metade desses conjuntos, G_g foi gerado a cada gene com distribuição $U(-1, 1)$, enquanto E_s foi gerado para cada *spot* com distribuição $N(0, 3^2)$. Nos demais conjuntos, porém, E_s foi gerado com distribuição $N(0, 0.7^2)$ e, para compensar essa redução na variância de E_s , G_g foi gerado como a soma de uma distribuição $U(-1, 1)$ com uma distribuição $N(0, 8.5)$. Desta forma, em ambos os casos, a soma $G_g + E_s$ teve distribuição $U(-1, 1) + N(0, 3^2)$.

D_d simulava uma tendência nos sinais de cada fluoróforo, com $d=1$ ou 2 . Esses valores tinham distribuição $N(0, 1)$ e foram gerados apenas uma vez a cada conjunto de dados simulados. Como os fluoróforos foram associados alternadamente aos grupos “tratado” e “controle” em cada hibridação, os dois valores D_1 e D_2 foram adicionados alternadamente aos logaritmos dos sinais provenientes de cada grupo.

GB_{gib} representava o logaritmo da real variação na expressão de cada gene g na réplica biológica b . Estes valores foram diferentes de 0 apenas para os genes diferencialmente expressos. Nesses casos, em primeiro lugar gerou-se um valor

médio para o logaritmo da razão de expressão do gene, R_g , a partir de uma distribuição gama com média 0,5 e desvio 0,2 (isto é, distribuição gama com parâmetro de localização igual a 6,25 e escala 0,08), sendo que para os genes reprimidos esses valores foram multiplicados por -1. Para metade dos genes diferencialmente expressos (escolhidos ao acaso), o valor assim obtido foi adotado como o logaritmo da razão de expressão em ambas as réplicas biológicas (digamos $R_{g1} = R_{g2} = R_g$). Para os demais, a cada réplica biológica esse valor foi ainda multiplicado por um número gerado com distribuição uniforme no intervalo [0,5,1,5] (ou seja $R_{gb} = R_g \times u$, com $b=1$ ou 2 e $u \sim U(0,5, 1,5)$), buscando com isso simular variações na razão de expressão do gene entre as réplicas biológicas. Uma vez definidos o logaritmo da razão de expressão para cada gene a cada réplica biológica, a metade desse valor foi adicionada ao logaritmo do sinal referente ao grupo “tratado” e metade foi subtraída do sinal referente ao grupo “controle”, em todos os *spots* correspondentes ao gene em ambas as réplicas técnicas da réplica biológica (isto é $GB_{gib} = R_{gb}/2$ para $i=\text{“tratado”}$ e $GB_{gib} = -R_{gb}/2$ para $i=\text{“controle”}$). Desta forma, nestes *spots* alterou-se o valor do logaritmo da razão de expressão (tratado/controle) sem alterar a intensidade de cada *spot*.

e_{gsdctb} simulava o erro no sinal S_{gsdctb} , gerado com distribuição normal com média 0 e desvio dependente da intensidade do spot, isto é, dependente do valor de $U + B_b + A_t + G_g + E_s + D_d + GB_{gib}$. Para tanto, a cada hibridação simulada definiu-se uma transformação linear que levava o mínimo destas intensidades a 0.5 e o máximo a 8:

$$T(x) = 0.5 + 7.5 \times \frac{x - \min(\text{intensidade})}{\max(\text{intensidade}) - \min(\text{intensidade})} \quad (6.22)$$

(os valores 0.5 e 8 foram arbitrários, escolhidos para que os dados gerados tivessem distribuição semelhante a dados reais de arranjos de DNA). Em seguida, gerou-se um valor inicial do erro na hibridação, e_0 , a partir de uma distribuição uniforme $U(3, 4)$. Adotou-se então o valor $\sigma = e_0/T(i)^2$ como desvio do erro de cada spot, onde i é a intensidade do spot. (ou seja, $e_{\text{gsdctb}} \sim N(0, \sigma^2)$).

Por fim, aos dados de cada hibridação simulada acrescentou-se ainda uma tendência dependente da intensidade, da mesma forma como descrito na seção *ISER: application to simulated data sets* do capítulo 5. Como exemplo, um conjunto de dados simulado pode ser visto na Figura 6.3.

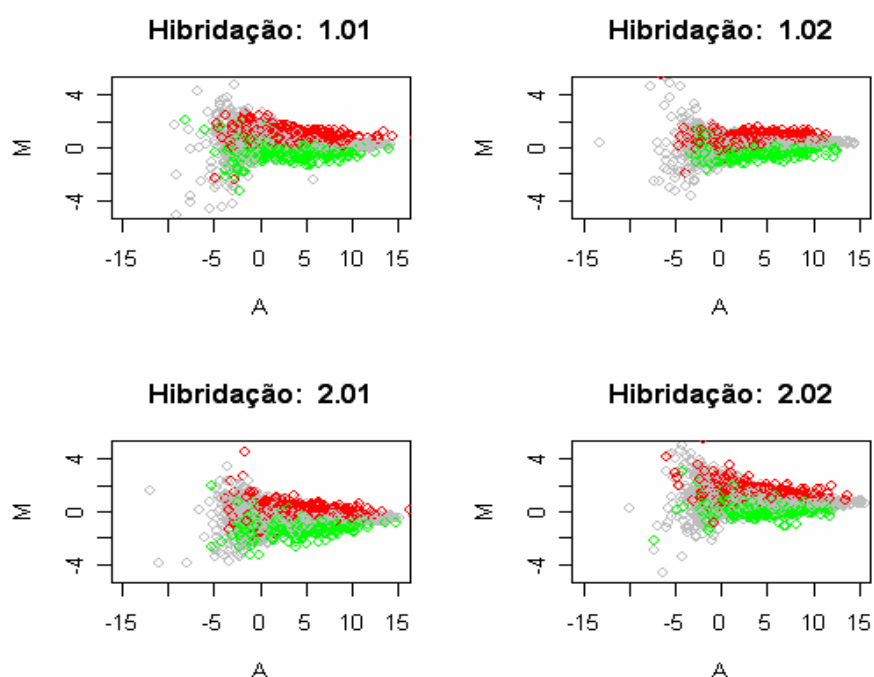


Figura 6.3: Gráficos dos logaritmos das razões (M) em função das intensidades (A , também em escala logarítmica) de um conjunto de dados simulados. Os pontos em vermelho são os genes induzidos e os em verde são os reprimidos, enquanto os pontos em cinza são os genes que não são diferencialmente expressos. Cada gráfico representa os dados de uma hibridação. Hibridação 1.01 e 1.02: representam réplicas técnicas de uma réplica biológica 1; Hibridação 2.01 e 2.02: representam réplicas técnicas da réplica biológica 2.

6.4.2 - Análise dos dados simulados

Apesar de o ISER2 dispensar normalizações prévias, foi conveniente normalizar os dados simulados através do algoritmo de loess (descrito no capítulo 2), de forma a não tornar tendenciosa a comparação do ISER2 com outras estratégias de análise. Este algoritmo foi aplicado independentemente aos dados de cada hibridação em cada conjunto de dados simulados.

Cada conjunto de dados simulados foi analisado com os algoritmos ISER2 e SAM (Tusher *et al.*, 2001), e com o teste t de Student para amostras pareadas. Após cada análise diversos valores de corte para os p-valores foram adotados na seleção de genes diferencialmente expressos, computando-se para cada seleção obtida as razões de positivos verdadeiros (fração dos genes diferencialmente expressos que é selecionada, também chamada de sensibilidade do método) e de falsos positivos (fração dos genes que não são diferencialmente expressos que é erroneamente selecionada como tal, equivalente a $1-\eta$, sendo η a especificidade do método). Esses valores foram utilizados para traçar as curvas ROC (*Receiver operating characteristic*, gráficos das razões de positivos verdadeiros contra as razões de falsos positivos), que evidenciam as performances relativas das várias estratégias de análise (Figuras 6.4 e 6.5). As razões FDR estimadas pelos algoritmos ISER2 e SAM também foram computadas, assim como os reais valores do FDR, comparando-se assim as eficiências desses algoritmos em estimar essa razão (Figuras 6.4 e 6.5).

6.4.3 – Resultados das simulações

Os resultados obtidos diferem bastante entre as situações onde o desvio na intensidade de cada spot (E_s) é igual a 0,7 e aquelas onde esse desvio é igual a 3. No primeiro caso, o valor pequeno desse desvio, aliado à transformação exponencial adotada para gerar os dados e à distribuição normal utilizada para gerar os erros nos logaritmos dos mesmos, faz que os dados gerados sejam adequados às condições necessárias para a aplicação do teste t de Student e do algoritmo SAM (que é uma variação do teste t, como descrito na seção 4.3). Desta forma, essas estratégias de análise apresentam o melhor desempenho nessas situações. O algoritmo ISER2, porém, mostra desempenho bastante próximo ao das demais estratégias, ainda que ligeiramente inferior, como pode ser visto na Figura 6.4.

Nas situações em que o desvio na intensidade de cada *spot* é igual a 3, torna-se clara o melhor desempenho do algoritmo ISER2 em relação às outras estratégias de análise testadas. Nesses casos, a estrutura dos dados é diferente das condições motivadoras do teste t e do SAM. Devido ao maior desvio nas intensidades dos *spots* de cada gene, podem-se obter, para um mesmo gene, sinais de diferentes *spots* com diferentes variabilidades, e nem o teste t como tampouco o SAM conseguem se adaptar a essas diferenças. Com isso, o melhor desempenho do ISER2 torna-se claro (Figura 6.5), evidenciando assim a maior flexibilidade do algoritmo. Em dados reais, espera-se que haja uma mistura de genes nos quais a variabilidade das intensidades dos sinais é pequena e de genes que apresentam uma maior variabilidade nesses sinais. Por exemplo, nos dados do experimento de deficiência de fosfato (Capítulo 7), a maioria dos genes

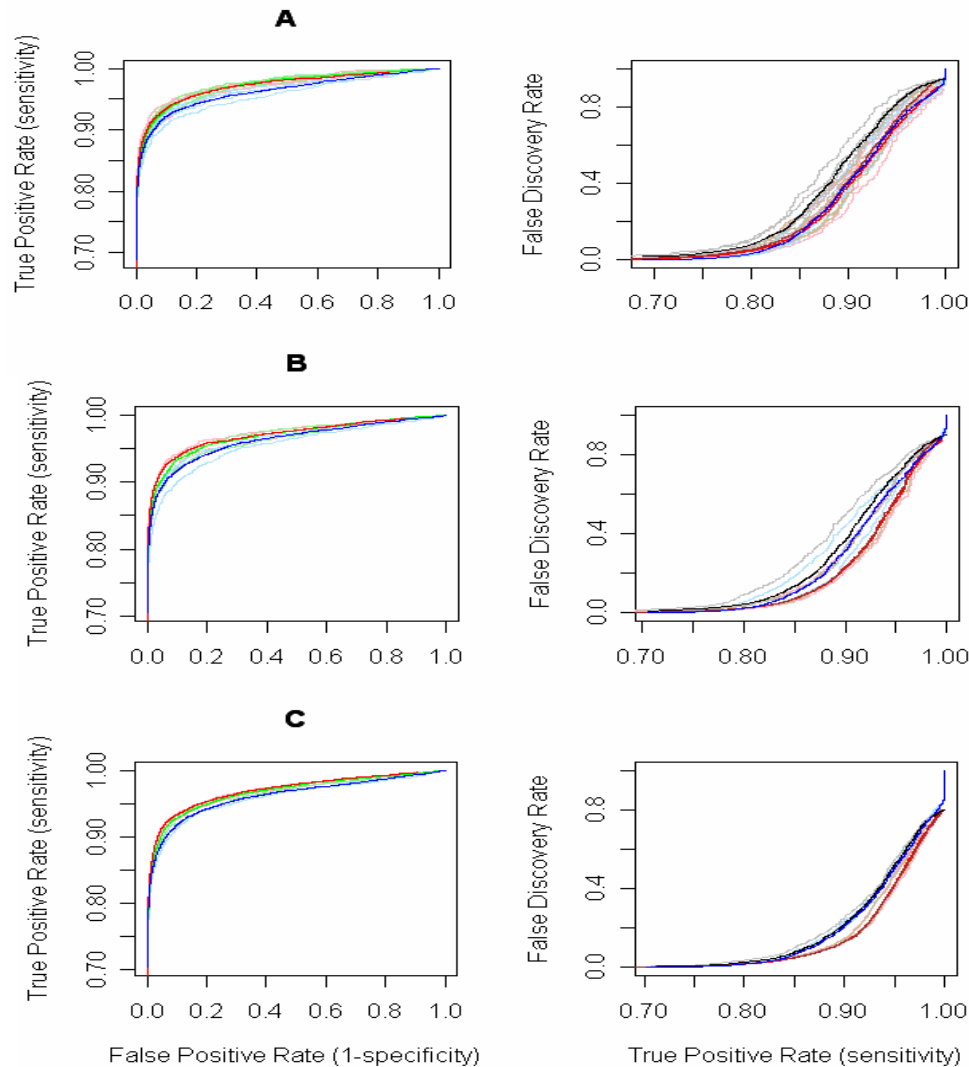


Figura 6.4: Resultados de simulações de dados com desvio na intensidade de cada spot é igual a 0.7. Os gráficos à esquerda mostram curvas ROC comparando o desempenho do algoritmo ISER2 (em azul), do SAM (em vermelho) e do teste t (em verde). As linhas mais fortes representam a média de 50 simulações, enquanto cada uma das linhas mais claras representam a média de 10 dessas simulações. Nos gráficos à direita encontram-se os FDRs reais obtidos pelos algoritmos ISER2 e SAM, bem como as estimativas dos mesmos encontradas pelos métodos, traçadas contra as sensibilidades dos algoritmos (True Positive Rates). O real FDR obtido pelo ISER encontra-se em preto, enquanto sua estimativa está em azul. O real FDR obtido pelo SAM encontra-se em marrom, enquanto sua estimativa está em vermelho. As linhas mais fortes representam a média de 50 simulações, enquanto cada uma das linhas mais claras representam a média de 10 dessas simulações. A) Simulações com 100 genes diferencialmente expressos. B) Simulações com 200 genes diferencialmente expressos. C) Simulações com 400 genes diferencialmente expressos.

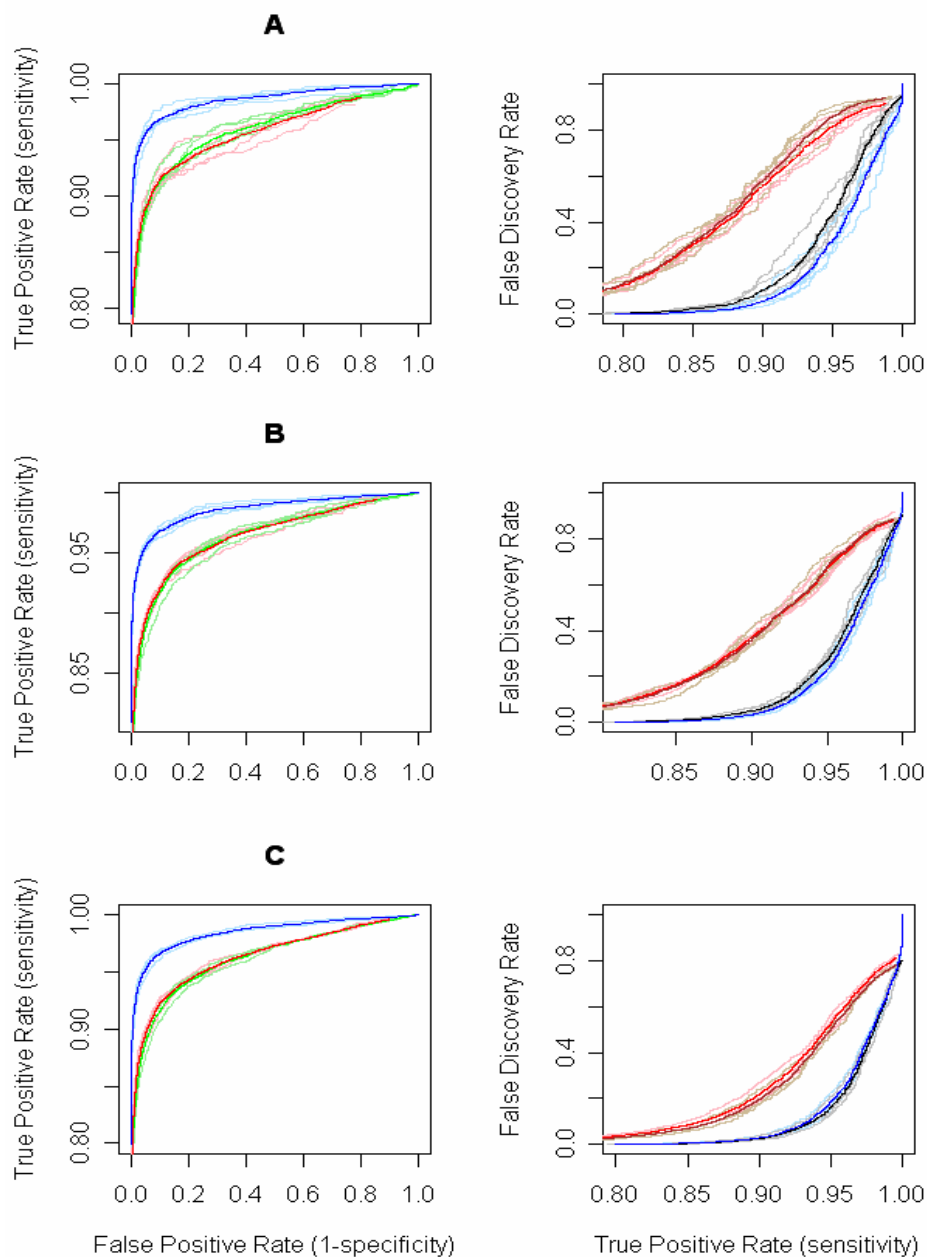


Figura 6.5: Resultados de simulações de dados com desvio na intensidade de cada spot é igual a 3. Os gráficos são análogos aos da Figura 6.4. A) Simulações com 100 genes diferencialmente expressos. B) Simulações com 200 genes diferencialmente expressos. C) Simulações com 400 genes diferencialmente expressos.

apresenta esse desvio abaixo de 0,7, porém em diversos genes esse valor é superior a 3, chegando até mesmo ao dobro disto. Nesse caso, a flexibilidade do algoritmo ISER2 será vantajosa em relação às demais estratégias de análise testadas, permitindo-o classificar corretamente os genes com maior ou menor variabilidade na intensidade de seus sinais.

