

A bioinformatic tool for analysis of EST transcript abundance during infection-related development by *Magnaporthe grisea*

DARREN M. SOANES AND NICHOLAS J. TALBOT*

School of Biological and Chemical Sciences, University of Exeter, Washington Singer Laboratories, Perry Road, Exeter EX4 4QG, UK

SUMMARY

Information regarding the levels of mRNA transcript abundance under different conditions, or in specific tissue types, can be obtained by analysis of the frequency of EST sequences in randomly sequenced cDNA libraries. Here we report a bioinformatics tool, which provides a means of identifying genes that are differentially expressed during pathogenesis-related development by the rice blast fungus *Magnaporthe grisea*. A total of 31 534 *M. grisea* ESTs were obtained from dbEST at NCBI, clustered into 8821 unique sequences (unisequences) and manually annotated. Transcript profiles were then calculated for 958 unigenes identified from eight different cDNA libraries. The data were integrated into the Consortium for Functional Genomics of Microbial Eukaryotes (COGEME) database (<http://cogeme.ex.ac.uk/>) and a web-based front end was designed to allow users to access and interrogate the generated datasets.

INTRODUCTION

In order to understand plant diseases at the molecular level, it will be necessary to define accurately the patterns of gene expression that are deployed by pathogenic micro-organisms during plant infection. Studying virulence-associated gene expression can be expected to lead to a deeper insight into the developmental and physiological programmes that pathogenic microbes use to invade and overcome their hosts, and will be important for development of effective control strategies (Yoder and Turgeon, 2001). In the case of fungal pathogens, draft genome sequences are being generated from several pathogenic species and are already available for the rice blast fungus *Magnaporthe grisea*, the head blight pathogen *Fusarium graminearum* and the corn smut fungus *Ustilago maydis* (Dean *et al.*, 2005; <http://www.broad.mit.edu>). The availability of these sequences is allowing microarray and SAGE (Serial Analysis of Gene Expression) analysis of plant infection to begin, although few studies have yet been published (Matsumura *et al.*, 2003;

Thomas *et al.*, 2002; Wan *et al.*, 2002). In the short term therefore there is a need to utilize current genomic data in a quantitative manner and to develop bioinformatics tools that will guide future studies.

Information regarding mRNA transcript abundance under different developmental conditions can be obtained from the quantitative analysis of EST sequences. If ESTs are randomly generated from non-normalized libraries, then the number of EST sequences representing a particular gene will be directly proportional to the mRNA abundance in the tissue from which the library was constructed. A comparison of the frequencies of EST sequences found in a number of different libraries can therefore be used to build up transcript profiles for large sets of genes (Audic and Claverie, 1997). Digital analysis of EST abundance has been carried out in a number of organisms for which large numbers of ESTs are available from a number of different libraries (Ewing *et al.*, 1999; Ogihara *et al.*, 2003). To facilitate such an analysis in *M. grisea* we have obtained 31 534 EST sequences deposited at dbEST (<http://www.ncbi.nlm.nih.gov/dbEST/index.html>). These ESTs have been sequenced from 16 different cDNA libraries (Ebbole *et al.*, 2004; Kim *et al.*, 2001) and are representative of *M. grisea* growing under a range of environmental conditions and distinct developmental stages. We have manually annotated these ESTs, and then used a statistical test to define patterns of differential gene expression predicted by EST abundance.

We describe a bioinformatics tool that can be used to analyse patterns of gene expression from the rice blast fungus *M. grisea* during infection-related development, starvation stress and sexual morphogenesis. This tool, which is freely available on the web, has been incorporated into the Consortium for Functional Genomics of Microbial Eukaryotes (COGEME) EST data warehouse (<http://cogeme.ex.ac.uk/>), an online comparative and functional genomics resource developed for the phytopathogenic fungi research community (Soanes *et al.*, 2002b).

RESULTS AND DISCUSSION

Generation of EST transcript profiles

A total of 31 534 *M. grisea* ESTs (Ebbole *et al.*, 2004; Kim *et al.*, 2001) were used to generate a non-redundant set of 8821 unique

*Correspondence: Tel.: +44 1392264673; Fax: +44 1392264668; E-mail: n.j.talbot@exeter.ac.uk

Library name	Number of ESTs	Strain (s)	Tissue
Mag02	2553	70-15	appressorium
Mag03	1466	70-15	mycelium (grown in minimal medium)
Mag04	3924	CP987	mycelium (grown in medium containing rice cell walls)
Mag06	3193	Guy11	mycelium grown in complete medium
Mag07	3405	Guy11	conidia
Mag08	4310	Guy11	mycelium (grown in nitrogen starvation medium)
Mag10	4421	$\Delta pmk1$	germinated conidia
Mag15	6521	4091-5-8, 4136-4-3	mixed mated culture

*All sequences were downloaded from the dbEST database at NCBI. Sequences were produced by D. J. Ebbole, Department of Plant Pathology & Microbiology, Texas A&M University, TX, except Mag02, which was produced by R. A. Dean, Fungal Genomics Laboratory, North Carolina State University, Raleigh, NC (Ebbole *et al.*, 2004).

sequences (unisequences) through cluster assembly. This represents approximately 79% of the genes in the *M. grisea* genome (Dean *et al.*, 2005). The majority of *M. grisea* ESTs held in dbEST are 5' end cDNA sequences, and in the examination of EST abundance 3' end cDNA sequences were deliberately excluded to avoid the possibility of misclassifying individual ESTs. Additionally, unisequences were selected for quantitative analysis only when they had been assembled from at least five individual ESTs obtained from non-normalized cDNA libraries, from which at least 1000 ESTs had been sequenced. These selection criteria were designed to restrict data analysis to ESTs that were sufficiently represented for subsequent statistical analysis. Of the 16 cDNA libraries from which *M. grisea* ESTs have been generated, only nine libraries had at least 1000 ESTs sequenced from them. One of these was a subtracted cDNA library, which meant that it was not suitable for further study (Kim *et al.*, 2001). The eight cDNA libraries used in this study are described in Table 1. These include cDNA libraries constructed from *M. grisea* appressoria, mated cultures, conidia and mycelium (grown in complete medium, minimal medium, nitrogen starvation medium and with rice cell walls as the sole carbon source, respectively). A cDNA library generated from germinated conidia of a $\Delta pmk1$ mutant was also analysed. *PMK1* encodes a MAP kinase, which is required for appressorium development and plant infection by *M. grisea* (Xu and Hamer, 1996).

