Mycol. Res. **108** (8): 853–857 (August 2004). © The British Mycological Society
DOI: 10.1017/S095375620400067X *Printed in the United Kingdom.*

Combining transcriptome data with genomic and cDNA sequence alignments to make confident functional assignments for *Aspergillus nidulans* genes

Andrew H. SIMS[1], Manda E. GENT[1], Geoffrey D. ROBSON[1], Nigel S. DUNN-COLEMAN[2] and Stephen G. OLIVER[1]*

[1] *School of Biological Sciences, University of Manchester, The Michael Smith Building, Oxford Road, Manchester M13 9PT, UK.*
[2] *Genencor International Inc., 925 Page Mill Road, Palo Alto, CA 94304, USA.*
E-mail: steve.oliver@man.ac.uk

Received 10 February 2004; accepted 16 April 2004.

Whole genome sequencing of several filamentous ascomycetes is complete or in progress; these species, such as *Aspergillus nidulans*, are relatives of *Saccharomyces cerevisiae*. However, their genomes are much larger and their gene structure more complex, with genes often containing multiple introns. Automated annotation programs can quickly identify open reading frames for hypothetical genes, many of which will be conserved across large evolutionary distances, but further information is required to confirm functional assignments. We describe a comparative and functional genomics approach using sequence alignments and gene expression data to predict the function of *Aspergillus nidulans* genes. By highlighting examples of discrepancies between the automated genome annotation and cDNA or EST sequencing, we demonstrate that the greater complexity of gene structure in filamentous fungi demands independent data on gene expression and the gene sequence be used to make confident functional assignments.

INTRODUCTION

The ongoing, rapid release of whole genome sequences, clone libraries, EST databases, and automated annotation provides large and valuable resources to the biological research community. The growing use of microarrays as a standard tool of choice for many molecular biologists, coupled with the falling costs of genome sequencing, means the need for confident functional assignments is becoming ever greater.

For organisms with limited genome annotation, the results of transcriptome analyses are likely to consist of a long list of 'unknown' sequences that are up- and down-regulated in response to an applied experimental condition. To bring meaning to transcriptome studies, we need to strive towards finding the identity of these 'unknowns'. Much of the information required for meaningful annotation of newly sequenced genomes is already present in publicly available databases such as NCBI and GenBank, as many genes are highly conserved across large evolutionary distances. Despite the advances of automated annotation, some level of manual curation is often required to increase the confidence of the true identity of hypothetical, putative or probable proteins. cDNA sequences can provide valuable supporting evidence for genome annotation (Oliver 1996), the translated sequence of hypothetical genes is crucial to determining function as errors in translation can lead to incorrect functional assignments. A genomics approach using relatively basic bioinformatic tools such as sequence comparisons (BLAST) and alignments (e.g. clustalW) in association with transcriptome data can produce highly valuable, confident functional assignments for genes that have not previously been isolated or characterised.

The genus *Aspergillus* is phylogenetically located in the *Ascomycota* and contains a particularly diverse range of species that includes examples of human and plant pathogens (Rustom 1997, Denning *et al.* 2002), industrially important organisms used in the production of enzymes and organic acids (Conesa *et al.* 2001) and, in *A. nidulans*, a model eukaryotic organism (Martinelli 1994).

In the sequencing efforts, *A. nidulans* was given high priority status and was the first from a candidate list of 15 fungal genomes to be sequenced by the Whitehead Institute (http://www-genome.wi.mit.edu/) under the Fungal Genome Initiative (FGI). The whole genome assembly (30.1 MB) was released in March 2003 and a limited, automated genome annotation (using

* Corresponding author.