A estratégia empregada pelo algoritmo ISER2 para estimar o FDR também foi comparada com a estimativa do FDR feita pelo algoritmo SAM. Em geral, as duas estratégias tiveram desempenhos semelhantes, ainda que o SAM tenha sido ligeiramente superior (Figuras 6.4 e 6.5). Porém, a estratégia utilizada pelo SAM para estimar o FDR é baseada em permutações dos dados, ou seja, os dados são dispostos em uma matriz com um gene por linha e os dados provenientes de cada amostra em cada coluna, e essas colunas são permutadas entre si (dados de diferentes *spots* vão ser dispostos em diferentes colunas, como se o sinal de cada *spot* fosse proveniente de uma amostra diferente). Com isso as colunas são aleatoriamente atribuídas ao grupo “tratado” ou ao grupo “controle”, e os dados são re-analisados, procurando assim obter uma aproximação da distribuição do valor do teste aplicado aos dados quando estes estão sob hipótese nula (a atribuição aleatória das colunas simula a hipótese nula). O problema dessa estratégia é que ela supõe que o número de dados é igual para todos os genes, pois caso contrário a matriz de dados teria linhas com números diferentes de entradas, e para dividi-la em colunas seria necessário trabalhar com valores faltantes. Em dados reais de arranjos de DNA, é freqüente a ocorrência de genes fixados em números diferentes de *spots*, como era o caso dos dados do experimento de deficiência de fosfato descrito no Capítulo 7. Além disso, devido

ao pré-processamento dos dados (descrito no Capítulo 2), os sinais provenientes de alguns *spots* dos arranjos podem ser eliminados das análises, gerando diferenças na quantidade de dados disponíveis para cada gene.

A estratégia utilizada pelo ISER2 para estimar o FDR depende apenas da distribuição global dos p-valores atribuídos aos genes, por isso ela apresenta a grande vantagem de não ser influenciada por diferenças no número de *spots* referentes a cada gene. Com isso, de maneira análoga a seleção de genes diferencialmente expressos, ainda que a estratégia adotada pelo ISER2 seja ligeiramente inferior à estratégia do SAM em situações ideais, em dados reais a sua flexibilidade torna-a mais robusta e, portanto, torna seu desempenho mais vantajoso. Além disso, como pode ser visto nas Figuras 6.4 e 6.5, a estimativa do FDR obtida pelo ISER2 é bastante acurada. Ainda que esta incorra numa pequena margem de erro (diferença entre o FDR estimado e o real), esta aparente desvantagem do algoritmo em relação ao SAM é compensada pela sua maior eficiência em selecionar os genes diferencialmente expressos, isto é, o melhor desempenho do SAM em estimar o FDR deixa de ser vantajoso frente aos seus valores de FDR sistematicamente maiores.

Concluimos assim que o algoritmo ISER2 é uma estratégia eficiente de análise de dados para a seleção de genes diferencialmente expressos, sendo capaz de estimar satisfatoriamente o FDR esperado nessa seleção. Devido a sua flexibilidade, esse algoritmo impõe menos condições aos dados que as outras estratégias de análise avaliadas (como as condições de os logaritmos das razões transformadas apresentarem distribuição aproximadamente normal, os diferentes sinais para um mesmo gene apresentarem variabilidades iguais ou o número de

dados disponíveis ser o mesmo para todos os genes) e, portanto, é mais eficiente para a análise de dados reais, onde tais condições frequentemente não são satisfeitas.

Capítulo 7: Efeito da deficiência de fósforo na expressão gênica de cana-de-açúcar

O fósforo é um elemento essencial à nutrição vegetal, sendo absorvido pelas plantas principalmente na forma de fosfato inorgânico (P_i). Devido a pouca disponibilidade desse nutriente em grande parte dos solos, principalmente nos solos ácidos das regiões tropicais e subtropicais, ao longo de sua evolução as plantas desenvolveram diversas estratégias de resposta à sua deficiência, incluindo mecanismos de regulação da expressão gênica. Em vista da importância econômica da cana-de-açúcar, a caracterização da resposta desta espécie à deficiência de fosfato é de grande interesse. A identificação de genes cuja expressão é regulada em face a pouca disponibilidade deste nutriente abre perspectivas para o melhoramento genético dessa espécie, seja através de procedimentos da genética clássica ou mesmo da engenharia genética. Esses resultados podem levar ao aprimoramento de linhagens mais adaptadas à deficiência de fosfato e portanto mais aptas à cultura em solos pobres em P_i , o que seria de grande valor para a agricultura.

Com o intuito de demonstrar a aplicação dos algoritmos de análise de dados desenvolvidos neste trabalho, os mesmos foram utilizados para identificar genes de cana-de-açúcar diferencialmente expressos em resposta à deficiência de fosfato. Para isso foi realizado um experimento com plantas cultivadas em hidroponia em diferentes regimes de fosfato, como descrito a seguir.

7.1 – Tratamento vegetal e extração de RNA

Plantas provenientes de culturas de meristemas de cana-de-açúcar (variedade SP80-3280 R11), na fase III de desenvolvimento, foram obtidas junto ao Centro de Tecnologia Canavieira (Piracicaba, SP). O cultivo e tratamento das plantas foi feito em colaboração com o grupo de pesquisa do professor Antônio Figueira (CENA-USP). Estabelecemos sistemas de soluções nutritivas com fosfato suficiente ou desprovida de fosfato, para cultivo em hidroponia (Hogland e Arnon, 1953; Almeida, 2002; Figura 7.1). Após aclimação as plantas foram divididas em dois grupos: um grupo controle (+P) foi cultivado em solução nutritiva completa por todo o experimento, e o outro grupo (-P) recebeu solução nutricional desprovida de fosfato. Amostras de tecidos de raízes foram extraídas de doze plantas de cada um dos dois grupos (+P e -P), em quatro momentos do experimento: 6, 12, 24 e 48 horas (plantas diferentes para cada ponto experimental).

Dentre as doze amostras de raízes coletadas a cada tempo e cada tratamento, selecionamos oito, de maneira a descartar as plantas cujos pesos frescos totais eram muito diferentes dos demais. Procuramos com isso restringir o trabalho a amostras provenientes de um conjunto mais homogêneo de plantas. Essas oito amostras foram maceradas individualmente e então agrupadas em quatro grupos de duas amostras. A cada grupo foi formado um "pool" contendo a mesma quantidade de tecido macerado de cada uma das duas plantas, para extração de RNA. O RNA total foi extraído separadamente de cada "pool" de amostras de raiz.



Figura 7.1 – Plantas no sistema de hidroponia.

7.2 – Híbridações de microarranjos de DNA

Com o intuito de identificar genes responsáveis pela coordenação dos mecanismos de resposta da cana-de-açúcar à deficiência de fosfato, as amostras de RNA obtidas foram utilizadas em hibridações de microarranjos de cDNA confeccionados pelo grupo de pesquisa da professora Gláucia Souza, o Laboratório de Transdução de Sinal (IQ-USP). Estes microarranjos contêm cDNAs obtidos a partir de ESTs de 1557 genes selecionados, sendo em sua maioria responsáveis por fatores de transdução de sinais incluídos no “*sugarcane signal trasduction (SUCAST) catalogue*” (Souza et al, 2001). Estes fatores são responsáveis pela sinalização celular que irá desencadear as respostas fisiológicas da planta aos estresses a que ela seja submetida, por isso a

caracterização do seu perfil de expressão durante a deficiência de fosfato é extremamente importante para elucidar a resposta da cana-de-açúcar a esse estresse nutricional. Os demais genes presentes nos microarranjos são proteínas de função desconhecida. Os arranjos foram fixados em lâminas de vidro (*slides*), sendo que a cada lâmina dois arranjos idênticos (réplicas técnicas) são depositados lado a lado. Com isso cada gene está representado em pelo menos dois *spots* nos slides, sendo que a maioria dos genes está representada em quatro ou mais *spots*, pois já se encontra fixada mais de uma vez em cada réplica do arranjo.

As hibridações de microarranjos de DNA foram realizadas em colaboração com o grupo da professora Gláucia Souza (Laboratório de Transdução de Sinal, IQ-USP). Sondas obtidas a partir de amostras de RNA de raízes coletadas no mesmo tempo experimental, sendo uma de plantas privada de fosfato e outra de plantas controle, foram marcadas com fluoróforos diferentes (*Cy3* e *Cy5*) e hibridadas no mesmo microarranjo. Desta forma comparamos sempre a expressão gênica de plantas desprovidas de fosfato com a das respectivas plantas cultivadas em solução completa, evitando que variações circadianas na expressão dos genes fossem confundidas com respostas a deficiência. A cada ponto experimental (6, 12, 24 e 48 h) foram realizadas duas hibridações de microarranjos, com sondas obtidas a partir de amostras de RNA independentes (réplicas biológicas). A associação entre fluoróforos e regimes de fosfato foi invertida de uma réplica para outra, isto é, em uma das hibridações marcamos a sonda proveniente de plantas privadas de fosfato com o fluoróforo *Cy3* e a sonda proveniente de plantas controle com o *Cy5*, enquanto na outra hibridação marcamos a sonda controle

com Cy3 e a sonda proveniente de plantas privadas de fosfato com Cy5. Desta forma, caso algum gene apresentasse maior afinidade por um fluoróforo que pelo outro, isso não geraria uma tendência nos dados obtidos para esse gene que pudesse ser interpretada erroneamente como sinal de expressão diferencial do mesmo.

As imagens provenientes das hibridações (Figura 7.2) foram quantificadas com o programa ArrayVision (APBiotech, Brasil). Este programa obtém, para cada ponto de fixação do DNA no arranjo, um valor proporcional a fluorescência total emitida pelo ponto e um valor relativo à emissão de fundo (*background*), proveniente da região ao redor do mesmo. As diferenças entre esses valores (*sinal-background*) foram então exportadas pelo programa e utilizadas nas análises subseqüentes (Figura 7.3). Os dados provenientes de cada uma das réplicas do arranjo presentes num mesmo *slide* são exportados em arquivos separados, desta forma possibilitando que sejam analisados separadamente.

Os dados obtidos nesse experimento foram incluídos em um artigo sobre os efeitos de diversos estresses e hormônios na expressão gênica de cana-de-açúcar, um resultado do projeto SUCEST-FUN (citado na Introdução): “Signal transduction-related responses to phytohormones and environmental challenges in sugarcane”, Flávia Riso Rocha, Flávia Stal Papini-Terzi, Milton Yutaka Nishiyama-Jr, Ricardo Zorzetto Nicoliello Vêncio, Renato Vicentini, Rodrigo Drummond Couto Duarte, Vicente Eugenio de Rosa Jr, Fabiano Vinagre, Carla Barsalobres, Ane Hackbart Medeiros, Fabiana Aparecida Rodrigues, Eugênio César Ulian, Sônia Marli Zingaretti di Mauro, João Antonio Galbiatti, Raul Santin Almeida, Antonio Vargas de Oliveira Figueira, Adriana Silva Hemerly, Marcio Castro Silva-Filho,

Marcelo Menossi e Gláucia Mendes Souza. O artigo foi submetido para publicação na revista *BMC Genomics*. Os dados deste e dos demais experimentos descritos no artigo foram analisados com algoritmos desenvolvidos pelo grupo de pesquisa da professora Gláucia Souza (Laboratório de Transdução de Sinal, IQ-USP), porém o algoritmo ISER também foi aplicado aos dados e os resultados obtidos foram acrescentados ao material suplementar da publicação.

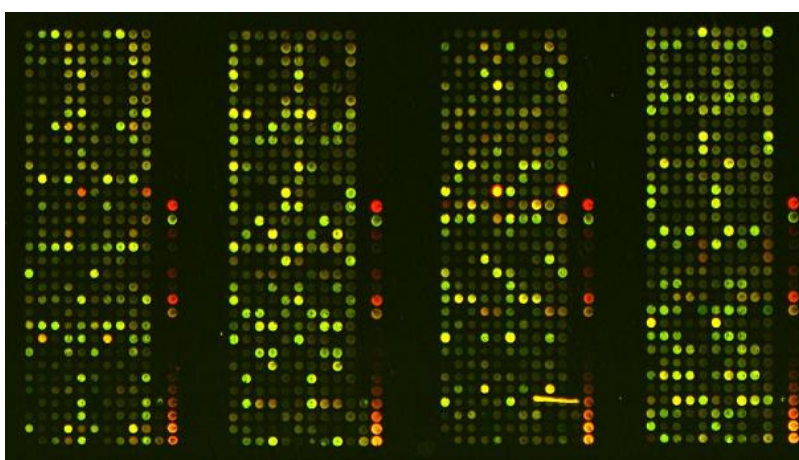


Figura 7.2: Imagem obtida após a leitura dos sinais fluorescentes de um microarranjo de DNA utilizado no experimento. A cor verde indica, a princípio, genes cuja expressão foi reprimida em resposta à deficiência de fosfato, enquanto a cor vermelha indica genes cuja expressão foi induzida. A cor amarela indica genes cuja expressão não foi alterada no experimento, enquanto pontos sem cor indicam ausência de expressão. Porém, vale observar que a seleção de genes induzidos ou reprimidos não pode ser feita baseada apenas na observação visual da imagem obtida, dependendo da normalização dos dados e de testes estatísticos.

7.3 – Análise preliminar dos dados

Os dados obtidos de todas as hibridações, já subtraídos do *background*, foram dispostos em gráficos de dispersão e gráficos MA (figura 8.3). Verificou-se que todas as hibridações apresentam uma tendência central dos dados, evidenciada por uma grande correlação entre tratado e controle ($R^2 > 0.84$ em todas as hibridações, $R^2 > 0.9$ em 81% das mesmas), o que confirma a suposição de que a expressão da maior parte dos genes não varia em resposta a deficiência de fosfato.

Em cada uma das hibridações, foram descartados os *spots* que apresentavam sinais iguais a zero em ambos os canais (tratado e controle) ou que apresentavam sinal igual a zero em apenas um canal, porém no outro canal apresentavam um valor da razão sinal/ruído menor que 1,96. Os valores das razões sinal/ruído, a cada canal de cada spot, são exportados pelo software utilizado para quantificar as membranas, sendo definidos como a média do sinal obtido no spot dividida pelo desvio dos valores dos sinais obtidos nos diversos pontos (*pixels*) que compõem a região circular em torno do *spot*, onde é computado o background. Esta razão mede o quanto o sinal no *spot* se destaca da flutuação do background ao seu redor. Assim, se supusermos que a distribuição dos sinais nos pontos em torno do *spot* tem distribuição normal e que o sinal no *spot* teria a mesma distribuição caso não houvesse expressão do respectivo gene, a chance de essa razão ser maior que 1,96 devido apenas a flutuações aleatórias seria menor que 5%, isto é, uma razão sinal/ruído acima de 1,96 indicaria que o sinal no spot é significativamente maior que zero. Esta estratégia procura evitar

que consideremos diferencialmente expresso um gene cujos respectivos *spots* apresentam sinais iguais a zero em um canal e sinais pequenos, maiores que zero por mero acaso, no outro canal. Vale observar que *spots* que apresentem sinal igual a zero em um dos canais, sinal diferente de zero no outro e não sejam eliminados durante a filtragem dos dados são invariavelmente marcados pelo ISER como induzidos - caso o canal cujo sinal é igual a zero seja o controle - ou reprimidos - caso o sinal do tratado seja igual a zero. Desta forma, se esse padrão se repetir na maioria dos *spots* referentes a algum gene, este pode ainda ser selecionado como diferencialmente expresso.

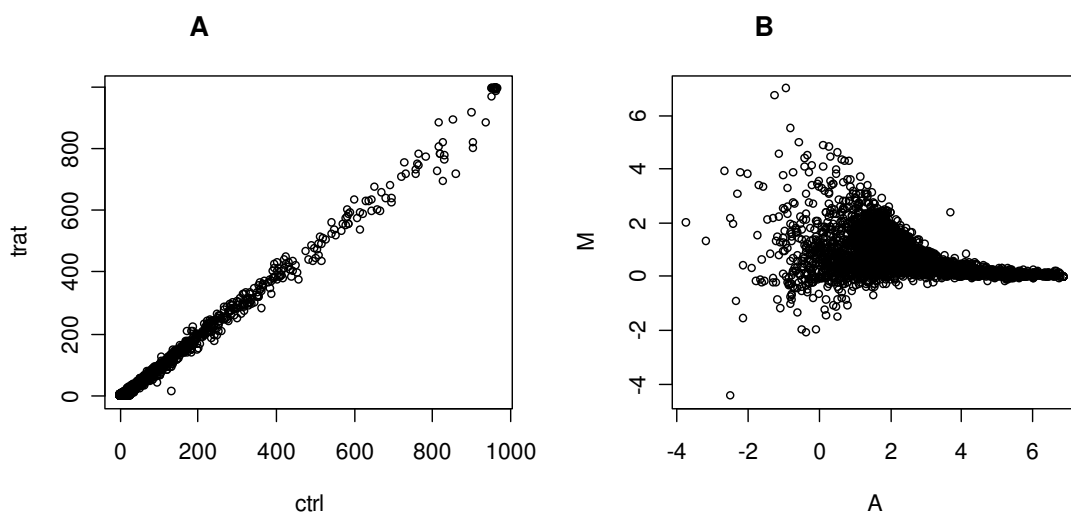


Figura 7.3: A) Gráfico de dispersão dos dados de uma hibridação com amostras de 12 h de deficiência de fosfato, $R^2=0.994$; B) Os mesmos dados no gráfico conhecido como MA plot - dispersão dos logaritmos das razões (M) em função das intensidades médias dos genes (A, também em escala logarítmica)

Os dados foram então normalizados com o algoritmo *lowess* (Cleveland, 1979) e analisados com os algoritmos ISER e ISER2, descritos nos Capítulos 5 e 6. Como já foi dito, esses algoritmos têm embutida a transformação de dados, além de uma segunda normalização dos mesmos.