A total of 1070 unisequences had at least five constituent ESTs from these eight libraries. By comparison of the unisequences with the draft *M. grisea* genome sequence, we found that ESTs representing one gene sometimes clustered into more than one unisequence. This may be due to there being no significant overlap between ESTs sequenced from the 5' and 3' ends of the gene or more than one mRNA species being transcribed from the gene due to differential splicing or alternative transcriptional start points (Ebbole *et al.*, 2004; Ogihara *et al.*, 2003). Unisequences representing the same gene were identified by comparison with the draft *M. grisea* genome sequence (<http://www-genome.broad.mit.edu/annotation/fungi/magnaporthe/>) and the transcript profiles for these unisequences

Table 1 Details of the cDNA libraries used for the analysis of transcript abundance in *Magnaporthe grisea*.*

were summed. Transcript profiles were produced for 958 unique genes (unigenes).

To analyse the data, statistical analysis was applied using a specific method developed to determine whether the difference in EST sequence abundance between two libraries is significant (Audic and Claverie, 1997). The procedure is based on determining the probability that a difference in the frequency of ESTs between two libraries is due to a difference in expression levels rather than random sampling. The theory links the threshold of selection of putatively regulated genes to the fraction of false positive clones one is willing to risk. We used this statistical method to produce a *P* value for each pair of cDNA libraries for each unigene. Software to perform this analysis is available at the homepage of Stéphane Audic (<http://igs-server.cnrs-mrs.fr/~audic/significance.html>). A UNIX program downloaded from this website was adapted to enable us to analyse large amounts of data automatically. The *P* value indicates the probability that differences in the number of EST transcripts, representing each unigene, sequenced from the two libraries are due to differential expression of the gene rather than chance sampling error (Audic and Claverie, 1997). The *P* value depends on the absolute number of ESTs sequenced corresponding to the gene under each condition. For example, if for a particular gene three ESTs were sequenced from one library and one from another, this would not be considered a significant difference, but 21 ESTs from one library and seven from another would be considered a significant difference even though the ratio is three to one in each case. A *P* value of 0.95 would mean that the difference in EST transcript abundance of a particular gene under two different conditions is likely to be due to a differential expression in 19 out of 20 cases. This method assumes that ESTs are randomly sequenced from non-normalized cDNA libraries (Audic and Claverie, 1997).

The data reported here have been integrated into the COGEME fungal EST database (<http://cogeme.ex.ac.uk/transcript.html>). Users are able to retrieve transcript profiles for unisequences via a link from the information page of each unisequence (Fig. 1). Alternatively, users can choose a pair of cDNA libraries and

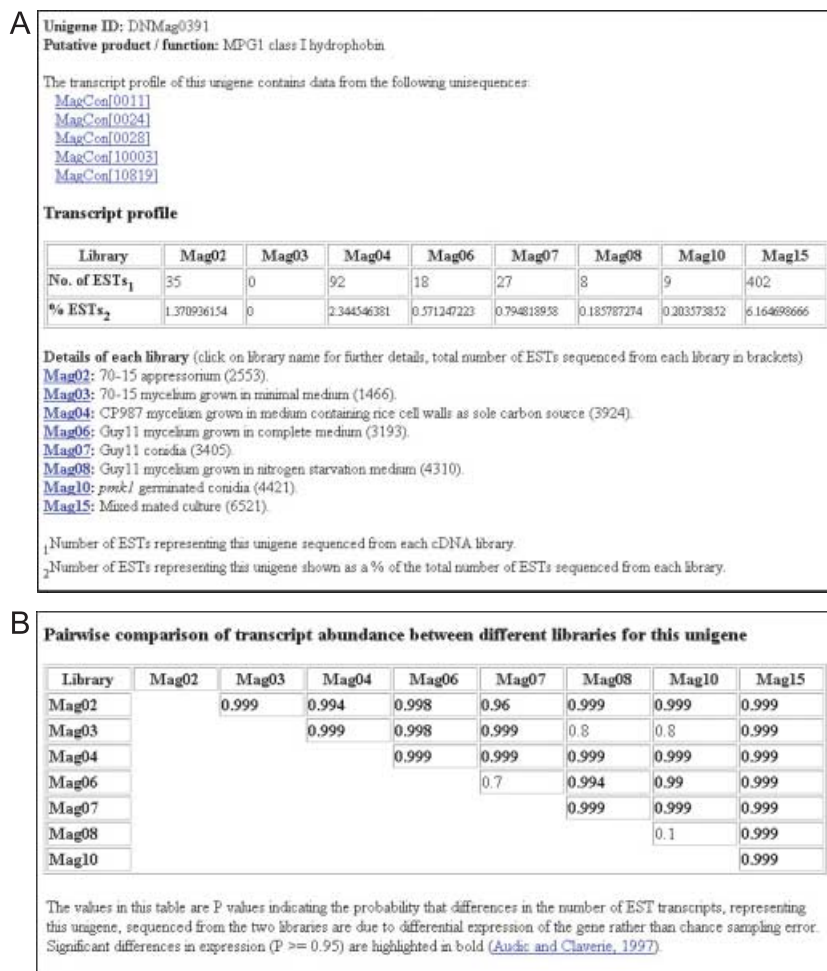


Fig. 1 Screen shots from the web-based front of the COGEME fungal EST database showing a transcript profile for the unigene that encodes the fungal hydrophobin *MPG1*. (A) This shows the number of ESTs representing this unigene sequenced from each cDNA library, as well as the percentage of the total ESTs sequenced from each library that this represents. (B) This matrix shows *P* values representing the probability that differences in EST abundance in pair-wise comparisons between each library are due to differential expression of the gene rather than chance sampling error.

display those unigenes that show significantly different levels of EST transcript abundance between any two cDNA libraries with a minimum *P* value of 0.99 or 0.95 (Fig. 2). This provides an online tool by which gene expression in *M. grisea* under distinct developmental and environmental conditions can be compared.