Genewise, FgeneSH, FgeneSH+) consisting of 9541 putative open reading frames (ORFs) was released in June 2003. We have produced *A. nidulans* microarrays representing approximately 5800 ESTs from conidial and stress response cDNA libraries, plus over 300 PCR products representing sequences in GenBank or other putative *A. nidulans* genes (Sims *et al.* 2004). However, less than 10% of the *A. nidulans* putative genes have been isolated or identified prior to whole genome sequencing, highlighting both the need and potential for functional genomics in this organism. Here we demonstrate that combining transcriptome data with sequence comparisons and alignments can produce functional assignments for genes.

MATERIALS AND METHODS

Wild-type *Aspergillus nidulans* (FGSC A1004) was used for the glucose up-shift experiment as described in Sims *et al.* (2004). The transcriptome of chymosin-producing pyrG-recombinant *A. nidulans* strain (Cullen *et al.* 1987) was compared to that of its parent strain transformed with an empty vector. Both strains were grown in chemostat cultures on 5% SCM media (Ward *et al.* 1990). Fermentations were performed in a temperature-, pH- and agitation-controlled Braun Biostat fermenter at 30 °C, pH at 5.5, 1000 rpm with a working volume of 2.1 l with on-line carbon dioxide monitoring. Dry weight samples from the vessel and overflow were taken to confirm steady-state growth (with a dilution rate of 0.1 h^{-1}). Chymosin activity was measured using a simple microtitre plate method based on Emtage *et al.* (1983). RNA extraction, labeling and hybridization were performed based upon methods described in Hedge *et al.* (2002). Differential expression was calculated after global normalization in MaxDView (http://www.bioinf.man.ac.uk/microarray/). Evolutionary distances were calculated based upon multiple sequence alignments using the MultAlin program (Corpet 1988; http://prodes.toulouse.inra.fr/multalin/multalin.html) with the default (blosum62) matrix. Putative identities of the genes represented by the ESTs were found by BLASTx (blosum62) comparisons to NCBI GenBank. The hybridizations for the secretion experiment were performed on a microarray containing 4100 ESTs from a conidial library (Sims *et al.* 2004), plus an additional 1700 ESTs from subtraction and stress response libraries (Ayoubi *et al.* 2002) and over 300 additional PCR products, amplified using primers designed from the Whitehead Institute *A. nidulans* genome sequence.

RESULTS AND DISCUSSION

Transcriptome data can be used as a starting point for deducing the coding sequences and functions of genes that have not previously been isolated or sequenced. BLAST searches of EST sequences are used to generate putative identities. For instance, translated sequence comparisons (BLASTx) suggested that two *A. nidulans* ESTs that were significantly down-regulated during a glucose up-shift experiment (Sims *et al.* 2004) had high sequence similarity (e-values <1e-10) to different isoforms of malate dehydrogenase (MDH) for many widely diverse organisms. In the model eukaryote, *S. cerevisiae*, there are three different isozymes of malate dehydrogenase, differentiated by their sub-cellular location, mitochondrial (Mdh1p), cytoplasmic (Mdh2p) and peroxisomal (Mdh3p) (Steffan & McAlister-Henn 1992).

The ORFs of the three yeast *MDH* genes were used as '*in-silico* probes' by comparing their nucleotide sequences against the *Aspergillus nidulans* Whitehead genome sequence using BLASTn. Three distinct regions of the *A. nidulans* genome were found that had significant homology to the three yeast ORFs, suggesting that *A. nidulans* also has three isozymes of MDH. A comparison of the genomic nucleotide sequences with the yeast genes was not sufficient to deduce which genomic sequence encoded which isozyme. If a phylogenetic tree of the nucleotide sequences is constructed, then the three yeast isoforms cluster together (Fig. 1A), showing greater similarity to one another than to any of the *A. nidulans* sequences. This is presumably due to the presence of introns in the *Aspergillus* genomic sequences.