7.4 – Resultados do algoritmo ISER

A cada ponto experimental, o algoritmo ISER foi aplicado independentemente nos dados provenientes de cada réplica biológica e de cada réplica técnica (arranjos replicados lado a lado em cada *slide*). A cada análise, *spots* diferentes contendo um mesmo gene também foram tratados independentemente. Uma margem de erro de 10% foi adotada, o que significa que, a cada subconjunto dos dados determinado pela janela deslizante do algoritmo, os valores de corte definidos para as razões transformadas, além dos quais os *spots* foram marcados como induzidos ou reprimidos, correspondiam respectivamente ao quinto e 95° quantis da distribuição normal.

Os genes selecionados como diferencialmente expressos foram aqueles identificados coerentemente como induzidos ou reprimidos em mais de 66% (2/3) de todos os seus respectivos *spots*. Além disso, exigiu-se que em cada réplica biológica esses genes fossem confirmados como induzidos ou reprimidos em mais de 50% dos respectivos *spots* (provenientes de ambas as réplicas técnicas, isto é, de todo o *slide*). Esta estratégia de seleção garante que os genes selecionados tenham sido coerentemente identificados como diferencialmente expressos na maioria de seus *spots*, além de evitar que um gene cujos *spots* foram eliminados

de uma das réplicas (por apresentarem sinal fraco), seja selecionado apenas por seu perfil de expressão na outra réplica. Considerando a margem de erro adotada na seleção de cada *spot*, a probabilidade de cada gene ser selecionado por engano torna-se inferior a 1,5%. Na verdade, essa estratégia de seleção é bastante restritiva e acaba por identificar um pequeno número de genes diferencialmente expresso. Os 19 genes selecionados encontram-se na Tabela 7.1.

ID privado	Categoria	Razão 6h	Razão 12h	Razão 24h	Razão 48h
168282	Protein kinases	3,602	0,889	1,125	0,657
174000	Stress	0,879	1,267	0,895	1,132
190593	Stress	0,961	1,327	0,899	0,630
184571	Transcription	0,908	1,218	0,896	0,775
166746	Pathogenicity	0,881	1,301	0,895	1,011
168387	Stress	1,396	1,639	1,636	1,054
187156	Stress	1,331	0,891	0,958	1,636
178047	Ubiquitination	0,649	0,580	1,013	8,445
182391	Hormone biosynthesis	0,980	1,251	0,895	0,649
184108	Pathogenicity	0,671	0,628	0,889	0,368
177059	Stress	0,731	0,728	0,347	0,572
179917	Receptors	1,014	0,593	0,582	1,143
184256	Transcription	0,514	0,752	0,895	0,779
176095*	Transcription	0,557	0,755	0,930	0,821
188864	Stress	0,576	0,741	0,985	1,120
191164	Hormone related	0,365	0,727	0,741	1,187
172386**	Others	0,472	1,009	0,847	0,658
176519	No matches	0,608	0,631	1,097	0,772
186421	No matches	0,592	0,664	0,902	0,467

Tabela 7.1: Genes identificados como diferencialmente expressos pelo algoritmo ISER. ID privado: identificador de cada gene nos bancos de dados do projeto SUCEST. Categoria: Classificação de acordo com o catálogo SUCAST. As razões marcadas em **laranja** foram identificadas como significativas de indução do gene, enquanto as marcadas em **verde** são significativas de repressão.

*Confirmado por *real-time PCR*. **Testado por *real-time PCR*, porém não confirmado.

7.5 – Resultados do algoritmo ISER2

O algoritmo ISER2 foi aplicado independentemente aos dados de cada ponto experimental, adotando-se a estratégia baseada em estatísticas de ordem, descrita no capítulo 6, para associar p-valores aos genes. As distribuições dos p-valores mostraram-se próximas de misturas beta-uniforme-beta (Figura 7.4), e esse modelo foi utilizado para estimar os q-valores associados aos genes, como descrito no capítulo 6. Por fim, foram selecionados os genes que apresentavam q-valores menores que 5%, esperando-se portanto um FDR de apenas 5% na seleção final. Em comparação com os resultados do ISER (seção 7.4), um número muito maior de genes foi selecionado: 166, indicados na Tabela 7.2 (A a D) e identificados de acordo com suas categorias no catálogo SUCAST (Souza *et al.*, 2001). Vale observar que todos os genes identificados pelo algoritmo ISER como diferencialmente expressos também foram selecionados pelo ISER2. Porém, como o ISER seleciona muito menos genes, sua utilização como estratégia de seleção de genes diferencialmente expressos em experimentos com réplicas, ainda que eficiente, é excessivamente restritiva.

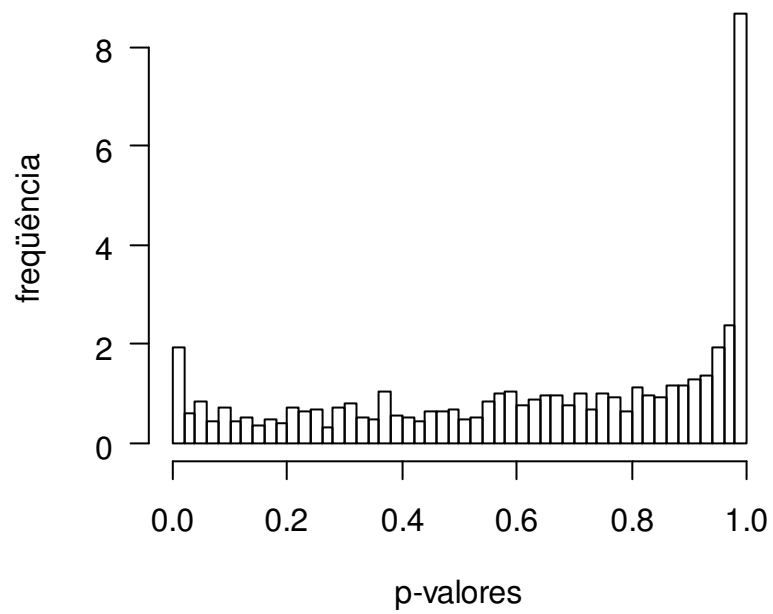


Figura 7.4 – Histograma dos p-valores para genes reprimidos em 24 h de deficiência de fosfato.

ID privado	Categoria	Razão 6h	Razão 12h	Razão 24h	Razão 48h
171407	Transcription factor	0,82	1,53	0,94	1,11
177825	Transcription factor	1,01	1,61	0,66	0,97
184571	Transcription factor	0,90	1,23	0,89	0,80
166973	Transcription factor	1,00	1,00	1,07	2,16
180808	Transcription factor	0,96	0,91	0,85	2,93
183414	Transcription factor	0,71	0,72	1,07	1,44
174870	Transcription factor	0,61	0,94	1,00	0,94
178172	Transcription factor	0,59	0,75	0,90	1,21
177939	Transcription factor	0,68	0,93	0,83	1,16
174393	Transcription factor	1,14	0,99	0,93	0,75
184256	Transcription factor	0,51	0,74	0,86	0,77
181308	Transcription factor	0,43	5,82	1,30	0,17
176095*	Transcription factor	0,55	0,77	0,97	0,81
173724	Transcription factor	0,64	1,61	1,27	1,01
183500	Transcription factor	0,71	1,08	0,98	1,14
177011	Transcription factor	0,70	0,41	5,16	0,96
172608	Transcription factor	0,54	1,15	1,06	0,97
184332	Transcription factor	0,81	0,84	0,95	0,69
187767	Transcription factor	0,84	0,80	0,75	0,89
187073	Transcription factor	1,21	0,73	0,78	1,21
167232	Transcription factor	0,86	0,90	0,71	1,16
187023	Transcription factor	0,97	0,59	0,73	0,94
173107	Transcription factor	1,07	0,96	0,95	0,69
175717	Transcription factor	1,15	0,88	1,04	0,10
186449	Transcription factor	1,00	0,95	0,90	0,81
170582*	Stress	1,94	1,22	0,77	1,16
190593	Stress	0,95	1,32	0,89	0,65
174000	Stress	0,88	1,28	0,89	1,25
169206	Stress	0,93	1,10	0,84	3,04
168387	Stress	1,57	1,25	1,60	1,19
176934	Stress	1,14	1,07	2,05	0,87
181102	Stress	0,69	0,94	1,45	1,48
187156	Stress	1,31	0,91	0,98	1,69
179423	Stress	1,16	1,05	0,96	1,80
173478	Stress	0,78	1,06	1,33	1,49
191826	Stress	0,70	1,01	1,18	1,03
178338	Stress	0,66	0,94	0,91	1,32
174153	Stress	0,78	1,00	1,00	0,98
175185	Stress	0,59	0,93	0,98	0,96
188864	Stress	0,57	0,74	0,99	1,06
178260	Stress	0,62	1,06	1,72	1,02
176597*	Stress	0,59	1,10	0,64	1,03
177059	Stress	0,82	0,81	0,35	0,66
189806	Stress	0,81	0,24	0,84	0,69
173652	Stress	1,25	1,17	0,73	0,94

Tabela 7.2-A: Genes identificados como diferencialmente expressos pelo algoritmo ISER2 - fatores de transcrição e genes relacionados à resposta a estresse. ID privado: identificador de cada gene nos bancos de dados do projeto SUCEST. Categoria: Classificação de acordo com o catálogo SUCAST. As razões marcadas em **laranja** foram identificadas como significativas de indução do gene, enquanto as marcadas em **verde** são significativas de repressão.

*Confirmado por *real-time PCR*.

ID privado	Categoria	Razão 6h	Razão 12h	Razão 24h	Razão 48h
173199	Protein kinase	1,90	0,88	0,89	0,35
168282	Protein kinase	5,17	0,89	1,08	0,75
183098	Protein kinase	0,80	0,70	1,04	1,88
187323	Protein kinase	0,75	1,48	0,90	2,51
192864	Protein kinase	1,06	1,22	1,22	2,24
183482	Protein kinase	0,64	0,99	0,95	1,11
185628	Protein kinase	1,09	0,70	0,97	0,64
181471	Protein kinase	0,63	1,93	1,03	0,87
184602	Protein kinase	0,67	1,19	0,85	1,47
181839	Protein kinase	0,63	0,43	1,01	0,75
169047	Protein kinase	0,12	1,05	0,85	0,87
179771	Protein Phosphatase	0,75	0,84	0,90	1,40
172820	Receptor	0,70	0,71	1,02	0,90
170454	Receptor	1,15	2,31	1,26	0,90
173150	Receptor	0,69	1,13	1,48	0,90
175630	Receptor	0,62	19,18	0,83	1,33
169318	Receptor	0,55	0,72	0,96	1,54
173273	Receptor	1,08	0,47	0,66	0,98
166730	Receptor	0,82	0,73	0,67	0,23
166394	Receptor	0,73	1,16	1,00	1,05
185831	Receptor	0,73	1,04	0,99	0,98
181293	Receptor	1,22	1,67	1,24	0,93
180170	Receptor	0,75	0,89	0,99	1,09
184974	Receptor	0,61	0,93	1,05	0,94
173789	Receptor	0,76	1,77	0,92	1,51
165945	Receptor	1,07	0,69	0,91	1,08
179917	Receptor	1,15	0,10	0,46	1,33
166621	Hormone biosynthesis	0,74	3,44	0,45	5,75
180272	Hormone biosynthesis	0,68	1,85	0,91	6,41
182391	Hormone biosynthesis	0,98	1,25	0,89	0,64
181090	Hormone biosynthesis	1,21	1,20	1,64	1,29
187863	Hormone related	0,63	0,83	0,90	1,38
177821	Hormone related	0,64	0,84	0,98	1,70
191164	Hormone related	0,37	0,76	0,73	1,17
186446	Pathogenicity	1,06	2,42	0,90	0,83
166746	Pathogenicity	0,89	1,31	0,89	0,96
173713	Pathogenicity	0,96	74,35	0,76	7,53
170121	Pathogenicity	0,75	1,03	0,92	1,03
184108	Pathogenicity	0,66	0,73	0,87	0,62
183222	Pathogenicity	0,38	0,83	0,93	1,06
188292	Pathogenicity	0,99	1,11	0,80	1,30
196255	Pathogenicity	0,98	0,65	1,21	0,49
180519	Pathogenicity	0,01	1,23	0,73	1,05
181855	Pathogenicity	0,25	0,54	0,65	0,10
170004	Pathogenicity	1,46	1,52	0,59	1,17
165908	Pathogenicity	0,40	1,05	0,70	0,80
191072	Pathogenicity	0,70	0,71	1,14	1,19
175426	Pathogenicity	1,10	0,61	0,83	3,48

Tabela 7.2-B: Genes identificados como diferencialmente expressos pelo algoritmo ISER2 - quinases, fosfatases, receptores e genes relacionados a hormônios e a resposta a patógenos. ID privado e Categoria: como na Tabela 7.2-A. As razões marcadas em **laranja** foram identificadas como significativas de indução do gene, enquanto as marcadas em **verde** são significativas de repressão.

ID privado	Categoria	Razão 6h	Razão 12h	Razão 24h	Razão 48h
167709	Ubiquitination	1,09	0,99	1,10	1,44
170200	Ubiquitination	0,96	0,95	0,91	1,54
178047	Ubiquitination	0,65	0,72	1,02	16,36
174244	Ubiquitination	0,76	0,95	1,20	1,38
171436	Ubiquitination	0,66	0,87	1,17	1,28
190350	Ubiquitination	0,64	0,91	1,26	1,40
174009	Ubiquitination	0,70	0,85	1,13	1,26
189367	Ubiquitination	0,66	0,78	1,00	0,77
183832	Ubiquitination	0,74	0,90	1,10	1,17
188484	Ubiquitination	0,73	0,99	1,23	1,10
176387	Ubiquitination	0,70	0,93	1,17	1,13
172901	Ubiquitination	0,70	0,96	1,09	1,09
166276	Ubiquitination	0,75	0,94	1,14	1,18
191576	Ubiquitination	0,69	0,95	1,32	1,09
191660	Ubiquitination	0,68	0,93	1,23	1,67
172927	Ubiquitination	0,86	0,00	0,75	1,32
176410	Small GTPase	1,43	0,48	0,90	1,60
187376	Small GTPase	1,34	2,92	1,16	0,89
178063	Small GTPase	0,72	0,84	0,86	1,02
190103	Small GTPase	0,99	0,62	0,83	1,04
181203	Small GTPase	0,93	0,45	0,73	0,72
184018	Calcium	1,80	1,15	0,84	1,14
166632	Calcium	0,80	1,01	0,97	1,32
170332	Calcium	0,67	0,96	1,08	1,10
171469	Calcium	0,64	2,20	1,12	1,03
184690	Carbohydrate metabolism	0,68	0,95	0,76	1,81
182947	Carbohydrate metabolism	1,12	0,92	0,94	2,22
168238	Adapters	0,99	0,91	1,03	1,89
180561	Development	0,82	0,80	0,84	0,71
206297	Two component	0,65	1,06	1,07	1,10
181668	Others	1,32	0,99	1,06	0,88
178299	Others	1,04	1,33	1,04	0,99
168151	Others	0,96	0,96	1,09	1,37
181937	Others	1,04	0,92	1,10	1,59
185279	Others	0,64	1,27	1,02	0,98
188486	Others	0,65	0,98	0,99	1,19
188387	Others	0,60	0,88	0,86	0,69
172386**	Others	0,47	1,01	0,86	0,64
185706	Others	0,60	0,41	0,51	0,82
168638*	Others	0,58	0,64	0,85	1,14
166567	Others	0,68	0,93	0,89	1,27
187959	Others	0,81	0,86	0,99	1,21
176870	Others	0,85	0,73	0,90	1,52
173962	Others	0,39	1,12	0,52	0,80

Tabela 7.2-C: Genes identificados como diferencialmente expressos pelo algoritmo ISER2 - ubiquitinas, GTPases, genes relacionados ao metabolismo de carboidratos e outros. ID privado e Categoria: como na Tabela 7.2-A. As razões marcadas em **laranja** foram identificadas como significativas de indução do gene, enquanto as marcadas em **verde** são significativas de repressão.

*Confirmado por *real-time PCR*. **Testado por *real-time PCR*, porém não confirmado.

ID privado	Categoria	Razão 6h	Razão 12h	Razão 24h	Razão 48h
188927	Unknown protein	1,30	3,01	1,00	1,44
166377*	Unknown protein	0,91	1,45	1,07	0,83
179943	Unknown protein	1,88	1,23	1,01	6,55
188472	Unknown protein	1,21	0,95	1,02	1,89
186358	Unknown protein	1,01	0,98	0,97	1,99
170072	Unknown protein	0,82	1,12	1,29	1,53
186397	Unknown protein	0,78	0,92	0,82	0,96
184573	Unknown protein	0,68	0,57	1,13	0,71
187729	Unknown protein	0,30	1,67	0,86	0,33
184402	Unknown protein	1,35	1,26	0,72	1,32
188544	Unknown protein	1,33	1,13	0,51	1,00
187871	No matches	1,49	1,06	1,04	0,97
166809	No matches	1,00	1,21	0,95	0,64
173519	No matches	1,00	1,31	0,98	1,02
176685	No matches	0,90	1,48	1,02	1,00
191337	No matches	1,25	1,11	0,92	3,82
192076	No matches	1,21	1,00	0,93	3,98
183451	No matches	0,78	1,00	1,05	1,06
183027	No matches	0,71	0,95	1,06	1,34
181625	No matches	0,63	1,24	0,90	1,05
188289	No matches	0,63	0,86	0,91	1,17
176519	No matches	0,61	0,65	1,06	0,80
186195	No matches	0,59	0,89	0,99	1,00
186421	No matches	0,59	0,66	0,90	0,53
191260	No matches	0,61	0,76	0,95	0,97
189818	No matches	0,82	0,47	0,47	0,63
187705	No matches	0,67	0,73	0,74	0,89
191112	No matches	0,65	0,41	0,71	0,19
188383	No matches	1,13	1,00	1,05	0,56

Tabela 7.2-D: Genes identificados como diferencialmente expressos pelo algoritmo ISER2 - proteínas desconhecidas e genes sem homólogos conhecidos. ID privado e Categoria: como na Tabela 7.2-A. As razões marcadas em **laranja** foram identificadas como significativas de indução do gene, enquanto as marcadas em **verde** são significativas de repressão.