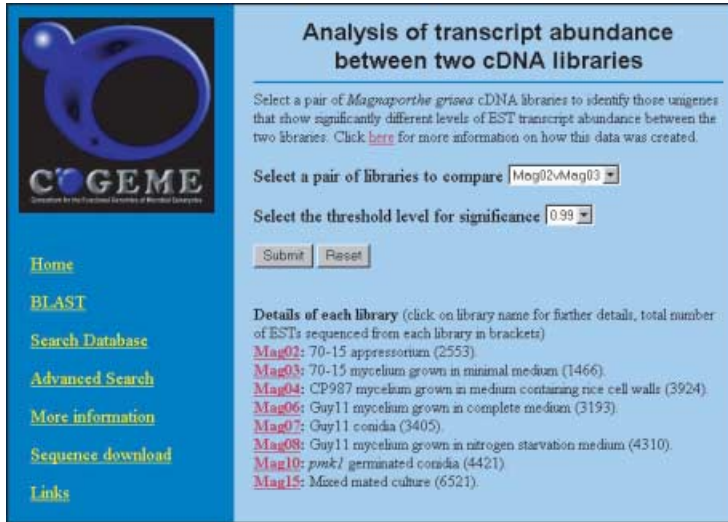
Pathogenesis-associated gene expression in *M. grisea* based on informatic analysis of ESTs

To assess patterns of gene expression that may be associated with plant infection by *M. grisea*, EST frequencies were compared from a cDNA library made from mycelium grown in a medium containing rice cell walls as the sole source of carbon and a cDNA library made from *M. grisea* mycelium grown in a nutrient-rich growth medium. The results from this comparison are shown in Tables 2 and 3. The *MPG1* (DNMag0391) hydrophobin-encoding gene was found to be significantly more highly expressed (P value ≥ 0.99) when *M. grisea* mycelium was grown in the presence of rice cell walls. This is consistent with previous studies, which showed that the gene is highly expressed during initial stages of

plant infection (Ebbole *et al.*, 2004; Matsumura *et al.*, 2003; Soanes *et al.*, 2002a; Talbot *et al.*, 1993, 1996), providing a useful test of the efficacy of the informatic tool. Interestingly, we also identified a number of genes encoding enzymes that may be involved in the breakdown of plant cell walls, such as alpha-L-arabinofuranosidase, endo-beta-1,4-D-xylanase and beta-glucosidase as well as two genes encoding putative sugar transporters that are more highly expressed in the presence of rice cell walls (Table 2). During growth of *M. grisea* in the host plant, extracellular depolymerizing enzymes are likely to be necessary to breakdown plant cell walls and specific transporters would facilitate uptake of the released monosaccharides for utilization by the fungus. In addition, genes that encode enzymes necessary for catabolism of simple sugars and amino acids, such as glyceraldehyde 3-phosphate dehydrogenase and glutamate dehydrogenase, and two chitinase-encoding genes were also up-regulated by the presence of rice cell walls. The latter may be involved in fungal cell wall remodelling during invasive growth of the fungus within the host plant (Rast *et al.*, 1991).

A gene that putatively encodes NAD-dependent formate dehydrogenase showed the highest quantitative difference in

A



Analysis of transcript abundance between two cDNA libraries

Select a pair of *Magnaporthe grisea* cDNA libraries to identify those unigenes that show significantly different levels of EST transcript abundance between the two libraries. Click [here](#) for more information on how this data was created.

Select a pair of libraries to compare

Select the threshold level for significance

Details of each library (click on library name for further details, total number of ESTs sequenced from each library in brackets)

Mag02: 70-15 appressorium (2553)

Mag03: 70-15 mycelium grown in minimal medium (1466)

Mag04: CP987 mycelium grown in medium containing rice cell walls (3924)

Mag06: Guy11 mycelium grown in complete medium (3193)

Mag07: Guy11 conidia (3405)

Mag08: Guy11 mycelium grown in nitrogen starvation medium (4310)

Mag10: *pmk1* germinated conidia (4421)

Mag15: Mated mated culture (6521)

Home

BLAST

[Search Database](#)

[Advanced Search](#)

[More information](#)

[Sequence download](#)

[Links](#)

B

Comparison of transcript profiles of *Mag02* and *Mag03*

Sequences significantly more highly expressed ($P \geq 0.99$) in *Mag02* than *Mag03*

Unigene ID	Putative product / function	Mag02 ₁	Mag03 ₁	P value ₂
DNMag0335	homologue of UVI-1 [<i>Bipolaris oryzae</i>], cell wall protein	30	0	0.999
DNMag0371	MAS3 protein	20	0	0.999
DNMag0391	MPG1 class I hydrophobin	35	0	0.999
DNMag0623	unknown	13	0	0.996
DNMag0815	unknown	11	0	0.991

Sequences significantly more highly expressed ($P \geq 0.99$) in *Mag03* than *Mag02*