The coding regions of the genomic sequences were translated and assembled based their similarity to the amino-acid sequence of each of the yeast MDH isoforms (using pairwise BLASTx). Multiple sequence alignments with hierarchical clustering were performed (Corpet 1988) to compare the translated genomic sequences with each yeast MDH isozyme to see which showed the highest level of identity (Fig. 1B). One genomic sequence was found to represent each yeast isoform, suggesting that *A. nidulans* also has three corresponding isoforms of MDH, tentatively named MdhA (mitochondrial), MdhB (cytoplasmic) and MdhC (peroxisomal).

In the University of Oklahoma cosmid and cDNA sequencing database (http://www.genome.ou.edu/fungal.html), a full-length cosmid sequence (contig 1670) and a partial cDNA sequence (m8h04a1.r1) were found that were exact matches to the genomic regions of the putative *mdhA* and *mdhC* genes. These were highly valuable to corroborate the positions of the four and eight introns respectively. Release of the *A. nidulans* predicted gene set identified the putative sequences *mdhA* and *mdhC* as the corresponding hypothetical genes AN6717.1 and AN6499.1 (Table 1). However, there appear to be two anomalies with the automated annotation, which failed to predict a 54-nucleotide intron in AN6717.1 (within exon 3) that would be anticipated by sequence similarity with the yeast *MDH1* ortholog and is confirmed by the cDNA sequence for *mdhA* (Fig. 2A). In addition, the hypothetical gene AN6499.1 contains only six introns, compared to the eight predicted by sequence alignment to *MDH3* and

Table 1. Functional assignments for *Aspergillus nidulans* genes based on genomic sequence and transcriptome data.

Gene name	Whitehead locus	Microarray data±	PCR/EST	Top BLAST results/expected ortholog
Glucose up-shift genes				
mdhA	AN6717.1	Down	N9-B10-SP6	malate dehydrogenase precursor [*Nucellala pillus*]
mdhB	AN5031.1			malate dehydrogenase precursor [*Schizosaccharomyces. pombe*]
mdhC	AN6499.1	Down	contig_1644*	malate dehydrogenase [*Mus musculus*]
pdhA	AN5162.1	Up	contig_4018*	pyruvate dehydrogenase E1 component alpha subunit [*Pichia stipitis*]
		Up	contig_2803*	pyruvate dehydrogenase E1 component alpha subunit [*Kluyveromyces lactis*]
mstB	AN2475.1	Up	contig_2004*	putative sugar transporter mstB [*Aspergillus nidulans*]
Secretion-related genes				
prpA	AN0248.1	Up	PCR	alternative protein disulphide isomerase [*Aspergillus niger*]
tigA	AN0075.1	Up	PCR	alternative protein disulphide isomerase [*Aspergillus niger*]
pdiA	AN7436.1	Up	PCR	protein disulphide isomerase [*Aspergillus niger*]
		Up	contig_3071*	ER resident chaperone bipA [*Aspergillus awamori*]
bipA	AN2062.1	Up	contig_3521*	ER resident chaperone bipA [*Aspergillus awamori*]
		Up	contig_240*	ER resident chaperone bipA [*Aspergillus awamori*]

* 1999Oct161435. PCR/EST sequences are available from Pipeonline 2.0 (http://bioinfo.okstate.edu/pipeonline/).