*Confirmado por *real-time PCR*.

7.6 – Confirmação da expressão diferencial por *real-time PCR*

Entre os genes selecionados como diferencialmente expressos pelo algoritmo ISER2, seis foram utilizados para confirmação dos resultados via *real-time PCR*, dois dos quais haviam sido também selecionados pelo algoritmo ISER. As reações de PCR foram feitas a partir de uma amostra independente de RNA,

utilizando um gene que codifica uma proteína 14-3-3 (SAS SCCCLR1048F12.g) como referência para normalização dos dados, o qual aparentemente não tem a expressão alterada pela deficiência de fosfato. Cada um desses seis genes teve sua expressão testada por *real-time PCR* em apenas um ponto experimental, no qual fora identificado como induzido ou reprimido em resposta a deficiência de fosfato, com exceção do gene induzido por fermento, que fora selecionado como reprimido em 6 e 24 horas de experimento e teve sua expressão avaliada nesses dois pontos. Desta forma, foi realizado um total de sete reações de *PCR*.

Os dados de *PCR* foram analisados com o método $2^{-\Delta\Delta Ct}$ (Livak & Schmittgen, 2001), e os resultados encontram-se na Tabela 7.3. Em apenas um caso a reação de *real-time PCR* não corroborou os resultados das análises dos dados de arranjos: a repressão do gene envolvido na estabilidade de RNA, em 6 horas de experimento. Nesse caso, os dados de *real time PCR* apresentaram grande variabilidade, como pode ser verificado pela amplitude do respectivo intervalo de confiança na Tabela 7.3, e o experimento não foi conclusivo. Além disso, ainda que esse gene não tenha sido reprimido na amostra analisada por *PCR*, isso pode dever-se à variabilidade biológica, e não significa que o gene não fora reprimido nas réplicas analisadas com arranjos de DNA. Por outro lado, as seleções de genes diferencialmente expressos obtidas com o algoritmo ISER2 tiveram FDR estimado em 5%, portanto a presença de um falso-positivo entre as sete expressões diferenciais testadas por *PCR* não invalida o método. Com isso, concluímos que os resultados das confirmações por *real-time PCR* encontram-se de acordo com o esperado, corroborando assim a eficiência do algoritmo ISER2 na seleção de genes diferencialmente expressos.

ID privado	Categoria	Resultado ISER2	Razão nos arranjos	Razão no PCR	Intervalo de confiança PCR*	
176095	Transcription factor	reprimido 6h	0,549	0,319	0,157	0,648
170582	Stress	induzido 6h	1,935	3,260	2,549	4,170
176597	Stress	reprimido 6h	0,592	0,406	0,325	0,508
176597	Stress	reprimido 24h	0,643	0,606	0,527	0,698
172386**	Others	reprimido 6h	0,468	1,079	0,308	3,785
168638	Others	reprimido 6h	0,576	0,276	0,175	0,435
166377	Unknown protein	induzido 12h	1,452	3,677	2,609	5,181

Tabela 7.3: Confirmação de alguns genes selecionados a partir dos dados de arranjos de DNA. Comparação dos resultados obtidos dos arranjos e das reações de real-time PCR. ID privado: identificador de cada gene nos bancos de dados do projeto SUCEST. Categoria: Classificação de acordo com o catálogo SUCAST. * Intervalo de confiança de 95% para a razão de expressão entre tratado (-P) e controle (+P). ** Único gene não confirmado como diferencialmente expresso.

7.7 – Discussão dos resultados

O algoritmo BLASTX (Altschul *et al.*, 1997) foi utilizado para buscar, nos bancos de dados do *National Center for Biotechnology Information* (NCBI, www.ncbi.nlm.nih.gov), genes já descritos com seqüências similares aos genes de cana-de-açúcar identificados como diferencialmente expressos neste experimento. Desta forma, a provável função da maioria dos genes selecionados (Tabelas 7.1 e 7.2) foi identificada. Os dados obtidos mostram que a cana-de-açúcar aciona uma ampla gama de respostas à deficiência de fosfato.

De acordo com a literatura, genes que respondem à deficiência de fosfato estão envolvidos em múltiplas vias metabólicas, o que implica na existência de um complexo sistema de regulação de P_i em plantas (Wu *et al.*, 2003). A análise de padrões de expressão em *Arabidopsis* durante a deficiência de fosfato (Hammond *et al.*, 2003) identificou a presença de pelo menos dois programas transcricionais de resposta. Um programa transiente, ativado durante os primeiros estágios da

deficiência, envolve genes relacionados à resposta geral a estresses (cuja expressão diferencial foi detectada após 4h de deficiência). Os genes relacionados especificamente à deficiência de fosfato são ativados num segundo programa transcricional, após um período prolongado de deficiência (identificados após 28 ou 100 h de deficiência no experimento de Hammond *et al.*, 2003), alterando a morfologia, fisiologia e o metabolismo da planta. Comparando esses resultados com os obtidos em cana-de-açúcar, nota-se que nessa espécie também há uma distinção clara entre a resposta primária à deficiência, ocorrida após 6 ou 12 h do experimento, onde a maioria dos genes diferencialmente expressos é reprimida, e a resposta tardia, após 24 h, quando o maior número de genes é induzido.

A expressão de genes de cana-de-açúcar similares a genes de outras espécies que codificam diversos fatores de transcrição foi afetada pela deficiência de fosfato. Entre eles, destaca-se a presença de quatro fatores WRKY, sendo três deles reprimidos e um induzido durante o experimento. Esses fatores de transcrição são relacionados a respostas a estresse (Eulgem *et al.*, 2000), e elementos cis-reguladores reconhecidos por eles foram identificados em 41,2% das regiões promotoras dos genes de raiz de *Arabidopsis* responsivos a deficiência de fosfato (Hammond *et al.*, 2004). Três genes cujos similares em outras espécies codificam fatores de transcrição com domínio AP2 foram reprimidos pela deficiência de fosfato, contrastando com os resultados de Wu *et al.* (2003), que detectou um fator com esse domínio induzido em raízes de *arabidopsis*. AP2 é um domínio de ligação a DNA altamente conservado em fatores de transcrição do grupo EREBP, muitos dos quais são reconhecidamente responsivos a estresses ambientais (Weigel, 1995). Nossos dados indicam que os

fatores AP2 podem fazer parte de uma via distinta de sinalização da deficiência de fósforo em cana-de-açúcar. Outro ponto em comum entre nossos dados e a literatura é a repressão de um gene cujo homólogo em milho codifica um fator de transcrição da família MYB. Essa classe de proteínas atua na transcrição de diversos genes e seus membros apresentam padrões de expressão bastante diversos em resposta à deficiência de P em *Arabidopsis*, sendo alguns induzidos e outros reprimidos (Wu *et al.*, 2003). É curioso que o promotor do gene homólogo de aveia (*MybHv1*) já foi caracterizado, sendo expresso somente no ápice radicular (Wissenbach *et al.*, 1993). Alterações no perfil de expressão do ápice radicular poderiam estar relacionadas às alterações morfológicas que o sistema radicular sofre em resposta aos baixos níveis de P. Além disso, dois fatores de transcrição já descritos como reguladores eucarióticos do metabolismo de fósforo pertencem à família MYB: Psr1 de *Chlamydomonas reinhardt* (Wykoff *et al.*, 1999) e PHR1 de *Arabidopsis* (Rubio *et al.*, 2001). Também foi induzido durante o experimento um gene cujo homólogo em milho codifica um fator de transcrição MADS. Fatores de transcrição dessa família já foram descritos em *Arabidopsis* como controladores das alterações na arquitetura radicular decorrentes de deficiências nutricionais (Zhang e Forde, 1998), portanto é possível que esse gene induzido em cana-de-açúcar esteja também exercendo esse papel.

Vários genes cujos similares em outras espécies participam da resposta a diferentes estresses foram diferencialmente expressos em cana-de-açúcar. Um gene cujo homólogo em milho codifica uma superóxido dismutase foi induzido no estágio inicial da deficiência, o que pode representar uma proteção contra o estresse oxidativo gerado pela deficiência de fósforo, a qual induz a produção de

espécies reativas de oxigênio (González-Meler *et al.*, 2001). Outros genes relacionados ao estresse oxidativo foram induzidos durante o experimento, como dois genes cujos homólogos em milho e trigo codificam glutathione S-transferases e dois similares a genes de tabaco e sorgo que codificam citocromos P450. Genes para citocromos P450 também foram induzidos por deficiência de fosfato em *Arabidopsis* (Hammond *et al.*, 2003) e em tremoço branco (Uhde-Stone *et al.*, 2003). A repressão de dois genes cujos homólogos em milho e trigo codificam tioredoxinas do tipo H confirma a existência de alterações nas condições de redox da raiz em resposta à ausência de P, já que essas enzimas são importantes reguladores do ambiente intracelular (Gelhaye *et al.*, 2004). A expressão diferencial de vários genes cujos similares em outras espécies são relacionados a outros estresses, causados por ferimento, seca e patógenos, mostra que as vias metabólicas envolvidas na resposta a deficiência de fosfato têm diversos pontos em comum com os mecanismos de resposta a esses estresses. O fato de a maioria desses genes ser reprimida, porém, é sinal da especificidade da resposta a deficiência de fosfato.

Diversos genes de cana-de-açúcar cuja expressão responde a deficiência de fosfato são similares a genes de outras espécies que codificam proteínas quinases, e um gene tem um homólogo em milho que codifica uma proteína fosfatase (PP2C). Essas proteínas têm papel na regulação de diversas vias de transdução de sinal e da síntese proteica, com isso atuando no controle do crescimento e da divisão celular (como é o caso da caseína quinase). A regulação desses genes é coerente com as consequências conhecidas da deficiência de fosfato, que implica em diversas alterações no crescimento radicular da planta e

na realocação das reservas deste nutriente. Um gene para proteína fosfatase 2C também foi reprimido em raízes de *Arabidopsis* privadas de fosfato, além de ser induzido em folhas nas mesmas condições (Wu *et al.*, 2003). Nesse mesmo estudo, foram encontrados genes de *Arabidopsis* responsivos a essa deficiência que codificam proteínas kinases componentes da cascata MAPK de transdução de sinal.

A sinalização envolvendo hormônios também foi afetada pela deficiência de fosfato, já que alguns genes de cana-de-açúcar que foram diferencialmente expressos devido ao estresse são similares a genes de outras espécies relacionados à biossíntese de hormônios e à resposta aos mesmos. Foram encontradas alterações nos níveis de expressão de um gene cujo homólogo em arroz codifica uma enzima supostamente envolvida na biossíntese de etileno, um gene cujo homólogo em arroz é induzido por giberelina e um gene similar a genes de outras espécies reprimidos por auxina, analogamente a resultados encontrados em *Arabidopsis* (Wu *et al.*, 2003). Outra evidência de que a via de resposta ao etileno é afetada pela deficiência de fosfato é a repressão, após 6 h de estresse, de um gene cujo homólogo em *Arabidopsis* codifica uma proteína semelhante ao fator de transcrição ETHYLENE INSENSITIVE3 (EIN3). De acordo com Mao *et al.* (2006), um homólogo de EIN3 em arroz, *OsEIL1*, age como um regulador positivo da resposta a etileno, e plantas de arroz transgênicas superexpressando *OsEIL1* apresentam fenótipo com raízes curtas, raízes primárias enroladas, parte aérea ligeiramente mais curta e elevada resposta a etileno exógeno. A repressão em cana-de-açúcar deste fator de transcrição pode estar ligada às mudanças na arquitetura radicular que ocorrem em resposta à deficiência de fosfato, nas quais

tanto o etileno quanto a auxina têm provável papel (Lynch e Brown, 1997). Um gene reprimido no primeiro estágio da deficiência (6 h) e induzido após 48 h de experimento codifica um homólogo da BLE1 de arroz, que é induzido por brassinolídeos. Plantas de arroz nocauteadas para a expressão do gene *OsBle1* apresentaram redução no crescimento (Yang e Komatsu, 2004). Assim, a repressão do homólogo de *OsBle1* de cana-de-açúcar logo na fase inicial do estresse poderia fazer parte da resposta da planta de frear o seu metabolismo para se adequar à situação de restrição de um elemento essencial. Essa hipótese está em linha com as observações em *Arabidopsis*, onde inclusive genes relacionados com a fotossíntese foram reprimidos (Wu *et al.*, 2003). Genes envolvidos na biossíntese de ácido jasmônico e ácido salicílico também foram induzidos pela deficiência de fosfato, indicando que a resposta da cana-de-açúcar ao estresse também tem intercessões com as vias de sinalização acionadas por esses hormônios.

Vias de sinalização envolvendo canais de cálcio e GTPases têm também seu papel na resposta da cana-de-açúcar à deficiência de fosfato. Isto é indicado pela expressão diferencial de cinco genes cujos homólogos em arroz codificam quatro GTPases e uma proteína ativadora de GTPase. Observamos também a indução de dois genes cujos homólogos em milho codificam uma calreticulina e um fator de alongação 1-alfa que se liga a calmodulina, além da repressão de um gene similar a um gene de *Arabidopsis* para uma Ca^{2+} ATPase e de um gene cujos homólogos em arroz e soja codificam enzimas com atividade fosfatase de nucleosídeo-trifosfato, também conhecidas como apirases. A apirase de ervilha está envolvida na maior absorção de fósforo, conforme observado em plantas

transgênicas de *Arabidopsis* (Thomas *et al.*, 1999). A repressão do homólogo de cana-de-açúcar deste gene só foi verificada no momento inicial do experimento (6 h), corroborando a hipótese de que o estresse por deficiência de P causa um grande ajuste global inicial.

A degradação e a síntese de proteínas também são afetadas pela deficiência de fosfato, como indicado pela regulação, em resposta a esse estresse, da expressão de genes de cana-de-açúcar cujos similares em outras espécies codificam ubiquitinas e proteínas ribossomais. Em *Arabidopsis*, diversos genes para ubiquitinas foram induzidos devido à deficiência de fosfato, supostamente envolvidos na via ubiquitina-proteasomo de degradação de proteínas (Wu *et al.*, 2003). Em arroz, genes para três proteínas ribossomais foram reprimidos em resposta a esse estresse, incluindo o gene para a proteína ribossomal 60S L23A (Wasaki *et al.*, 2003), cujo homólogo em cana-de-açúcar foi também reprimido. O dobramento de proteínas é também afetado pela deficiência de fosfato, como indicado pela repressão de um gene cujo homólogo em arroz codifica uma chaperonina.

O metabolismo de carboidratos também parece ser influenciado pela deficiência de fosfato. Dois genes induzidos após 48 h de experimento têm homólogos em milho e arroz que codificam, respectivamente, uma frutose bisfosfate aldolase e uma trealase, ambas envolvidas no metabolismo de carboidratos. A indução de um gene para frutose bisfosfate aldolase coincide com os resultados encontrados em tremoço branco submetido à deficiência de fosfato (Uhde-Stone *et al.*, 2003).

Vários outros genes de cana-de-açúcar foram diferencialmente expressos em resposta à deficiência de fosfato. Entre eles, muitos apresentam similaridades apenas com genes que codificam proteínas de funções desconhecidas e outros não apresentam similaridade com nenhum gene já descrito.

Em suma, os dados obtidos indicam que as células radiculares de cana-de-açúcar alteram diversos aspectos do seu metabolismo em resposta à deficiência de fosfato. A identificação de um grande número de genes envolvidos na transdução de sinal cuja expressão é regulada em resposta a esse estresse é um primeiro passo na direção de elucidar as estratégias adotadas pela cana-de-açúcar para adaptar-se a pouca disponibilidade desse nutriente. O aprofundamento da caracterização desses genes poderá abrir perspectivas de como manipular as cultivares de cana-de-açúcar de forma a obter melhorias do ponto de vista agrônomo.

Capítulo 8: Agrupamentos dos genes de cana-de-açúcar cuja expressão responde à deficiência de fosfato.

Algoritmos de agrupamento são ferramentas de análise exploratória de dados, cujo objetivo é identificar grupos (ou seja, subconjuntos dos dados) com padrões empíricos semelhantes. Os grupos se formam devido à similaridade entre seus elementos e à dissimilaridade entre estes e elementos de outros grupos. Estes algoritmos vêm sendo largamente utilizados em dados de arranjos de DNA (Chu *et al.*, 1998, Eisen *et al.*, 1998, Tamayo *et al.*, 1999), procurando identificar grupos de genes com perfis de expressão semelhantes nos experimentos estudados. Além disso, a aplicação desses algoritmos aos dados de arranjos de DNA fornece uma maneira de visualizar os resultados dos experimentos, facilitando assim a sua interpretação.