Unigene ID	Putative product / function	Mag02 ₁	Mag03 ₁	P value ₂
DNMag0052	60S acidic ribosomal protein P2	0	7	0.999
DNMag0054	60S ribosomal protein L10	0	7	0.999
DNMag0166	calciotum subunit B	0	9	0.999
DNMag0208	coproporphyrinogen oxidase, sixth step in heme biosynthetic pathway	0	7	0.999
DNMag0232	D-hydroxyacid dehydrogenase / D-lactate dehydrogenase	0	10	0.999
DNMag0293	glyceraldehyde 3-phosphate dehydrogenase	3	21	0.999
DNMag0438	pathogenesis related (SnoDProt1)	0	8	0.999
DNMag0592	unknown	8	55	0.999
DNMag0612	unknown	2	55	0.999
DNMag0645	unknown	0	23	0.999
DNMag0650	unknown	0	18	0.999
DNMag0658	unknown	0	8	0.999
DNMag0673	unknown	0	7	0.999
DNMag0074	60S ribosomal protein L26	0	6	0.998
DNMag0260	extracellular chitinase	0	6	0.998
DNMag0560	translation elongation factor-1 gamma (EF-1-gamma)	0	6	0.998
DNMag0670	unknown	0	6	0.998
DNMag0079	60S ribosomal protein L30	1	7	0.996
DNMag0700	unknown	1	7	0.996
DNMag0023	40S ribosomal protein S12	0	5	0.995
DNMag0067	60S ribosomal protein L19	0	5	0.995
DNMag0070	60S ribosomal protein L22	0	5	0.995
DNMag0083	60S ribosomal protein L35	0	5	0.995
DNMag0107	adenosine kinase	0	5	0.995
DNMag0249	enolase (2-phosphoglycerate dehydratase)	0	5	0.995
DNMag0573	ubiquinol-cytochrome c reductase	0	5	0.995
DNMag0849	unknown	0	5	0.995
DNMag0850	unknown	0	5	0.995
DNMag0610	unknown	10	17	0.994
DNMag0036	40S ribosomal protein S25	1	6	0.99
DNMag0712	unknown	1	6	0.99

Fig. 2 Screen shots from the web-based front of the COGEME fungal EST database showing a comparison of EST transcript profiles between libraries *Mag02* and *Mag03*. (A) The researcher can use this screen to specify which pair of libraries to compare and set the minimum *P* value for consideration of a difference in EST frequency as significant (0.99 or 0.95). (B) A list of unigenes that are significantly more highly expressed in one library than the other, split into those that are more highly expressed in *Mag02* (top) and those that are more highly expressed in *Mag03* (bottom).

Table 2 Unigenes that are significantly more highly expressed in library Mag04 (mycelium grown in medium containing rice cell walls) than in Mag06 (mycelium grown in complete medium).

Unigene ID	Putative product/function	Mag04*	Mag06*	P value
DNMag0125	alpha-L-arabinofuranosidase	19	0	0.999
DNMag0127	alternative oxidase	27	5	0.999
DNMag0136	aspartate aminotransferase	32	4	0.999
DNMag0163	beta-glucosidase	79	0	0.999
DNMag0165	C-4 sterol methyl oxidase	28	2	0.999
DNMag0186	cell wall protein	33	5	0.999
DNMag0194	chitinase	18	1	0.999
DNMag0208	coproporphyrinogen oxidase	60	12	0.999
DNMag0232	D-hydroxyacid/b-lactate dehydrogenase	46	0	0.999
DNMag0247	endo-beta-1,4-D-xylanase	31	0	0.999
DNMag0260	extracellular chitinase	23	3	0.999
DNMag0262	fatty acid hydroxylase	14	0	0.999
DNMag0267	flavo-hemoglobin	101	0	0.999
DNMag0315	hexose transporter	74	0	0.999
DNMag0390	monosaccharide transporter	18	1	0.999
DNMag0391	MPG1 class I hydrophobin	92	18	0.999
DNMag0397	NAD-dependent glutamate dehydrogenase	18	0	0.999
DNMag0398	NAD-dependent formate dehydrogenase	291	1	0.999
DNMag0599	unknown	49	1	0.999
DNMag0606	unknown	42	9	0.999
DNMag0645	unknown	128	9	0.999
DNMag0651	unknown	29	1	0.999
DNMag0657	unknown	38	0	0.999
DNMag0660	unknown	12	0	0.999
DNMag0661	unknown	19	0	0.999
DNMag0662	unknown	13	0	0.999
DNMag0670	unknown	47	3	0.999
DNMag0674	unknown	30	0	0.999
DNMag0706	unknown	19	1	0.999
DNMag0765	unknown	16	0	0.999
DNMag0785	unknown	13	0	0.999
DNMag0935	unknown (homologue of Ynr018wp, <i>S. cerevisiae</i>)	23	3	0.999
DNMag0293	glyceraldehyde 3-phosphate dehydrogenase	46	15	0.998
DNMag0421	nodulin-like	11	0	0.998
DNMag0643	unknown	11	0	0.998
DNMag0703	unknown	11	0	0.998
DNMag0704	unknown	18	2	0.998
DNMag0724	unknown	11	0	0.998
DNMag0842	unknown	11	0	0.998
DNMag0863	unknown	11	0	0.998
DNMag0797	unknown	14	1	0.997
DNMag0623	unknown	10	0	0.996
DNMag0014	30-kDa heat shock protein (HSP30)	9	0	0.994
DNMag0231	delta-9 fatty acid desaturase, stearoyl-CoA desaturase	20	4	0.994
DNMag0552	transcription factor (homologue of wc 1, <i>N. crassa</i>)	9	0	0.994
DNMag0893	unknown	9	0	0.994
DNMag0641	unknown	8	0	0.99
DNMag0672	unknown	8	0	0.99
DNMag0679	unknown	8	0	0.99
DNMag0738	unknown	8	0	0.99
DNMag0794	unknown	8	0	0.99

*Number of ESTs representing each unigene sequenced from each cDNA library.

Table 3 Unigenes that are significantly more highly expressed in library Mag06 (mycelium grown in complete medium) than in Mag04 (mycelium grown in medium containing rice cell walls).

Unigene ID	Putative product/function	Mag04*	Mag06*	P value
DNMag0111	aflatoxin biosynthesis ketoreductase	0	37	0.999
DNMag0206	copper transport protein	3	22	0.999
DNMag0207	copper transporter	1	29	0.999
DNMag0238	DnaJ-like protein, cochaperone	0	17	0.999
DNMag0246	elongation factor 1-alpha (EF-1-alpha)	3	22	0.999
DNMag0287	glutathione S-transferase	0	26	0.999
DNMag0366	manganese-superoxide dismutase	0	9	0.999
DNMag0373	methyltransferase	0	141	0.999
DNMag0399	NADH oxidase	0	38	0.999
DNMag0422	norsolorinic acid reductase	0	178	0.999
DNMag0477	pyridoxal reductase (pyridoxine 4-dehydrogenase)	0	13	0.999
DNMag0482	quinone oxidoreductase	0	18	0.999
DNMag0505	scytalone dehydratase	0	23	0.999
DNMag0546	tetrahydroxynaphthalene reductase	0	11	0.999
DNMag0547	tetra/tri hydroxynaphthalene reductase	3	40	0.999
DNMag0592	unknown	4	28	0.999
DNMag0601	unknown	0	13	0.999
DNMag0610	unknown	3	24	0.999
DNMag0626	unknown	0	11	0.999
DNMag0633	unknown	0	14	0.999
DNMag0635	unknown	0	9	0.999
DNMag0649	unknown	4	64	0.999
DNMag0753	unknown	0	15	0.999
DNMag0783	unknown	0	10	0.999
DNMag0791	unknown	0	9	0.999
DNMag0826	unknown	0	11	0.999
DNMag0861	unknown	0	11	0.999
DNMag0006	14-3-3 protein	0	8	0.998
DNMag0286	glutathione S-transferase	0	8	0.998
DNMag0734	unknown	0	8	0.998
DNMag0857	unknown	0	8	0.998
DNMag0326	homologue of CipA (<i>E. nidulans</i>)	0	7	0.996
DNMag0330	homologue of <i>E. nidulans</i> ESDC, involved in sexual development	2	11	0.996
DNMag0343	intracellular protease/amidase	0	7	0.996
DNMag0439	pathogenesis-related protein	0	7	0.996
DNMag0517	short chain dehydrogenase (keto acyl reductase)	0	7	0.996
DNMag0699	unknown	0	7	0.996
DNMag0093	6-phosphogluconate dehydrogenase	0	6	0.993
DNMag0114	alcohol dehydrogenase	0	6	0.993
DNMag0218	cytochrome b5 reductase	0	6	0.993
DNMag0272	gamma actin	0	6	0.993
DNMag0521	small GTP-binding protein YPTI	0	6	0.993
DNMag0561	translation initiation factor 3 (eIF3)	0	6	0.993
DNMag0579	ubiquinone/menaquinone biosynthesis methyltransferase	0	6	0.993
DNMag0629	unknown	0	6	0.993
DNMag0634	unknown	0	6	0.993
DNMag0690	unknown	0	6	0.993

*Number of ESTs representing each unigene sequenced from each cDNA library.

expression between the two libraries (Table 2). In methylotrophic yeast species such as *Candida boidinii*, formate dehydrogenase (FDH) is involved in the detoxification of formate produced by metabolism of methanol, choline and methylamine, as well as contributing significantly to energy yield during growth on methanol by production of NADH (Sakai *et al.*, 1997). FDH is highly expressed during anoxic conditions in plants and may have a role in anaerobic metabolism by producing NADH (Bykova *et al.*, 2003). The high levels of expression of FDH during growth of *M. grisea* in the presence of rice cell walls may be indicative of a switch in fungal metabolism during growth in planta due to anaerobic conditions or available nutrients, or alternatively, the need to detoxify formate produced by the breakdown of antifungal toxins synthesized by the host plant.

By contrast, unigenes that are more significantly highly expressed in mycelium grown in nutrient-rich medium when compared with mycelium grown in the presence of rice cell walls are shown in Table 3. These include a number of genes involved in the synthesis of melanin (such as scytalone dehydratase, tetra/tri hydroxynaphthalene reductase-encoding genes) (Chumley and Valent, 1990), two glutathione S-transferases (Choi *et al.*, 1998), as well as manganese-superoxide dismutase (involved in the breakdown of superoxide, a reactive oxygen species) (Mayer *et al.*, 2001) and two genes that may be involved in the synthesis of a secondary metabolite similar to aflatoxin (aflatoxin biosynthesis ketoreductase, norsolorinic acid reductase) (Cary *et al.*, 1996). Taken together, the pattern of gene expression indicates that secondary metabolism and melanin production accompany rapid hyphal proliferation in axenic culture in a manner that does not occur when the fungus is exposed to the complex rice cell wall substrate that it would encounter during pathogenesis.

To bring about plant infection, *M. grisea* forms specialized infection structures known as appressoria (Dean, 1997; Talbot, 2003). These dome-shaped cells are used to breach the rice leaf cuticle, sending a narrow penetration hypha into the plant. *M. grisea* conidia will form appressoria within 8 h when allowed to germinate on a hydrophobic surface in the presence of water (Lee and Dean, 1994). *PMK1* encodes a mitogen activated protein kinase (MAPK) that is a functional homologue of the MAPK genes *FUS3* and *KSS1* from *Saccharomyces cerevisiae* (Xu and Hamer, 1996). $\Delta pmk1$ strains of *M. grisea* are non-pathogenic and fail to form appressoria on an inductive surface (Talbot, 2003; Xu and Hamer, 1996). The cDNA library Mag02 was constructed using conidia from a wild-type strain of *M. grisea* that were germinated on an inductive surface and formed appressoria (Ebbole *et al.*, 2004). By contrast, cDNA library Mag10 was constructed using conidia from a $\Delta pmk1$ strain of *M. grisea* that had germinated on an inductive surface, but not formed appressoria (Ebbole *et al.*, 2004). Comparison of EST transcript abundances between these two libraries is shown in Tables 4 and 5. Among the most highly abundant ESTs from Mag02 were several cDNAs that putatively encode cell wall