O embasamento da aplicação dos algoritmos de agrupamento a dados de expressão gênica começa pela definição matemática da similaridade entre os padrões de expressão de dois genes quaisquer. Para isso, definimos para cada gene um “vetor de expressão”, formado pelos logaritmos das razões de expressão do gene em diferentes pontos experimentais (momentos em uma série temporal, diferentes condições experimentais, diferentes tecidos etc.). Podemos então encarar estes vetores como pontos em um espaço vetorial, no qual cada dimensão corresponde a um ponto experimental. A similaridade entre dois genes pode então ser definida a partir de uma noção de distância nesse espaço vetorial, isto é, quanto menor a distância entre dois pontos no espaço, maior a similaridade entre os respectivos genes. Esse espaço vetorial, chamado de “espaço de

expressão”, pode ser munido de diversas métricas (distâncias); a mais freqüentemente empregada é a distância euclidiana:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}, \quad (8.1)$$

na qual n é a dimensão do espaço. Outra medida de similaridade bastante utilizada é o coeficiente de correlação:

$$corr(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sigma_x \cdot \sigma_y}, \quad (8.2)$$

em que $\bar{x}$ e $\bar{y}$ são as médias dos vetores x e y e σ_x e σ_y são seus desvios.

Após a escolha da medida de distância a ser utilizada nos agrupamentos, outro ponto importante na aplicação dos mesmos é a padronização dos dados. Esta estratégia pode ser empregada para evitar que genes com valores mais extremos dos logaritmos das razões de expressão dominem as análises, obscurecendo os demais. Além disso, muitas vezes deseja-se identificar semelhanças entre os padrões de expressão dos genes enfocando mais os valores relativos desses logaritmos que seus valores absolutos. Uma padronização usual aos dados de arranjos de DNA é obter, para cada gene, a média de suas coordenadas no espaço de expressão, e subtrair esse valor dessas mesmas coordenadas (técnica conhecida na literatura como “mean centring”, Quackenbush, 2001). Geometricamente, essa técnica equivale a projetar, no espaço de expressão, os pontos representativos dos genes no hiperplano definido por $\sum_i x_i = 0$. Essa padronização é especialmente útil em experimentos de série

temporal, onde se deseja encontrar genes que apresentem, ao longo do tempo, variações semelhantes nas suas expressões, em relação ao seu nível basal.

Outra padronização usual, em geral aplicada após a subtração da expressão média de cada gene, é a divisão dos logaritmos das razões de expressão de cada gene pelo desvio dos mesmos, de forma que os dados padronizados de cada gene tenham, além de média igual a zero, desvio igual a um. Geometricamente, essa padronização é equivalente a projetar, no espaço de expressão, os pontos representativos dos genes (já subtraídos de sua média) na esfera centrada na origem de raio igual a $\sqrt{n-1}$. Apesar de essas padronizações apresentarem a vantagem de destacar o perfil de expressão de cada gene, facilitando assim a identificação de grupos entre os mesmos, uma desvantagem das mesmas é a perda de informação quanto aos valores absolutos dos logaritmos das razões de expressão. Por exemplo, no caso da subtração da média, um gene que seja homogeneamente induzido em todo o experimento (logaritmos das razões de expressão iguais a 2, por exemplo) e um gene que seja homogeneamente reprimido podem acabar tendo perfis de expressão padronizados extremamente semelhantes.

Uma vez definidas a padronização dos dados e a medida de distância a serem utilizadas, diferentes algoritmos de agrupamento podem ser aplicados aos dados. As estratégias de agrupamentos dividem-se em aglomerativas e divisivas. Algoritmos aglomerativos consideram, a princípio, cada elemento como um grupo separado, e então gradualmente unem esses grupos, de acordo com sua semelhança, para formar grupos maiores. Por outro lado, algoritmos divisivos

começam considerando todos os elementos como um só grupo, que é gradualmente dividido em grupos menores.

Algoritmos de agrupamento podem ser ainda divididos em hierárquicos e não hierárquicos. Os primeiros são estratégias aglomerativas que vêm sendo largamente aplicadas a dados de arranjos de DNA (Eisen *et al.*, 1995), já que são simples e apresentam resultados de fácil visualização. Nestes algoritmos os perfis de expressão dos genes são unidos gradualmente formando grupos, que por sua vez são unidos para formar grupos maiores, e assim sucessivamente, formando uma árvore hierárquica (dendrograma) semelhante às árvores filogenéticas, na qual quanto mais alto o nó que liga dois grupos, menor a similaridade entre eles. Um ponto delicado na aplicação de algoritmos hierárquicos é a definição da distância entre grupos, a partir da distância entre elementos já definida. Diferentes definições da distância $d(A,B)$ entre os grupos A e B podem ser utilizadas (Figura 8.1), e as mais frequentemente aplicadas a dados de arranjos de DNA são (Quackenbush, 2001):

- Ligação simples - $d(A,B) = \min\{d(x,y) | x \in A, y \in B\}$,
- Ligação completa - $d(A,B) = \max\{d(x,y) | x \in A, y \in B\}$,
- Ligação média - $d(A,B) = \text{média}\{d(x,y) | x \in A, y \in B\}$,
- Ligação média ponderada - semelhante à ligação média, porém com os tamanhos dos grupos (números de elementos) sendo usados como pesos no cálculo da distância,

- Distância entre centróides (*within-groups clustering*) – a média de cada grupo (centróide) é calculada e a distância entre os grupos é definida como a distância entre esses centróides.

Uma vez definida essa distância, os algoritmos hierárquicos funcionam da mesma forma: a princípio cada gene é considerado como um grupo de um só elemento, as distâncias entre os genes são computadas e o par de genes mais próximos (ou seja, cujos padrões de expressão são mais semelhantes) são unidos num grupo, recalculando-se então as distâncias dos demais genes a esse grupo. A cada passo subsequente do algoritmo os grupos mais próximos são unidos e as distâncias do grupo resultante dessa união aos demais grupos são calculadas, e assim sucessivamente até todos os genes encontrarem-se reunidos num grupo único (Figura 8.1). A grande desvantagem dos algoritmos hierárquicos, porém, é que a árvore hierárquica gerada não fornece uma divisão precisa dos dados em grupos distintos. Uma vez que o último passo do algoritmo sempre reúne todos os elementos num único grupo, resta ao usuário fazer uma inspeção visual e arbitrar os ramos da árvore que devem ser considerados como grupos distintos.

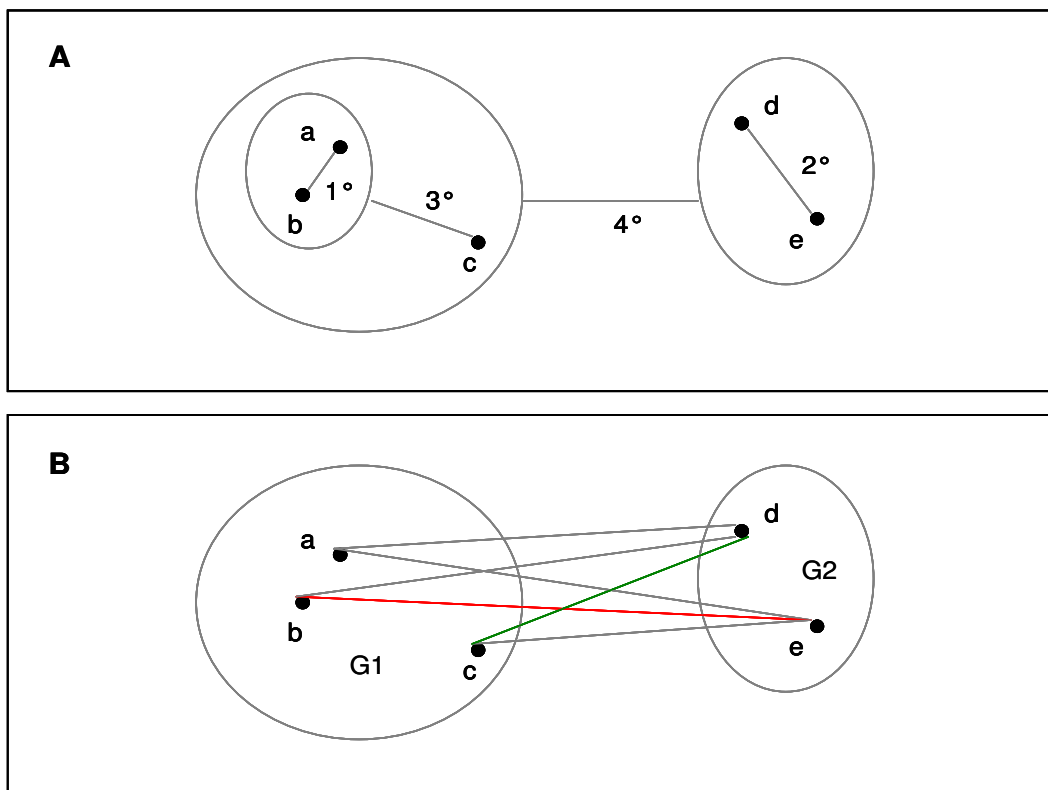


Figura 8.1: Exemplo de agrupamento hierárquico. A) A aplicação de um algoritmo hierárquico de agrupamento aos dados representados pelos pontos a, b, c, d e e , utilizando a distância euclidiana (que coincide com a noção usual de distância entre pontos), irá unir em primeiro lugar os pontos a e b , mais próximos entre si. Os passos subseqüentes do algoritmo encontram-se representados pelas linhas numeradas na figura, onde uma linha unindo pontos ou grupos indica a união destes num mesmo grupo (devido à simplicidade deste exemplo, a ordem em que essas associações se dariam seria a mesma caso adotássemos qualquer uma das definições de distância entre grupos citadas). Como pode ser observado, o último passo do algoritmo une todos os pontos num único grupo. B) Utilizando a distância euclidiana para medir a dissimilaridade entre os pontos, pode-se definir a distância entre os grupos $G1$ e $G2$, indicados na figura, de diversas formas: a ligação simples é a menor distância entre um ponto de $G1$ e um ponto de $G2$, indicada pela linha verde; a ligação completa é a maior distância entre pontos de $G1$ e $G2$, indicada pela linha vermelha; a ligação média é obtida calculando a média de todas as distâncias entre pontos de $G1$ e pontos de $G2$, indicadas por todas as linhas entre os grupos.

Diversos algoritmos não hierárquicos também vêm sendo aplicados a dados de arranjos de DNA, como k-médias (*k-means*, Johnson e Wichern, 2001) e mapas auto organizáveis (*self-organizing maps*, Tamayo *et al.*, 1999). Esses algoritmos dividem os dados em grupos sem se preocupar em estabelecer a relação entre cada par de elementos. Para isso, em geral tais algoritmos geram as coordenadas dos centróides (pontos médios) de cada grupo, que são então alteradas ao longo do algoritmo, de maneira a adequar os grupos à estrutura dos dados. Por fim, cada gene é associado ao grupo de cujo centróide encontra-se mais próximo. Desta forma, esses algoritmos não precisam calcular as distâncias entre todos os pares de genes, computando apenas as distâncias entre estes e cada centróide dos grupos, o que diminui o seu gasto computacional. Uma desvantagem desses algoritmos, porém, é que o número de grupos precisa ser determinado *a priori*, o que nem sempre é uma tarefa simples. Para contornar esse problema foi proposta a aplicação de outras estratégias matemáticas, como a análise de componentes principais (PCA, Johnson e Wichern, 2001) e a associação de técnicas de bootstrap aos algoritmos de agrupamento (Bodt *et al.*, 2002), porém em geral essas estratégias dependem de uma inspeção visual nos gráficos gerados.

8.1 – Algoritmos hierárquicos de agrupamento na visualização de dados de expressão gênica

Os algoritmos hierárquicos de agrupamento são os mais utilizados para visualizar dados de expressão gênica (Eisen *et al.*, 1995). A árvore hierárquica gerada por esses algoritmos fornece uma ordenação dos dados que, aplicada a matriz formada pelos valores dos logaritmos das razões de expressão, permite visualizar claramente os padrões empíricos existentes entre os genes. Para isso, gera-se um gráfico a partir dessa matriz, no qual o valor de cada logaritmo é representado por um retângulo colorido, de forma que logaritmos positivos (genes induzidos) são representados em vermelho e negativo (reprimidos) em verde, com a intensidade das cores correspondendo a magnitude dos logaritmos (Figura 8.2).

A aplicação de algoritmos de agrupamento aos dados do experimento de deficiência de fosfato (Capítulo 7) foi restrita a uma seleção dos genes, procurando com isso não sobrecarregar os algoritmos com genes que não apresentaram variações na sua expressão durante o experimento. Foram selecionados os genes que apresentaram, em qualquer um dos pontos experimentais, um p-valor menor que 5%, de acordo com o algoritmo ISER2. Vale observar que essa seleção foi mais abrangente que a apresentada no Capítulo 7, podendo incorrer num FDR de até 47%, enquanto naquele capítulo o FDR foi limitado em 5%. A escolha de critérios menos rigorosos para esta seleção justifica-se pelo fato de que nesse caso deseja-se detectar padrões de expressão nos dados, e posteriormente pode-se fazer uma inspeção visual dos grupos obtidos e selecionar aqueles que representam padrões mais interessantes.

A métrica adotada para realizar os agrupamentos foi a distância euclidiana (8.1), e a distância entre grupos foi computada de acordo com a ligação média (*average-linkage clustering*). Para a aplicação do algoritmo hierárquico, nenhuma padronização foi efetuada nos dados, de forma que as o sistema de cores da Figura 8.2 reflete exatamente os valores dos logaritmos das razões de expressão médias computadas pelo algoritmo ISER2.

A ordenação dos genes e o dendrograma obtidos encontram-se na Figura 8.2. Pode-se observar que os genes mais fortemente reprimidos em resposta a deficiência de fosfato encontram-se na parte superior da figura, enquanto os mais induzidos encontram-se na parte inferior. O esquema de cores, à esquerda da figura, fornece uma maneira eficiente de visualizar os dados, porém o dendrograma não indica uma divisão única dos genes em grupos.

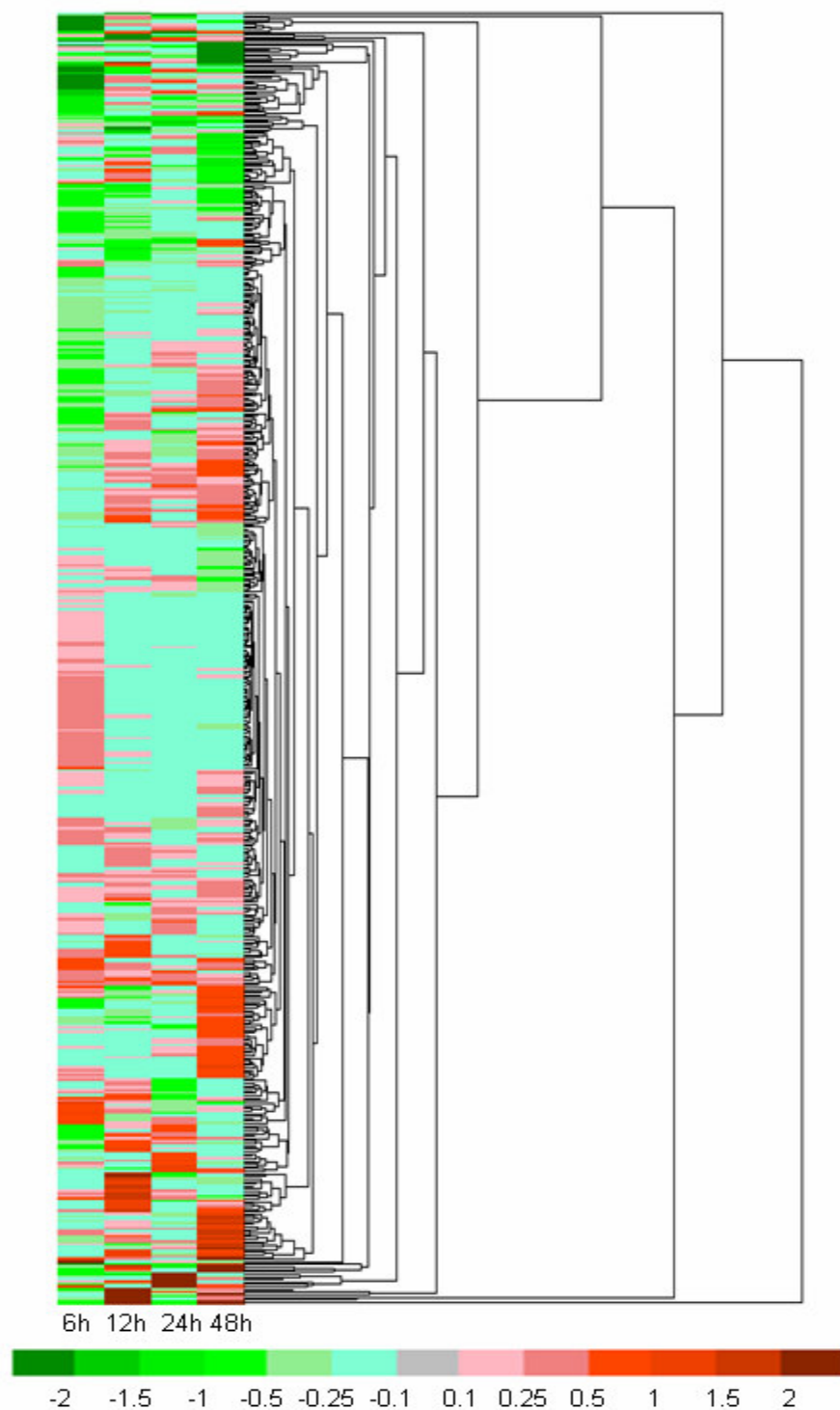


Figura 8.2: Resultados da aplicação de algoritmo hierárquico aos dados de fosfato. Cada linha no esquema de cores corresponde a um gene, e a distância horizontal de cada nó do dendrograma indica a distância entre os componentes do respectivo grupo. O esquema de cores indica os valores dos logaritmos das razões de expressão nos quatro tempos do experimento (6, 12, 24 e 48 h), de acordo com os limites dos intervalos indicados na barra inferior.