proteins that may be structural components of the appressorium, such as a homologue of the cell wall protein UVI-1 from *Bipolaris oryzae* (Kihara *et al.*, 2001) and the clock-controlled protein 6, which is expressed during conidiogenesis of *Neurospora crassa* (Bell-Pedersen *et al.*, 1996). Transcripts corresponding to the fungal virulence genes *MAS1* (*GAS2*) and *MAS3* (*GAS1*) were also significantly more abundant in the wild-type appressorial library than the $\Delta pmk1$ library. This is consistent with previous studies that have shown that these two genes are expressed only in appressoria and not expressed in $\Delta pmk1$ mutants (Ebbole *et al.*, 2004; Xue *et al.*, 2002). The hydrophobin-encoding gene *MPG1* is also more highly expressed during appressorium formation than in a $\Delta pmk1$ library, consistent with studies that have shown that *MPG1* is positively regulated by Pmk1 (Soanes *et al.*, 2002a). Table 5 shows a list of unigenes that are significantly more highly expressed in Mag10 ($\Delta pmk1$ germinated conidia) than in Mag02 (wild-type appressorium). These include a large number of ribosomal protein-encoding genes as well as a translation elongation factor, suggesting that protein synthesis is relatively active in germ tubes of $\Delta pmk1$ mutants. This may be due to the fact that germ tubes produced by wild-type *M. grisea* conidia cease elongation after 4–8 h and differentiate to form appressoria, whereas germ tubes produced by $\Delta pmk1$ conidia do not form appressoria but carry on elongation (Xu and Hamer, 1996). Two genes encoding enzymes involved in melanin biosynthesis (tetrahydroxynaphthalene reductase and tetra/tri hydroxynaphthalene reductase) are also more highly expressed in the $\Delta pmk1$ library than in the wild-type library. This seems counter-intuitive because melanin is synthesized during appressorium formation and forms a layer that is essential for maintenance of turgor pressure in appressoria (Chumley and Valent, 1990; de Jong *et al.*, 1997). It is possible, however, that regulation of melanin biosynthetic genes occurs prior to appressorium maturation and is perhaps misregulated in the $\Delta pmk1$ mutant, which is unable to form appressoria.

As well as identifying genes in which a putative cellular function can be inferred by similarity to known genes, our *in silico* analysis has also identified genes that have no similarity to known genes (so-called orphans) with differential expression profiles (Schmid and Aquadro, 2001). For example, Table 2 lists 27 unigenes of unknown function that are significantly more highly expressed in liquid culture containing rice cell walls than in complete medium. Table 4 details eight, clear appressorium-specific orphan genes and Table 5 lists three orphans that are de-repressed in *pmk1* mutants.

CONCLUDING REMARKS

In silico analysis of EST transcript abundance provides a low-cost way of identifying genes that are highly expressed in certain tissues, or under specific conditions. Such analyses have been conducted in a number of organisms including rice (Ewing *et al.*, 1999), potato (Ronning *et al.*, 2003), rye (Milla *et al.*, 2002) and wheat (Ogihara

Table 4 Unigenes that are significantly more highly expressed in library Mag02 (appressorium) than in Mag10 (*pmk1*, germinated conidia).

Unigene ID	Putative product/function	Mag02*	Mag10*	P value
DNMag0006	14-3-3 protein	12	3	0.999
DNMag0335	homologue of UVI-1 [<i>Bipolaris oryzae</i>], cell wall protein	30	20	0.999
DNMag0356	laccase	7	0	0.999
DNMag0370	MAS1 protein	8	0	0.999
DNMag0371	MAS3 protein	20	0	0.999
DNMag0391	MPG1 class I hydrophobin	35	9	0.999
DNMag0623	unknown	13	2	0.999
DNMag0815	unknown	11	1	0.999
DNMag0951	vacuolar-ATPase	17	1	0.999
DNMag0328	homologue of clock controlled protein 6 (<i>N. crassa</i>) (cell surface glycoprotein)	11	3	0.998
DNMag0595	unknown	10	2	0.998
DNMag0610	unknown	10	2	0.998
DNMag0186	cell wall protein	13	5	0.997
DNMag0817	unknown	7	1	0.996
DNMag0883	unknown	5	0	0.995
DNMag0906	unknown	5	0	0.995
DNMag0622	unknown	10	4	0.991
DNMag0303	GTPase Rho3	6	1	0.99

*Number of ESTs representing each unigene sequenced from each cDNA library.

Table 5 Unigenes that are significantly more highly expressed in library Mag10 (*pmk1*, germinated conidia) than in Mag02 (appressorium).

Unigene ID	Putative product/function	Mag02*	Mag10*	P value
DNMag0051	60S acidic ribosomal protein P1	0	18	0.999
DNMag0188	cell wall protein (homologue of PhiA, <i>E. nidulans</i> , essential for phialid development)	2	30	0.999
DNMag0241	dTDP-D-glucose 4,6-dehydratase	2	41	0.999
DNMag0454	pol polyprotein, gypsy retrotransposon	0	31	0.999
DNMag0547	tetra/tri hydroxynaphthalene reductase	4	43	0.999
DNMag0614	unknown	5	48	0.999
DNMag0619	unknown	2	25	0.999
DNMag0636	unknown	0	20	0.999
DNMag0029	40S ribosomal protein S18	0	15	0.998
DNMag0242	dTDP-glucose 4,6-dehydratase	0	13	0.996
DNMag0442	peroxisomal hydratase-dehydrogenase-epimerase (multifunctional beta-oxidation protein)	0	13	0.996
DNMag0052	60S acidic ribosomal protein P2	0	12	0.994
DNMag0054	60S ribosomal protein L10	0	12	0.994
DNMag0557	translation elongation factor eEF-1 beta chain	0	12	0.994
DNMag0588	ubiquitin-conjugating enzyme E2-16 kDa (ubiquitin-protein ligase)	0	12	0.994
DNMag0150	ATP synthase subunit 9, lipid-binding protein, mitochondrial	7	35	0.993
DNMag0108	ADP/ATP carrier protein	3	22	0.991
DNMag0546	tetrahydroxynaphthalene reductase	0	11	0.991
DNMag0079	60S ribosomal protein L30	1	15	0.99

*Number of ESTs representing each unigene sequenced from each cDNA library.

et al., 2003). In some cases hierarchical clustering has then been used to identify genes that show similar patterns of gene expression across a number of libraries, or tissues that show similar patterns of expression for a number of genes (Ewing *et al.*, 1999). Rigorous statistical analysis has made it possible to identify changes in EST transcript abundance that are significant and not due to

random sampling error (Audic and Claverie, 1997). There are, of course, limitations to this type of study. Large numbers of ESTs are needed from a number of libraries as smaller EST collections (containing fewer than 400 ESTs) tend to give unreliable results (Ewing *et al.*, 1999). EST transcript analysis therefore only gives reliable data for a small number of highly expressed genes. In this

study, for example, we have presented transcript profiles for only 958 unigenes. The results obtained in this study are consistent with studies of gene expression using Northern blots or reporter genes (Soanes *et al.*, 2002a; Talbot *et al.*, 1993; Xue *et al.*, 2002), indicating the robustness of the informatic process utilized. All the data generated in this study has been integrated into the COGEME fungal EST database and can be freely accessed. Analysis of transcript abundance for other plant pathogenic fungal species will be entered into the database in this format as the data become available. It also may be possible to integrate expression profiles generated by microarray and SAGE analysis with those generated by analysis of EST transcript abundance in the future.