8.2 – Mapas auto organizáveis (*self-organizing maps*)

Um mapa auto-organizável (*self-organizing map* ou *SOM*, Kohonen, 1995) é um algoritmo divisivo, não hierárquico e baseado em conceitos de redes neurais. Nele, os centróides de um número pré-definido de grupos são inicialmente dispostos em uma grade retangular (cujas dimensões também precisam ser pré-definidas) no plano gerado pelas duas primeiras componentes principais do espaço onde se encontram os dados em estudo. Esses centróides (que também são pontos do espaço) são então ajustados iterativamente de maneira a se ajustarem a estrutura dos dados. A cada iteração do algoritmo, um elemento do conjunto de dados é selecionado ao acaso, e os centróides próximos a ele são deslocados ligeiramente em sua direção. Após um grande número de iterações, os centróides migram para as regiões do espaço onde a concentração de pontos é maior, e então o conjunto de dados é dividido associando-se cada elemento ao grupo de cujo centróide encontra-se mais próximo.

A estratégia adotada pelo SOM é extremamente eficiente para adaptar a localização dos centróides à estrutura dos dados, identificando assim os grupos presentes nos mesmos. Porém, para que essa estratégia seja realmente eficaz é preciso que o número de grupos seja corretamente definido, o que constitui um ponto fraco do método. Para contornar essa dificuldade e aumentar a confiabilidade dos resultados do SOM, Bodt *et al.* (2002) propôs uma estratégia baseada na aplicação do algoritmo a conjuntos de re-amostras *bootstrap* dos dados iniciais: os elementos (pontos) do conjunto de dados são re-amostrados com reposição, tendo cada re-amostra o mesmo número de elementos do

conjunto inicial, e o SOM é aplicado a cada re-amostra utilizando-se as mesmas dimensões da grade de centróides (isto é, o mesmo número de grupos). Desta forma, a cada re-amostra são obtidos n centróides de grupos, $(G_1, \dots, G_n)$, e calcula-se a soma de quadrados intra-classe (ou soma dos erros de quantização):

$$SSIntra(G_1, \dots, G_n) = \sum_{i=1}^n \sum_{x_j \in V_i} d^2(x_j, G_i), \quad (8.3)$$

em que V_i é a região do espaço mais próxima de G_i que de qualquer outro centróide e d é a distância euclidiana. Uma vez obtidos esses valores, calcula-se então seu coeficiente de variação:

$$CV(SSIntra) = \frac{\sigma_{SSIntra}}{\mu_{SSIntra}}, \quad (8.4)$$

no qual $\mu_{SSIntra}$ e $\sigma_{SSIntra}$ são respectivamente a média e o desvio dos valores de $SSIntra$ obtidos em todas as re-amostras *bootstrap*. Esse processo é repetido com diferentes números de grupos no SOM e, por fim, desenhando-se o gráfico dos valores do coeficiente de variação contra os números de grupos (Figura 8.3) percebe-se que aqueles crescem gradativamente em função destes, apresentando porém um salto após certo número de grupos. Bodt *et al.* (2002) argumenta que esse salto deve-se ao fato de que, deste ponto em diante, o número de grupos utilizados no SOM torna-se excessivo para a estrutura dos dados em estudo, recomendando portanto que seja adotado o número de grupos imediatamente anterior ao salto.

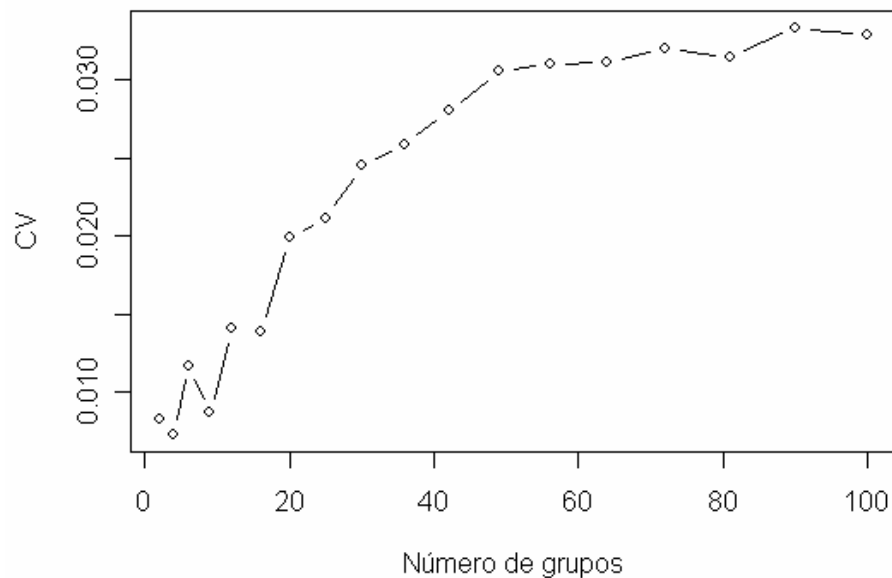


Figura 8.3: Teste para determinar o número de grupos a ser utilizado no algoritmo SOM. Os números de grupos testados correspondem a grades com dimensões $n \times (n-1)$ ou $n \times n$, com $n=1, \dots, 10$. Devido a acréscimo nos valores de CV após o ponto correspondente a 16 grupos (grade 4×4), adotou-se essa quantidade de grupos para a aplicação do algoritmo aos dados.

Outro ponto delicado na aplicação do SOM é a seleção ao acaso de um ponto do conjunto de dados a cada iteração do algoritmo. Devido à aleatoriedade dessa escolha, sucessivas aplicações do SOM a um mesmo conjunto de dados podem gerar resultados diferentes, o que põe em dúvida a confiabilidade do algoritmo. Para resolver esse problema e aumentar a capacidade do SOM em detectar a estrutura de grupos intrínseca aos dados, Rousset *et al.* (2005) propôs um método estocástico baseado em re-amostras *bootstrap*. A estratégia divide-se em dois passos: em primeiro lugar, re-amostras do conjunto de dados, com o mesmo número de elementos deste, são tomadas com reposição. A cada re-

amostra é aplicado o algoritmo SOM, sempre com as mesmas dimensões pré-fixadas da grade de centróides. Ao fim da análise de cada re-amostra, uma vez definidas as coordenadas finais dos centróides, todos os elementos do conjunto de dados (e não apenas os que faziam parte daquela re-amostra) são divididos entre os grupos, com cada elemento sendo situado no grupo de cujo centróide encontra-se mais perto. Computa-se então uma matriz V^r , onde r é o índice da re-amostra, definida por:

$$V^r(i, j) = \begin{cases} 1 & \text{se } x_i \text{ e } x_j \text{ são vizinhos} \\ 0 & \text{em caso contrário} \end{cases}, \quad (8.5)$$

em que os elementos x_i e x_j são considerados vizinhos se estão contidos no mesmo grupo ou em grupos vizinhos, isto é, grupos cujos centróides originalmente eram adjacentes na grade inicial.

Ao fim de um número pré-estabelecido B de re-amostras (Rousset et al., 2005, recomenda fixar B entre 200 e 500), calcula-se a matriz V_{Boot} , contendo as probabilidades empíricas de cada par de elementos serem vizinhos:

$$V_{\text{boot}}(i, j) = \frac{\sum_{r=1}^B V^r(i, j)}{B}. \quad (8.6)$$

Em uma segunda etapa, o método executa diversas vezes o algoritmo SOM com os dados originais (agora sem fazer re-amostras), calculando, para cada resultado obtido, uma matriz V^s definida analogamente a matriz V^r (8.5). Considerando então que a matriz V_{Boot} reflete a estrutura de grupos intrínseca aos dados, busca-se entre as matrizes V^s aquela que mais se assemelha a esta, ou seja, aquela cuja distância a V^s é menor. A distância entre matrizes é computada de acordo com a norma de Frobenius:

$$D(V^r, V_{boot}) = \frac{1}{N^2} \sqrt{\sum_{i,j=1}^N (V^r(i, j) - V_{boot}(i, j))^2}, \quad (8.7)$$

na qual N é o número total de elementos no conjunto de dados. O resultado do SOM correspondente à matriz V^s selecionada é então aceito como aquele que melhor representa a estrutura de grupos dos dados originais.

Para a aplicação do algoritmo SOM aos dados do experimento de deficiência de fósforo (Capítulo 7) foi feita uma seleção prévia dos genes análoga à realizada para a aplicação do algoritmo hierárquico ($p\text{-valor} < 0.05$). Com o objetivo de facilitar a identificação dos grupos presentes nos dados, decidimos adotar como padronização dos mesmos o procedimento de subtrair, das coordenadas de cada gene, a média das mesmas (*mean centering*) e dividir os valores obtidos pelo seu desvio. Porém, como já foi observado acima, esse procedimento causa perda de informação, e para minimizar esse problema adotamos um artifício matemático.

Como até o início do experimento de deficiência de fósforo os grupos *tratado* e *controle* passaram exatamente pelas mesmas condições experimentais, permitindo até mesmo que associássemos ao acaso as plantas a um grupo ou outro, é razoável supor que, em média, o logaritmo da razão de expressão de cada gene fosse igual a zero (razão igual a um) num momento exatamente anterior ao início da deficiência, que chamaremos de 0 h. Acrescentando esse ponto aos dados antes da padronização, passamos a ter uma referência nas coordenadas de cada gene, impedindo assim que a subtração da média das mesmas leve à perda de informação. Essa estratégia foi aplicada aos dados do

experimento, que assim passaram a ter cinco dimensões (0, 6, 12, 24 e 48 h). As cinco coordenadas de cada gene foram então padronizadas pela média e desvio.

O algoritmo SOM foi aplicado aos dados, utilizando a distância Euclidiana como medida da similaridade entre os genes. A estratégia proposta por Bodt *et al.*, (2002) foi empregada para definir o numero de grupos a ser utilizado no algoritmo. O resultado dos testes (Figura 8.3) indica um maior acréscimo nos valores do coeficiente de variação (CV, definido na equação 8.4) após o ponto referente a 16 grupos (grade 4x4). Com isso, o algoritmo SOM foi aplicado aos dados com uma grade inicial de 4x4 grupos.

A estratégia proposta por Rousset *et al.*, (2005) foi também utilizada na aplicação do algoritmo SOM aos dados. Quinhentas re-amostras bootstrap foram utilizadas para obter a matriz V_{Boot} . O algoritmo foi aplicado quinhentas vezes aos dados, selecionando-se os resultados cuja matriz de vizinhanças V^s melhor aproximou V_{Boot} . Os resultados do agrupamento que, de acordo com esse critério, melhor reflete a estrutura de vizinhanças dos dados encontram-se na Figura 8.4.

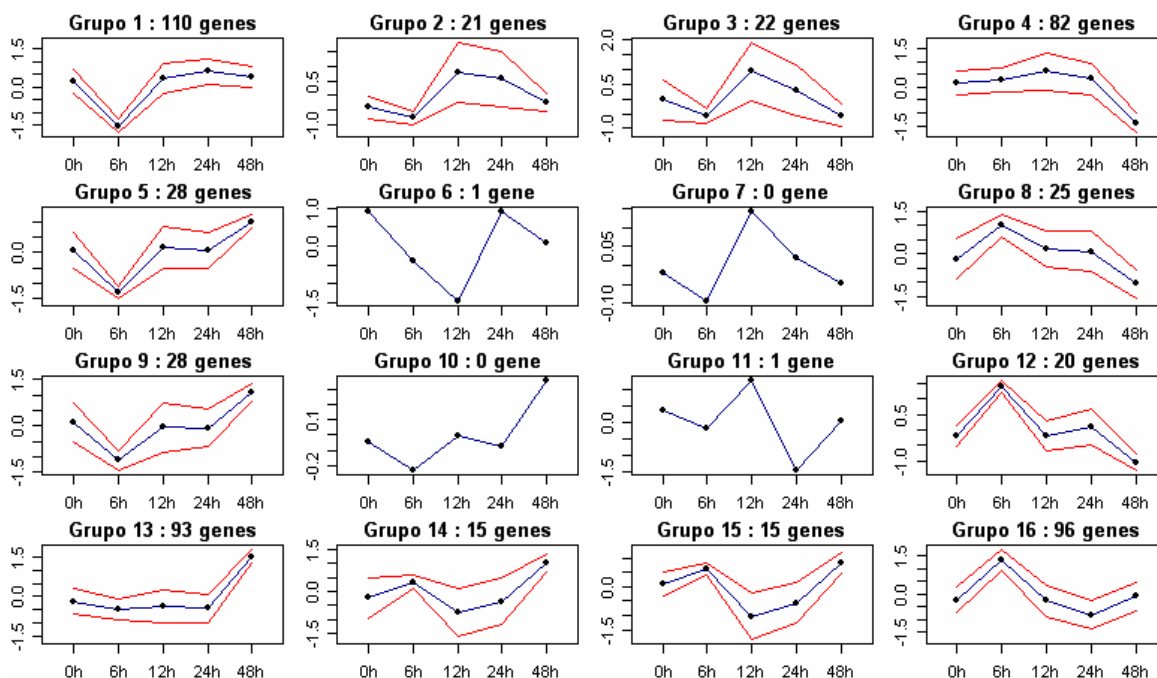


Figura 8.4: Agrupamento obtido nos dados de fosfato com o algoritmo SOM. Em cada grupo, a linha em azul representa a média do grupo (centróide) e as linhas em vermelho representam essa média acrescida (ou subtraída) do desvio padrão das coordenadas dos genes contidos no grupo. Há apenas um gene no grupo 6 e no 11, e os perfis destes são representados pelas linhas em azul. Nos grupos 7 e 10 as linhas azuis representam os centróides gerados pelo algoritmo, aos quais nenhum gene foi associado.

Analisando a Figura 8.4, nota-se em primeiro lugar a presença de dois grupos com apenas um gene em cada um (6 e 11) e de dois centróides os quais não foi associado nenhum gene (“grupos” 7 e 10). A presença desses últimos é um indício de que o número de grupos utilizado no algoritmo foi suficiente para captar a estrutura dos dados, sugerindo até que poderíamos ter utilizado menos grupos. A presença destes grupos adicionais, porém não prejudica as conclusões que se pode extrair dos demais grupos formados.

Os genes que formaram grupos de um só elemento apresentam perfis de expressão diferentes dos demais. O gene que codifica uma proteína quinase

(grupo 6) foi reprimido durante os estágios iniciais da deficiência de fósforo (6 e 12 h) e retornou a níveis de expressão próximos aos obtidos em plantas controle nos estágios subsequentes (24 e 48 h), o que provavelmente indica que a repressão desse gene faz parte da resposta genérica a estresse, ativada no princípio da deficiência (como citado no Capítulo 7), porém não participa da resposta específica à privação de fósforo. O gene que forma o grupo 11 codifica uma proteína relacionada à resposta a patógenos e apresenta um perfil singular, aparentemente sendo levemente induzido em 12 h de experimento e intensamente reprimido em 24 h. Esse comportamento pode ser associado a um estágio da resposta a deficiência de fósforo ativado especificamente nestes pontos experimentais, havendo ainda a possibilidade de esse gene ser um falso positivo, uma vez que ele não faz parte da seleção mais restrita de genes diferencialmente expressos descrita no Capítulo 7.

Entre os demais grupos identificados pelo SOM, diversos apresentam perfis de expressão interessantes, caracterizando os diferentes padrões de resposta a deficiência de fósforo. Observando os genes contidos em cada um desses grupos, de acordo com sua classificação no catálogo SUCAST (Souza *et al.*, 2001), pode-se obter um panorama das diversas fases da resposta da cana-de-açúcar a esse estresse nutricional, como descrito a seguir.

O grupo 1 é o mais numeroso, representando os genes cuja expressão é reprimida em 6 h de experimento e volta a níveis próximos dos observado em plantas controle nos demais pontos experimentais. Os genes nesse grupo fazem parte da resposta inicial à deficiência de fósforo, o que pode indicar que eles estão relacionados a uma resposta genérica a estresse (como citado no Capítulo 7).

Estão presentes nesse grupo quatorze genes classificados como responsáveis por fatores de transcrição, incluindo fatores dos tipos Myb, WRKY e MADS. Também fazem parte desse grupo doze genes classificados como relacionados a respostas a estresses, como os que codificam tioredoxinas, uma peroxidase, um citocromo P450 e proteínas envolvidas nas respostas à seca. Ainda se encontram nesse grupo genes envolvidos na resposta a patógenos, assim como genes responsáveis por diversas proteínas quinases, receptores Ser/Thr quinase, uma proteína fosfatase, diversas ubiquitinas, proteínas envolvidas na biossíntese de hormônios e no metabolismo de cálcio, além de diversas proteínas desconhecidas.

Com um comportamento oposto ao do grupo 1, os genes contidos no grupo 16 mostram-se induzidos em 6 h de experimento e retornam a níveis de expressão próximos aos observado em plantas controle nos demais pontos experimentais. Com isso, os genes contidos nesse grupo também fazem parte da resposta inicial à deficiência de fosfato. Encontram-se aí genes classificados como responsáveis por treze fatores de transcrição e por doze proteínas relacionadas à resposta a estresses, incluindo duas superóxido dismutases (relacionadas ao estresse oxidativo) e proteínas induzidas por fermento e seca. Genes que codificam uma proteína fosfatase e diversas proteínas quinases também estão contidos nesse grupo, incluindo duas quinases pertencentes à cascata MAPK de transdução de sinal, além de diversos receptores Ser/Thr quinase. Ainda compõem esse grupo genes responsáveis por pequenas GTPases, ubiquitinas, proteínas relacionadas a resposta a patógenos, ao metabolismo de cálcio e a vias de sinalização por hormônios, além de proteínas desconhecidas.

Os grupos 5 e 9 contêm genes com perfis de expressão bastante semelhantes, cuja expressão é reprimida em 6 horas de experimento e induzida após 48 h. Com isso, podemos considerar esses dois grupos como um só, cujos genes exercem papéis na resposta inicial (provavelmente mais genérica) à deficiência de fosfato e na resposta mais tardia (mais específica) a esse estresse, porém com comportamentos opostos nos dois estágios. Estão presentes nesses grupos quatro genes responsáveis por fatores de transcrição, incluindo um fator de transcrição *heat shock*, e oito genes classificados como envolvidos na resposta a estresses, incluindo dois genes responsivos a seca e frio. Também fazem parte desses grupos duas proteínas fosfatases e quatro proteínas quinases, incluindo uma quinase pertencente a cascata MAPK de transdução de sinal, além de diversos receptores Ser/Thr quinase. Compõem ainda esse grupo genes relacionados à resposta a patógenos e genes que codificam diversas ubiquitinas, uma proteína relacionada à via de sinalização da auxina e uma pequena GTPase.

Apresentando um comportamento oposto ao dos grupos 5 e 9, os grupos 8 e 12 são também semelhantes entre si. Os genes contidos nesses grupos são induzidos em 6 h de experimento e reprimidos em 48 h, indicando assim que também têm papéis nesses dois estágios da resposta à deficiência de fosfato. Pertencem a esses grupos sete genes responsáveis por fatores de transcrição, incluindo um fator do tipo Myb e um do tipo WRKY, e seis genes relacionados a resposta a estresses, sendo dois induzidos por ferimento e três responsivos a seca e frio. Genes responsáveis por uma proteína fosfatase e por seis proteínas quinases, das quais duas fazem parte da cascata MAPK de transdução de sinal, também fazem parte desses grupos, além de genes para quatro receptores

Ser/Thr quinase. Encontram-se ainda nesses grupos cinco genes relacionados à resposta a patógenos e genes que codificam uma pequena GTPase e uma proteína relacionada a via de sinalização do ácido jasmônico.

Os perfis de expressão dos genes contidos nos grupos 2 e 3 são menos homogêneos que nos grupos já descritos, como pode ser observado pelo maior desvio nos gráficos correspondentes a esses grupos (Figura 8.4). Em geral, porém, os genes contidos nesses dois grupos são induzidos em 12 e 24 h, indicando assim que fazem parte de uma resposta intermediária a deficiência de fosfato. Encontram-se nesses grupos genes que codificam dois fatores de transcrição, sendo um do tipo WRKY, e quatro proteínas relacionadas à resposta a estresses, sendo duas delas citocromos P450 e duas envolvidas na resposta à seca. Genes para seis proteínas quinases e seis receptores Ser/Thr quinase também pertencem a esses grupos, assim como cinco genes envolvidos na resposta a patógenos. Estão contidos também nesses grupos genes que codificam uma ubiquitina, uma pequena GTPase, uma proteína envolvida na biossíntese de etileno e duas proteínas relacionadas ao metabolismo de carboidratos.

Assim como os grupos 2 e 3, os grupos 14 e 15 contêm genes com perfis de expressão menos homogêneos, como evidenciado pelos desvios relativamente maiores mostrados nos respectivos gráficos (Figura 8.4). Em geral, os genes contidos nesses grupos apresentam-se reprimidos em 12 e 24 h de experimento, além de ligeiramente induzidos após 48 h. Analogamente aos grupos 2 e 3, os genes contidos nesses grupos aparentam participar de uma resposta intermediária a deficiência de fosfato. Encontram-se aí genes que codificam dois fatores de

transcrição e quatro proteínas relacionadas à resposta a estresses, dentre elas uma tioredoxina e uma glutational S-transferase. Genes para quatro proteínas quinases, uma delas pertencente à cascata MAPK de transdução de sinal, pertencem também a esses grupos, assim como genes para um receptor Ser/Thr quinase e para uma proteína relacionada à resposta a patógenos. Compõem ainda esses grupos genes que codificam três pequenas GTPases, uma ubiquitina, uma proteína do tipo 14-3-3 e uma proteína relacionada a sinalização do hormônio auxina.

O grupo 13 contém genes cuja expressão mantém-se inalterada nos estágios iniciais da deficiência, mas que são induzidos em 48 h do experimento. Esse perfil indica que os genes contidos nesse grupo têm papel numa resposta mais tardia à deficiência de fósforo, e assim provavelmente fazem parte de mecanismos específicos de resposta a esse estresse. Pertencem a esse grupo genes que codificam doze fatores de transcrição, incluindo um fator do tipo WRKY, e sete proteínas relacionadas à resposta a estresses, que incluem quatro citocromos P450 e uma glutational S-transferase. Também se encontram nesse grupo genes para três proteínas fosfatases, dez proteínas quinases e oito receptores Ser/Thr quinase, além de genes para cinco ubiquitinas. Compõem ainda esse grupo cinco genes envolvidos na resposta a patógenos, assim como genes que codificam três pequenas GTPases e proteínas relacionadas à biossíntese de ácido jasmônico e à via de sinalização do hormônio giberelina. Além disso, dois genes relacionados ao metabolismo de carboidratos e um gene que codifica uma proteína do tipo 14-3-3 fazem parte desse grupo.

O grupo 4 mostra um perfil oposto ao do grupo 13, contendo genes cuja expressão, após manter-se inalterada durante os primeiros estágios da deficiência, é reprimida em 48 h de experimento. Desta forma, assim como os genes do grupo 13, os genes nesse grupo desempenham papel em um estágio mais avançado da deficiência de fósforo, provavelmente participando de mecanismos da resposta específica a esse estresse nutricional. Pertencem a esse grupo genes que codificam onze fatores de transcrição e seis proteínas relacionadas à resposta a estresses, incluindo um citocromo P450 e uma glutathione S-transferase. Também fazem parte desse grupo genes para uma proteína fosfatase e quatorze proteínas quinases, incluindo uma quinase da cascata MAPK de transdução de sinal, além de treze receptores Ser/Thr quinase. Encontram-se ainda nesse grupo genes que codificam duas ubiquitinas, cinco proteínas relacionadas à resposta a patógenos, uma pequena GTPase e um proteína do tipo 14-3-3, além de proteínas relacionadas à biossíntese do ácido salicílico e à via de sinalização do ácido jasmônico.

Em suma, a aplicação do algoritmo SOM aos dados de fósforo permitiu a identificação dos diversos padrões de expressão dos genes cuja expressão é regulada pela deficiência de fósforo, facilitando assim a caracterização dos diversos estágios da resposta a essa deficiência.

Conclusão

Os métodos de análise de dados provenientes de arranjos de DNA constituem uma área de pesquisa em desenvolvimento. Diversas estratégias de análise já se encontram estabelecidas na literatura, e outras vêm sendo constantemente desenvolvidas. O desenvolvimento de métodos de análise voltados para experimentos com poucas réplicas constitui um problema ainda mais complexo, dada a dificuldade de aplicar testes estatísticos convencionais a essas situações. Nesse cenário, os algoritmos ISER e ISER2, desenvolvidos neste trabalho, apresentam-se como uma eficiente opção para a análise de dados sem réplicas ou com poucas réplicas biológicas. O ISER fornece um meio de selecionar genes diferencialmente expressos a partir de experimentos sem réplicas, valendo-se da distribuição global das razões de expressão, de maneira a otimizar os resultados obtidos. O ISER2 oferece um método robusto de seleção de genes diferencialmente expressos em experimentos onde há poucas réplicas biológicas, fornecendo ainda uma estimativa da razão de falsos positivos (FDR) esperada na seleção obtida.

A aplicação desses métodos de análise ao experimento de deficiência de fosfato em cana-de-açúcar possibilitou a identificação de diversos genes diferencialmente expressos em resposta a esse estresse. Os resultados obtidos indicam que a cana-de-açúcar aciona uma ampla gama de mecanismos para enfrentar essa deficiência nutricional, que passam tanto por respostas genéricas a estresses quanto por respostas específicas a privação de fosfato, regulando a

expressão de genes que, de acordo com a literatura, provavelmente têm papel na maior absorção e no melhor aproveitamento das reservas de fosfato disponíveis.

Os algoritmos de agrupamento mostraram-se ferramentas de extrema utilidade para visualizar e interpretar a grande quantidade de dados gerados pelos experimentos com arranjos de DNA, principalmente quando estes envolvem séries temporais. A aplicação desses algoritmos aos dados gerados no experimento de deficiência de fosfato possibilitou a identificação de diversos grupos de genes com padrões semelhantes de expressão ao longo do tempo, levando assim a uma melhor caracterização da resposta da cana-de-açúcar a esse estresse.

Os resultados desse trabalho representam um primeiro passo para a compreensão dos mecanismos genéticos de resposta da cana-de-açúcar à deficiência de fosfato. Os genes identificados como envolvidos nessa resposta podem ser utilizados futuramente em experimentos empregando engenharia genética para verificar se sua superexpressão ou seu silenciamento em plantas de cana-de-açúcar as tornam mais tolerantes a essa restrição nutricional. Esses resultados podem conduzir ao desenvolvimento de linhagens de cana-de-açúcar com melhor desempenho em solos pobres de fosfato, o que seria de extremo interesse agrônomo.

Material e Métodos

Desenvolvimento de algoritmos de análise: os algoritmos foram desenvolvidos em Matlab (Mathworks, USA) e em seguida convertidos ao ambiente estatístico R (<http://www.R-project.org>) ou foram desenvolvido diretamente em R. Esses algoritmos, juntamente com outros *scripts* desenvolvidos em linguagem PERL e CGI, estão disponibilizados na nossa página <http://ipe.cbmeg.unicamp.br/>.

Material Vegetal: o tratamento de plantas foi feito em colaboração com o grupo de pesquisa coordenado pelo professor Antonio Vargas de Oliveira Figueira, do Centro de Energia Nuclear na Agricultura da USP. O material vegetal foi cedido pelo Centro de Tecnologia Canavieira (Piracicaba, SP), proveniente de culturas de meristemas de cana-de-açúcar da linhagem SP80-3280 R11. As plantas foram fornecidas na fase III (caracterizada pela organogênese e estabelecimento do sistema radicular) e permaneceram em sistema de solução nutritiva completa, baseada em Hogland & Arnon (1953), por um período de adaptação de 28 dias. Após a aclimação, dividimos as plantas em dois grupos: o grupo +P (controle) permaneceu em solução nutritiva completa, enquanto o grupo -P foi transferido para solução nutritiva desprovida de fosfato (Almeida, 2002), onde permaneceu durante todo o experimento. Amostras de tecido de raiz de doze indivíduos por tratamento foram coletadas após 6, 12, 24 e 48 horas (indivíduos diferentes para cada coleta). Em cada ponto experimental, selecionamos oito das doze amostras, descartando as plantas cujos pesos frescos totais eram muito diferentes dos

demais. As oito amostras selecionadas foram então maceradas individualmente e agrupadas em quatro grupos de duas amostras. A cada grupo foi formado um "pool" contendo a mesma quantidade de tecido macerado de cada uma das duas plantas, para extração de RNA.

Extração de RNA e produção de sondas: o RNA total foi extraído das amostras de tecido com Trizol (Life Technologies, EUA), conforme recomendações do fabricante.

A partir de cada amostra foi obtida uma sonda de cDNA marcada com fluoróforo através da incorporação de Cy3-dCTP ou Cy5-dCTP (Amersham Pharmacia Biotech, EUA) durante transcrição reversa de 10µg de RNA total usando SuperScript II (Gibco, EUA) e primer oligo dT ancorado. Após a síntese, híbridos de RNA/cDNA foram desnaturados e o RNA foi hidrolisado com NaOH. Em seguida, esse material foi purificado em placas de filtração Multiscreen Millipore e o fluoróforo total incorporado foi quantificado medindo-se DO em 550nm (Cy3) e em 650nm (Cy5).

Confecção dos microarranjos: os microarranjos de DNA utilizados foram produzidos no Laboratório de Transdução de Sinal, através da deposição de produtos de PCR dos clones de interesse em lâminas espelhadas tipo 7 silanadas, usando para isso o equipamento *Generation III Microarrays Spotter™* (Amersham Pharmacia Biotech). 1557 ESTs de cana-de-açúcar, selecionados do banco de ESTs do projeto SUCEST, foram fixados em duplicata nos arranjos. Após a

deposição dos fragmentos de DNA, as lâminas são submetidas a 50mJ de luz UV, para que as fitas de DNA sejam ligadas covalentemente às lâminas, e armazenadas em dessecador.

Hibridação das sondas aos microarranjos: Quantidades iguais de cada fluoróforo foram utilizadas nas hibridizações dos microarrays. Hibridações na presença de 50% formamida foram realizadas durante a noite a 42 °C no hibridizador automatizado de lâminas (ASP™, Amersham Pharmacia Biotech, EUA). Após algumas lavagens com SSC/SDS (1x/0,2% e 0,1x/0,2%) as lâminas foram secas sob nitrogênio e analisadas no *Generation III Scanner™* (scanner confocal) para obtenção das imagens que refletem a intensidade de fluorescências dos *spots* do *microarray*. As hibridações foram feitas em duplicatas. Em cada um dos dois conjuntos de hibridações, utilizamos as sondas de cDNA provenientes de um *pool* independente de plantas de cada tratamento/coleta. Sondas provenientes da mesma coleta, sendo uma de planta privada de fosfato e outro de planta controle, foram hibridadas no mesmo microarranjo, marcadas com fluoróforos diferentes (Cy3 e Cy5).

Obtenção e quantificação de imagens: o programa ArrayVision (APBiotech, Brasil) foi utilizado para quantificar as imagens obtidas, obtendo assim dados sobre a expressão relativa de cerca de 1550 genes selecionados de cana-de-açúcar em quatro momentos do experimento, em dois tratamentos diferentes.

Análise de agrupamentos (clustering) de expressão: foram usadas ferramentas disponíveis no ambiente R (Yan, 2004) e algoritmos baseados na literatura (Eisen *et al.*, 1998; Bodt *et al.*, 2002; Rousset e Millet, 2005), também implementados em R, para identificar grupos de genes (*clusters*) cuja expressão apresenta perfis semelhantes durante o experimento de deficiência de fosfato.

Confirmação das variações na expressão, via transcrição reversa e PCR em tempo real: 5 µg de RNA total de cada amostra foram tratados com DNase I, Amplification Grade (Invitrogen) de acordo com as instruções do fabricante e uma alíquota de 7,5 µl do RNA tratado foi utilizado para reação de transcrição reversa usando-se o kit SuperScript First-Strand Synthesis System for RT-PCR (Invitrogen). As reações de PCR em tempo real foram realizadas utilizando-se o SYBR Green PCR Master Mix (Applied Biosystems) em um equipamento GeneAmp 5700 Sequence Detection System (Applied Biosystems). Os primers específicos foram desenhados usando o Primer Express 2.0 Software (Applied Biosystems). Todas as reações de PCR em tempo-real foram realizadas em triplicata e contêm 2 µl do cDNA sintetizado diluído 1:10, primer R (reverse) 600 nM, primer F(forward) 600 nM, 12,5 µl de SYBR Green PCR Master Mix e água milli-Q autoclavada em quantidade suficiente para 25 µl. O ciclo para as reações de PCR em tempo real foram: 50 °C por 2 minutos, 95 °C por 10 minutos e 40 ciclos de desnaturação a 95 °C por 15 segundos seguido de anelamento a 60 °C por 1 minuto. A especificidade dos produtos amplificados foi analisada pelas curvas de dissociação geradas pelo equipamento. Foram utilizados controles negativos em paralelo para confirmar a ausência de qualquer tipo de contaminante

300

no meio reacional. A razão de expressão de determinado gene alvo entre a amostra experimental e a controle foi determinada baseado no método de $2^{-\Delta\Delta C_t}$ (Livak & Schmittgen, 2001).

Referências

- Allison, D.B., Gadbury, G.L., Heo, M., Fernandez, J.R., Lee, C.K., Prolla, T.A. e Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Comput. Stat. Data Anal.* 39: 1-20.
- Allison, D.B., Cui, X., Page, G.P. e Sabripour, M. (2005). Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.* 7: 55-65.
- Almeida, R.S. (2002). Identificação e caracterização de genes de transportadores de fosfato em cana-de-açúcar (*Saccharum* spp.). *Dissertação de mestrado* - Escola Superior de Agricultura Luiz de Queiroz
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. e Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.
- Araujo, P.G., Rossi, M., de Jesus, E.M., Saccaro, N.L. Jr., Kajihara, D., Massa, R., de Felix, J.M., Drummond, R.D., Falco, M.C., Chabregas, S.M., Ulian, E.C., Menossi, M. e Van Sluys, M.A. (2005). Transcriptionally active transposable elements in recent hybrid sugarcane. *Plant J.* 44: 707-717.
- Bariola, P.A., MacIntosh, G.C. e Green, P.J. (1999). Regulation of S-Like Ribonuclease Levels in Arabidopsis. Antisense Inhibition of *RNS1* or *RNS2* Elevates Anthocyanin Accumulation. *Plant Physiol.* 119: 331-342.
- Benjamini, Y. e Hochberg, Y. (1995). Controlling the False Discovery Rate: a Pratical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* 57: 289-300.
- Bernardo, J.M. e Smith, A.F.M. (2000). *Bayesian Theory*. John Wiley, New York.
- Bieleski, R.L. (1973). Phosphate pools, phosphate transport, and phosphate availability. *Ann. Rev. Plant Physiol.* 24: 225-252.
- Bodt, E., Cottrell, M. e Verleysen, M. (2002). Statistical tools to assess the reliability of self-organizing maps. *Neural Netw.* 15: 967-978.

- Bosse, D. e Köck, M. (1998). Influence of phosphate starvation on phosphohydrolases during development of tomato seedlings. *Plant Cell Environ.* 21: 325-332.
- Box, G.E.P. e Cox, D.R. (1964). An analysis of transformations. *J. R. Stat. Soc. Ser. B* 26: 211–252.
- Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O. e Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science* 282: 699-705.
- Churchill, G.A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.* 32: 490-495.
- Cleveland, W.S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *J. Am. Stat. Assoc.* 74: 829-836.
- Colantuoni, C., Henry, G., Zeger, S. e Pevsner, J. (2002). Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts. *Biotechniques* 32: 1316-1320.
- Daram, P., Brunner, S., Rausch, C., Steiner, C., Amrhein, N. e Bucher, M. (1999). *Pht2;1* Encodes a Low-Affinity Phosphate Transporter from *Arabidopsis*. *Plant Cell* 11: 2153-2166.
- Datta, S. e Datta, S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics* 19: 459-466.
- Desprez, T., Amselem, J., Caboche, M. e Höfte, H. (1998). Differential gene expression in *Arabidopsis* monitored using cDNA arrays. *Plant J.* 14: 643-652.
- Drummond, R.D., Pinheiro, A., Rocha, C.S. e Menossi, M. (2005). ISER: selection of differentially expressed genes from DNA array data by nonlinear data transformation and local fitting. *Bioinformatics* 21: 4427-4429.
- Dudoit, S., Yang, Y.H., Callow, M.J. e Speed, T.P. (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Technical report* 578, Dept. Statistics, U. Cal. Berkeley.