EXPERIMENTAL PROCEDURES

A total of 31 534 EST sequences were downloaded from the dbEST database at NCBI (<http://www.ncbi.nlm.nih.gov/dbEST/index.html>). These had been sequenced from 16 different cDNA libraries (Ebbola *et al.*, 2004; Kim *et al.*, 2001). Sequencher (Gene Codes Corp., MI) was used to cluster the ESTs into 8821 unisequences. These were assigned a putative function and classified using the MIPS (Munich Information Center for Protein Sequences) system, as described previously (Soanes *et al.*, 2002b). All data were uploaded into a version of the COGEME database written using Microsoft Access (Microsoft Corp., WA). For each unisequence, the number of corresponding 5' ESTs sequenced from each non-normalized cDNA library, which contained at least 1000 ESTs, was calculated to generate a transcript profile for each unisequence, using SQL queries written using Microsoft Access. A total of 1070 unisequences were constructed from at least five ESTs from eight cDNA libraries, and were selected for further study. The BLASTN algorithm (Altschul *et al.*, 1990, 1997) was used to compare the 1070 unisequences with the *M. grisea* genomic sequence and a set of putative open reading frames (downloaded from the Broad Institute; <http://www-genome.broad.mit.edu/annotation/fungi/magnaporthe/>), to identify those unisequences that correspond to the same unique gene (Dean *et al.*, 2005). The transcript profiles for these unisequences were summed and transcript profiles were calculated for 958 unigenes. A statistical method was used to identify unigenes that had significant differences in EST transcript abundance in pair-wise comparisons between the eight libraries (Audic and Claverie, 1997) using a UNIX program downloaded from the personal website of Stéphane Audic (<http://igs-server.cnrs-mrs.fr/~audic/significance.html>). This program was adapted to allow it to process large batches of unigenes. Data were generated for each unigene, comprising the EST frequency for each library expressed as an absolute value and as a percentage of the total number of sequences from the library, and a set of *P* values for each pair-wise comparison of the eight libraries. These data were integrated into the COGEME fungal EST database (Soanes *et al.*, 2002b; <http://cogeme.ex.ac.uk/>

transcript.html) and bespoke cgi-perl scripts written to allow users to select transcript profiles for individual unigenes as well as carry out pair-wise comparisons between libraries. A web interface was designed and the database made available on the internet.

ACKNOWLEDGEMENTS

This work was supported by a grant to N.J.T. from the Biotechnology and Biological Sciences COGEME project, which is part of the BBSRC Investigating Gene Function Initiative. Data presented here will form part of the BBSRC e-fungi data warehouse project, integrating distinct functional genomic information from a range of fungal species.

REFERENCES

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
- 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**, 3389–3402.
- Audic, S. and Claverie, J.M. (1997) The significance of digital gene expression profiles. *Genome Res.* **7**, 986–995.
- Bell-Pedersen, D., Shinohara, M.L., Loros, J.J. and Dunlap, J.C. (1996) Circadian clock-controlled genes isolated from *Neurospora crassa* are late night- to early morning-specific. *Proc. Natl Acad. Sci. USA*, **93**, 13096–13101.
- Bykova, N.V., Stensballe, A., Egsgaard, H., Jensen, O.N. and Møller, I.M. (2003) Phosphorylation of formate dehydrogenase in potato tuber mitochondria. *J. Biol. Chem.* **278**, 26021–26030.
- Cary, J.W., Wright, M., Bhatnagar, D., Lee, R. and Chu, F.S. (1996) Molecular characterization of an *Aspergillus parasiticus* dehydrogenase gene, norA, located on the aflatoxin biosynthesis gene cluster. *Appl. Environ. Microbiol.* **62**, 360–366.
- Choi, J.H., Lou, W. and Vancura, A. (1998) A novel membrane-bound glutathione S-transferase functions in the stationary phase of the yeast *Saccharomyces cerevisiae*. *J. Biol. Chem.* **273**, 29915–29922.
- Chumley, F.G. and Valent, B. (1990) Genetic analyses of melanin-deficient, non-pathogenic mutants of *Magnaporthe grisea*. *Mol. Plant–Microbe Interact.* **3**, 135–143.
- Dean, R.A. (1997) Signal pathways and appressorium morphogenesis. *Annu. Rev. Phytopathol.* **35**, 211–234.
- Dean, R.A., Talbot, N.J., Ebbola, D.J., Farman, M.L., Mitchell, T., Orbach, M.J., Thon, M., Kulkarni, R., Xu, J.R., Pan, H., Read, N.D., Lee, Y.H., Carbone, I., Brown, D., Oh, Y.Y., Donofrio, N., Jeong, J.S., Soanes, D.M., Djonovic, S., Kolomiets, E., Rehmeier, C., Li, W., Harding, M., Kim, S., Lebrun, M.H., Bohnert, H., Coughlan, S., Butler, J., Calvo, S., Ma, L.J., Nicol, R., Purcell, S., Nusbaum, C., Galagan, J.E. and Birren, B.W. (2005) The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature*. **434**, 980–986.
- Ebbola, D.J., Jin, Y., Thon, M., Pan, H.Q., Bhattarai, E., Thomas, T. and Dean, R.A. (2004) Gene discovery and gene expression in the rice blast fungus, *Magnaporthe grisea*: analysis of expressed sequence tags. *Mol. Plant–Microbe Interact.* **17**, 1337–1347.