- Duff, S.M.G., Plaxton, W.C. e Lefebvre, D.D. (1991). Phosphate-starvation response in plant cells: *De novo* synthesis and degradation of acid phosphatases. *Proc. Natl. Acad. Sci. USA*. 88: 9538-9542.
- Dumont, F., Joris, B., Gumusboga, A., Bruyninx, M. e Loppes, R. (1993). Isolation and characterization of cDNA sequences controlled by inorganic phosphate in *Chlamydomonas reinhardtii*. *Plant Sci*. 89: 55-67.
- Efron, B. e Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.
- Efron, B., Tibshirani, R., Storey, J.D. e Tusher, V.G. (2001). Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* 96: 1151-1160.
- Eisen, M.B., Spellman, P.T., Brown, P.O. e Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*. 95: 14863-14868.
- Eulgem, T., Rushton, P.J., Robatzek, S. e Somssich, I.E. (2000). The WRKY superfamily of plant transcription factors. *Trends Plant Sci*. 5:199-206.
- Everitt, B. (1987). *Introduction to optimization methods and their application in statistics*. Chapman and Hall.
- Ezaki, B., Yamamoto, Y. e Matsumoto, H. (1995). Cloning and sequencing of the cDNAs induced by aluminum treatment and Pi starvation in cultured tobacco cells. *Physiol. Plant*. 93: 11-18.
- Freeman, W.M., Robertson, D.J. e Vrana, K.E. (2000). Fundamentals of DNA Hybridization Arrays for Gene Expression Analysis. *Biotechniques* 29: 1042-1055.
- Friddle, C.J., Koga, T., Rubin, E.M. e Bristow, J. (2000). Expression profiling reveals distinct sets of genes altered during induction and regression of cardiac hypertrophy. *Proc. Natl. Acad. Sci. USA*. 97: 6745-6750.
- Gamerman, D. (1997). *Markov Chain Monte Carlo*. CRC Press.
- Gelhaye, E., Rouhier, N. e Jacquot, J.P. (2004). The thioredoxin H system of higher plants. *Plant Physiol. Biochem*. 42: 265-271.
- Gentle, J.E. (2003) *Random number generation and Monte Carlo methods*. Springer, New York.

- González-Meler, M.A., Giles, L., Thomas, R.B. e Siedow, J.N. (2001). Metabolic regulation of leaf respiration and alternative pathway activity in response to phosphate supply. *Plant Cell Environ.* 24: 205-215.
- Grivet, L. e Arruda, P. (2001). Sugarcane genomics: depicting the complex genome of an important tropical crop. *Curr. Opin. Plant Biol.* 5:122–127.
- Hammond, J.P., Bennett, M.J., Bowen, H.C., Broadley, M.R., Eastwood, D.C., May, S.T., Rahn, C., Swarup, R., Woolaway, K.E. e White, P.J. (2003). Changes in gene expression in Arabidopsis shoots during phosphate starvation and the potential for developing smart plants. *Plant Physiol.* 132: 578-596.
- Hammond, J.P., Broadley, M.R. e White, P.J. (2004). Genetic responses to phosphorus deficiency. *Ann. Bot.* 94: 323-332.
- Herwig, R., Aanstad, P., Clark, M. e Lehrach, H. (2001). Statistical evaluation of differential expression on cDNA nylon arrays with replicated experiments. *Nucleic Acids Res.* 29: E117.
- Hogland, D.R. e Arnon, D.I. (1953) The water culture method for growing plants without soil. California Agriculture Experimental Station. *University of California Berkeley*, Circ. 347.
- Hoyle, D.C., Rattray, M., Jupp, R. e Brass, A. (2002). Making sense of microarray data distributions. *Bioinformatics* 18: 576-584.
- Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. e Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 18 Suppl 1: S96-104.
- Johnson, R.A. e Wichern, D.W. (2001). *Applied Multivariate Statistical Analysis*. Prentice Hall.
- Kerr, M.K. e Churchill, G.A. (2001). Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl. Acad. Sci. USA.* 98: 8961-8965.
- Kerr, M.K., Martin, M. e Churchill, G.A. (2000). Analysis of variance for gene expression microarray data. *J. Comput. Biol.* 7: 819-837.

- Lange, K. (1997). *Mathematical and statistical methods for genetic analysis*. Springer, New York.
- Lee, M.L.T., Kuo, F.C., Whitmore, G.A. e Sklar, J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. USA*. 97: 9834-9839.
- Lehmann, E.L. (1983). *Theory of Point Estimation*. John Wiley, New York.
- Liu, C., Muchhal, U.S. e Raghothama, K.G. (1997). Differential expression of TPSI1, a phosphate starvation-induced gene in tomato. *Plant Mol. Biol.* 33: 867-874.
- Livak, K.J. e Schmittgen, T.D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta Ct}$ method. *Methods* 25: 402-408.
- Loguinov, A.V., Mian, I.S. e Vulpe, C.D. (2004). Exploratory differential gene expression analysis in microarray experiments with no or limited replication. *Genome Biol.* 5: R18.
- Lynch, J. e Brown, K.M. (1997). Ethylene and plant responses to nutritional stress. *Physiol. Plant.* 100: 613-619.
- Maleck, K., Levine, A., Eulgem, T., Morgan, A., Schmid, J., Lawton, K.A., Dangl, J.L. e Dietrich, R.A. (2000). The transcriptome of *Arabidopsis thaliana* during systemic acquired resistance. *Nat. Genet.* 26: 403-410.
- Mao, C., Wang, S., Jia, Q. e Wu, P. (2006). OsEIL1, a rice homolog of the *Arabidopsis EIN3* regulates the ethylene response as a positive component. *Plant Mol. Biol.* 61:141-152.
- Marques, E.R., Ferreira, M.E., Drummond, R.D., Felix, J.M., Menossi, M., Savoldi, M., Travassos, L.R., Puccia, R., Batista, W.L., Carvalho, K.C., Goldman, M.H. e Goldman, G.H. (2004). Identification of genes preferentially expressed in the pathogenic yeast phase of *Paracoccidioides brasiliensis*, using suppression subtraction hybridization and differential macroarray analysis. *Mol. Genet. Genomics* 271: 667-677.

- Mondego, J.M.C., Guerreiro-Filho, O., Bengtson, M.H., Drummond, R.D., Felix, J.M., Duarte, M.P., Ramiro, D., Maluf, M.P., Sogayar, M.C. e Menossi, M. (2005). Isolation and characterization of *Coffea* genes induced during coffee leaf miner (*Leucoptera coffeella*) infestation. *Plant Sci.* 169: 351–360.
- Moraes, M.A.F.D. (2005). Número e qualidade dos empregos na agroindústria da cana-de-açúcar. in *A energia da cana-de-açúcar – Doze estudos sobre a agroindústria da cana-de-açúcar no Brasil e sua sustentabilidade*. Isaías de Carvalho Macedo (organizador). UNICA União da Agroindústria Canavieira do Estado de São Paulo, São Paulo, pp 207-213.
- Muchhal, U.S. e Raghothama, K.G. (1995). Phosphate starvation induced changes in the expression of tomato Ca^{2+} -ATPase. *Plant Physiol.* 108: 112-112.
- Mukatira, U.T., Chunming, L., Varadarajan, D.K. e Raghothama, K.G. (2001). Negative regulation of phosphate-induced genes. *Plant Physiol.* 127: 1854-1862.
- Mutch, D.M., Berger, A., Mansourian, R., Rytz, A. e Roberts, M.A. (2002). The limit fold change model: A practical approach for selecting differentially expressed genes from microarray data. *BMC Bioinformatics* 3: 17.
- Newton, M.A., Noueiry, A., Sarkar, D. e Ahlquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* 5: 155-176.
- Nogueira, F.T., De Rosa, V.E. Jr., Menossi, M., Ulian, E.C. e Arruda, P. (2003). RNA expression profiles and data mining of sugarcane response to low temperature. *Plant Physiol.* 132:1811-1824.
- Pawitan, Y., Murthy, K.R., Michiels, S. e Ploner, A. (2005). Bias in the estimation of false discovery rate in microarray studies. *Bioinformatics* 21: 3865-3872.
- Pessoa-Jr, A., Roberto, I.C., Menossi, M., dos Santos, R.R., Ortega Filho, S. e Penna, T.C.V. (2005) Perspectives on Bioenergy and Biotechnology in Brazil. *Appl. Biochem. Biotechnol.* 121–124: 59-70.

- Pounds, S. e Morris, S.W. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics* 19: 1236-1242.
- Puri, M.L. e Sen, P.K. (1971). *Nonparametric methods in multivariate analysis*. John Wiley, New York.
- Quackenbush, J. (2001). Computational analysis of microarray data. *Nat. Rev. Genet.* 2: 418-427.
- Quackenbush, J. (2002). Microarray data normalization and transformation. *Nat. Genet.* 32 *Suppl.*: 496-501.
- R Development Core Team (2006). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Raghothama, K.G. (1999). Phosphate acquisition. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 50: 665-693.
- Raghothama, K.G. (2000). Phosphate transport and signaling. *Curr. Opin. Plant Biol.* 3: 182-187.
- Rao, C.R. (2002). *Linear statistical inference and its applications*. John Wiley, New York.
- Rocke, D.M. e Durbin, B. (2003). Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics* 19: 966-972.
- Rousset, P. e Maillet, B. (2005). Increasing reliability of SOMs' neighbourhood structure with a bootstrap process. in *ICANN 2005, LNCS 3696*. W. Duch *et al* (Eds). pp. 433-438.
- Rubio, V., Linhares, F., Solano, R., Martin, A.C., Iglesias, J., Leyva, A. e Paz-Ares, J. (2001). A conserved MYB transcription factor involved in phosphate starvation signaling both in vascular plants and in unicellular algae. *Genes Dev.* 15: 2122-2133.
- Sagi, M., Davydov, O., Orazova, S., Yesbergenova, Z., Ophir, R., Stratmann, J.W. e Fluhr, R. (2004). Plant respiratory burst oxidase homologs impinge

on wound responsiveness and development in *Lycopersicon esculentum*. *Plant Cell* 16: 616-628.

- Salzman, R.A., Brady, J.A., Finlayson, S.A., Buchanan, C.D., Summer, E.J., Sun, F., Klein, P.E., Klein, R.R., Pratt, L.H., Cordonnier-Pratt, M.M. e Mullet, J.E. (2005). Transcriptional profiling of sorghum induced by methyl jasmonate, salicylic acid, and aminocyclopropane carboxylic acid reveals cooperative regulation and novel gene responses. *Plant Physiol.* 138: 352-368.
- Santana, D.P. e Sans, L.M.A. (1999). The distribution and nature of soils with low phosphorous in the brazilian savannas. in *Workshop on Improving Phosphorous Acquisition efficiency in marginal soils*. Embrapa Milho e Sorgo.
- Schachtman, D.P., Reid, R.J. e Ayling, S.M. (1998). Phosphorus Uptake by Plants: From soil to Cell. *Plant Physiol.* 116: 447-453.
- Schena, M., Shalon, D., Davis, R.W. e Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270: 467-470.
- Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O. e Davis, R.W. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. USA.* 93: 10614-10619.
- Schena, M., Heller, R.A., Theriault, T.P., Konrad, K., Lachenmeier, E. e Davis, R.W. (1998). Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol.* 16: 301-306.
- Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H. e Herzel, H. (2000). Normalization strategies for cDNA microarrays. *Nucleic Acids Res.* 28: E47.
- Schummer, M., Ng, W.V., Baumgarner, R.E., Nelson, P.S., Schummer, B., Bednarski, D.W., Hassell, L., Baldwin, R.L., Karlan, B.Y. e Hood, L. (1999). Comparative hybridization of an array of 21500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. *Gene* 238: 375-385.

- Smyth, G.K. e Speed, T. (2003). Normalization of cDNA microarray data. *Methods* 31: 265-273.
- Souza, G.M., Simões, A.C.Q., Oliveira, K.C., Garay, H.M., Fiorini, L.C., Gomes, F.S., Nishiyama-Junior, M.Y. e da Silva, A.M. (2001). SUCAST: prospecting signal transduction in sugarcane. *Genet. Mol. Biol.* 24: 25-34.
- Storey, J.D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B* 64: 479-498.
- Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. e Golub, T.R. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA*. 96: 2907-2912.
- Theodorou, M.E. e Plaxton, W.C. (1994). Induction of Ppi-dependent phosphofructokinase by phosphate starvation in seedlings of *Brassica nigra*. *Plant Cell Environ.* 17: 287-294.
- Thomas, C., Sun, Y., Naus, K., Lloyd, A. e Roux, S. (1999). Apyrase functions in plant phosphate nutrition and mobilizes phosphate from extracellular ATP. *Plant Physiol.* 119: 543-552.
- Tusher, V.G., Tibshirani, R. e Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*. 98: 5116-5121.
- Uhde-Stone, C., Zinn, K.E., Ramirez-Yáñez, M., Li, A., Vance, C.P. e Allan, D.L. (2003). Nylon Filter Arrays Reveal Differential Expression in Proteoid Roots of White Lupin in Response to Phosphorous Deficiency. *Plant Physiol.* 131: 1064-1079.
- 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., Coutinho, L.L., Nóbrega, M.P., Carrer, H., Franca, S.C., Bacci Jr, M., Goldman, M.H., Gomes, S.L., Nunes, L.R., Camargo, L.E., Siqueira, W.J., Van Sluys, M.A., Thiemann, O.H., Kuramae, E.E., Santelli, R.V., Marino, C.L., Targon, M.L., Ferro, J.A., Silveira, H.C., Marini, D.C., Lemos, E.G., Monteiro-Vitorello, C.B., Tambor, J.H., Carraro, D.M., Roberto, P.G., Martins, V.G., Goldman,

- G.H., de Oliveira, R.C., Truffi, D., Colombo, C.A., Rossi, M., de Araújo, P.G., Sculaccio, S.A., Angella, A., Lima, M.M., de Rosa Jr, V.E., Siviero, F., Coscrato, V.E., Machado, M.A., Grivet, L., Di Mauro, S.M., Nóbrega, F.G., Menck, C.F., Braga, M.D., Telles, G.P., Cara, F.A., Pedrosa, G., Meidanis, J. e Arruda, P. (2003). Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. *Genome Res.* 13: 2725-2735.
- Wasaki, J., Yonetani, R., Kuroda, S., Shinano, T., Yazaki, J., Fujii, F., Shimbo, K., Yamamoto, K., Sakata, K., Sasaki, T., Kishimoto, N., Kikuchi, S., Yamagishi, M. e Osaki, M. (2003). Transcriptomic analysis of metabolic changes by phosphorus stress in rice plant roots. *Plant Cell Environ.* 26: 1515-1523.
- Weigel, D. (1995). The APETALA2 domain is related to a novel type of DNA binding domain. *Plant Cell* 7: 388-389.
- Wissenbach, M., Uberlacker, B., Vogt, F., Becker, D., Salamini, F. e Rohde, W. (1993). Myb genes from *Hordeum vulgare*: tissue-specific expression of chimeric Myb promoter/Gus genes in transgenic tobacco. *Plant J.* 4: 411-422.
- Wu, P., Ma, L., Hou, X., Wang, M., Wu, Y., Liu, F. e Deng, X.W. (2003). Phosphate Starvation Triggers Distinct Alterations of Genome Expression in *Arabidopsis* Roots and Leaves. *Plant Physiol.* 132: 1260-1271.
- Wykoff, D.D., Grossman, A.R., Weeks, D.P., Usuda, H. e Shimogawara, K. (1999). Psr1, a nuclear localized protein that regulates phosphorus metabolism in *Chlamydomonas*. *Proc. Natl. Acad. Sci. USA.* 96: 15336-15341.
- Yan, J. (2004). *som: Self-Organizing Map*. R package version 0.3-4.
- Yang, G. e Komatsu, S. (2004). Molecular cloning and characterization of a novel brassinolide enhanced gene OsBLE1 in *Oryza sativa* seedlings. *Plant Physiol. Biochem.* 42: 1-6.
- Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J. e Speed, T.P. (2002). Normalization for cDNA microarray data: a robust composite method

- addressing single and multiple slide systematic variation. *Nucleic Acid Res.* 30: e15.
- Yang, Y.H. e Speed, T. (2002). Design issues for cDNA microarray experiments. *Nat. Rev. Genet.* 3: 579-588.
- Yoshida, K., Kuromitsu, Z., Ogawa, N. e Oshima, Y. (1989). Mode of expression of the positive regulatory genes *PHO2* and *PHO4* of the phosphate regulon in *Saccharomyces cerevisiae*. *Mol. Gen. Genet.* 217: 31-39.
- Yu, H.J., Hogan, P. e Sundaresan, V. (2005). Analysis of the Female Gametophyte Transcriptome of Arabidopsis by Comparative Expression Profiling. *Plant Physiol.* 139: 1853-1869.
- Zhang, H. e Forde, B.G. (1998). An Arabidopsis MADS box gene that controls nutrient-induced changes in root architecture. *Science* 279: 407-409.
- Zhao, N., Hashida, H., Takahashi, N., Misumi, Y. e Sakaki, Y. (1995). High-density cDNA filter analysis: a novel approach for large-scale, quantitative analysis of gene expression. *Gene* 156: 207-213