- Ewing, R.M., Kahla, A.B., Poirot, O., Lopez, F., Audic, S. and Claverie, J.M. (1999) Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res.* **9**, 950–959.
- de Jong, J., McCormack, B.J., Smirnoff, N. and Talbot, N.J. (1997) Glycerol generates turgor in rice blast. *Nature*, **389**, 245–246.
- Kihara, J., Sato, A., Okajima, S. and Kumagai, T. (2001) Molecular cloning, sequence analysis and expression of a novel gene induced by near-UV light in *Bipolaris oryzae*. *Mol. Genet. Genomics*, **266**, 64–71.
- Kim, S., Ahn, I.P. and Lee, Y.H. (2001) Analysis of genes expressed during rice–*Magnaporthe grisea* interactions. *Mol. Plant–Microbe Interact.* **14**, 1340–1346.
- Lee, Y.H. and Dean, R.A. (1994) Hydrophobicity of contact surface induces appressorium formation of *Magnaporthe grisea*. *FEMS Microbiol. Lett.* **115**, 71–76.
- Matsumura, H., Reich, S., Ito, A., Saitoh, H., Kamoun, S., Winter, P., Kahl, G., Reuter, M., Krüger, D.H. and Terauchi, R. (2003) Gene expression analysis of plant host–pathogen interactions by SuperSAGE. *Proc. Natl Acad. Sci. USA*, **100**, 15718–15723.
- Mayer, A.M., Staples, R.C. and Gil-ad, N.L. (2001) Mechanisms of survival of necrotrophic fungal plant pathogens in hosts expressing the hypersensitive response. *Phytochemistry*, **58**, 33–41.
- Milla, M.A., Butler, E., Huete, A.R., Wilson, C.F., Anderson, O. and Gustafson, J.P. (2002) Expressed sequence tag-based gene expression analysis under aluminum stress in rye. *Plant Physiol.* **130**, 1706–1716.
- Ogihara, Y., Mochida, K., Nemoto, Y., Murai, K., Yamazaki, Y., Shin, I.T. and Kohara, Y. (2003) Correlated clustering and virtual display of gene expression patterns in the wheat life cycle by large-scale statistical analyses of expressed sequence tags. *Plant J.* **33**, 1001–1111.
- Rast, D.M., Horsch, M., Furter, R. and Gooday, G.W. (1991) A complex chitinolytic system in exponentially growing mycelium of *Mucor rouxii*: properties and function. *J. Gen. Microbiol.* **137**, 2797–2810.
- Ronning, C.M., Stegalkina, S.S., Ascenzi, R.A., Bougri, O., Hart, A.L., Utterbach, T.R., Vanaken, S.E., Riedmuller, S.B., White, J.A., Cho, J., *et al.* (2003) Comparative analyses of potato expressed sequence tag libraries. *Plant Physiol.* **131**, 419–429.
- Sakai, Y., Murdanoto, A.P., Konishim, T., Iwamatsu, A. and Kato, N. (1997) Regulation of the formate dehydrogenase gene, FDH1, in the methylotrophic yeast *Candida boidinii* and growth characteristics of an FDH1-disrupted strain on methanol, methylamine, and choline. *J. Bacteriol.* **179**, 4480–4485.
- Schmid, K.J. and Aquadro, C.F. (2001) The evolutionary analysis of 'orphans' from the *Drosophila* genome identifies rapidly diverging and incorrectly annotated genes. *Genetics*, **159**, 589–598.
- Soanes, D.M., Kershaw, M.J., Cooley, R.N. and Talbot, N.J. (2002a) Regulation of the MPG1 hydrophobin gene in the rice blast fungus *Magnaporthe grisea*. *Mol. Plant–Microbe Interact.* **15**, 1253–1267.
- Soanes, D.M., Skinner, W., Keon, J., Hargreaves, J. and Talbot, N.J. (2002b) Genomics of phytopathogenic fungi and the development of bioinformatic resources. *Mol. Plant–Microbe Interact.* **15**, 421–427.
- Talbot, N.J. (2003) On the trail of a cereal killer: investigating the biology of *Magnaporthe grisea*. *Annu. Rev. Microbiol.* **57**, 177–202.
- Talbot, N.J., Ebbole, D.J. and Hamer, J.E. (1993) Identification and characterisation of MPG1 a gene involved in pathogenicity from the rice blast fungus *Magnaporthe grisea*. *Plant Cell*, **5**, 1575–1590.
- Talbot, N.J., Kershaw, M.J., Wakley, G., de Vries, O.M.H., Wessels, J.G.H. and Hamer, J.E. (1996) MPG1 encodes a fungal hydrophobin involved in surface interactions during infection-related development by the rice blast fungus *Magnaporthe grisea*. *The Plant Cell*, **8**, 985–999.
- Thomas, S.W., Glaring, M.A., Rasmussen, S.W., Kinane, J.T. and Oliver, R.P. (2002) Transcript profiling in the barley mildew pathogen *Blumeria graminis* by serial analysis of gene expression (SAGE). *Mol. Plant–Microbe Interact.* **15**, 847–856.
- Wan, J., Dunning, F.M. and Bent, A.F. (2002) Probing plant–pathogen interactions and downstream defense signaling using DNA microarrays. *Funct. Integr. Genomics*, **2**, 259–273.
- Xu, J.R. and Hamer, J.E. (1996) MAP kinase and cAMP signalling regulate infection structure formation and growth in the rice blast fungus *Magnaporthe grisea*. *Genes Dev.* **10**, 2696–2706.
- Xue, C., Park, G., Choi, W., Zheng, L., Dean, R.A. and Xu, J.R. (2002) Two novel fungal virulence genes specifically expressed in appressoria of the rice blast fungus. *Plant Cell*, **14**, 2107–2119.
- Yoder, O.C. and Turgeon, B.G. (2001) Fungal genomics and pathogenicity. *Curr. Opin. Plant Biol.* **4**, 315–321.