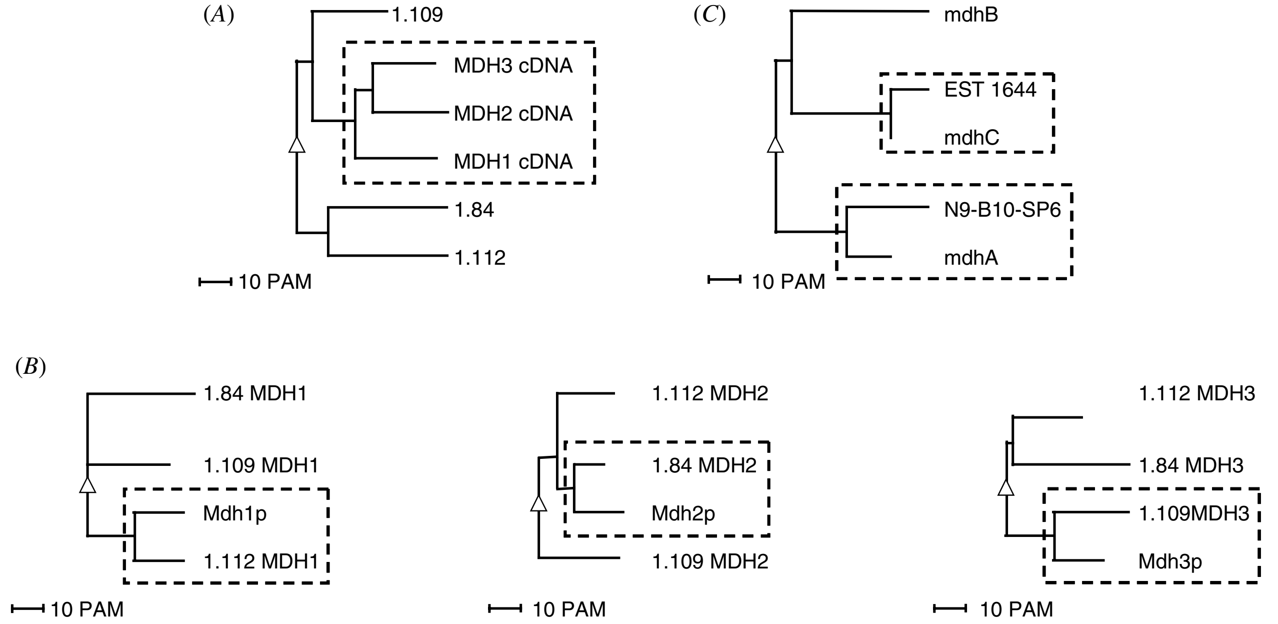

Fig. 1. Hierarchical clustering of sequence alignments representing malate dehydrogenase isozymes. (*A*) Nucleotide sequences of yeast malate dehydrogenase isozymes (cDNA) and *Aspergillus nidulans* genomic sequences (contigs). (*B*) Yeast MDH and translated *A. nidulans* genomic amino acid sequences. (*C*) Nucleotide sequences of yeast cDNA and *A. nidulans* EST sequences. Relative evolutionary distances are shown in PAM units based on multiple sequence alignments, performed using MultAln program (Corpet 1988; http://prodes.toulouse.inra.fr/multalin/multalin.html) with default (blosum62) matrix. The dashed boxes highlight sequences with greatest similarity. Triangles represent the roots of the trees.

the cDNA sequence (Fig. 2B). The deduced amino-acid sequence of *A. nidulans* MdhC protein has a PKL tripeptide, similar to the characteristic SKL of carboxy-terminal peroxisomal targeting sequences (McAlister-Henn *et al.* 1995). The hypothetical gene AN5031.1 is an exact match to the sequence predicted by sequence alignments to *MDH2*, although no sizable cDNA sequence was found to corroborate this assignment. The nucleotide sequences of the two ESTs represented on the microarray were compared with the *A. nidulans* putative cDNA sequences and found to be perfect matches to *mdhA* and *mdhC* (Fig. 1C). Both the mitochondrial and peroxisomal MDH genes would be expected to be down-regulated in the presence of glucose (McAlister-Henn *et al.* 1995), so the transcriptome data also supports the functional assignments (Table 1).

Other ESTs up-regulated during the glucose up-shift experiment include a sequence provisionally identified by BLASTx as the *A. nidulans* putative sugar transporter, MstB (Ventura *et al.*, unpubl), GenBank submission ANI278285. The automated annotation for the corresponding hypothetical gene (AN2475.1) failed to predict the 5′ exon and intron (Fig. 2D), but the transcriptome data supports the functional assignment, since this gene was up-regulated in the presence of glucose.

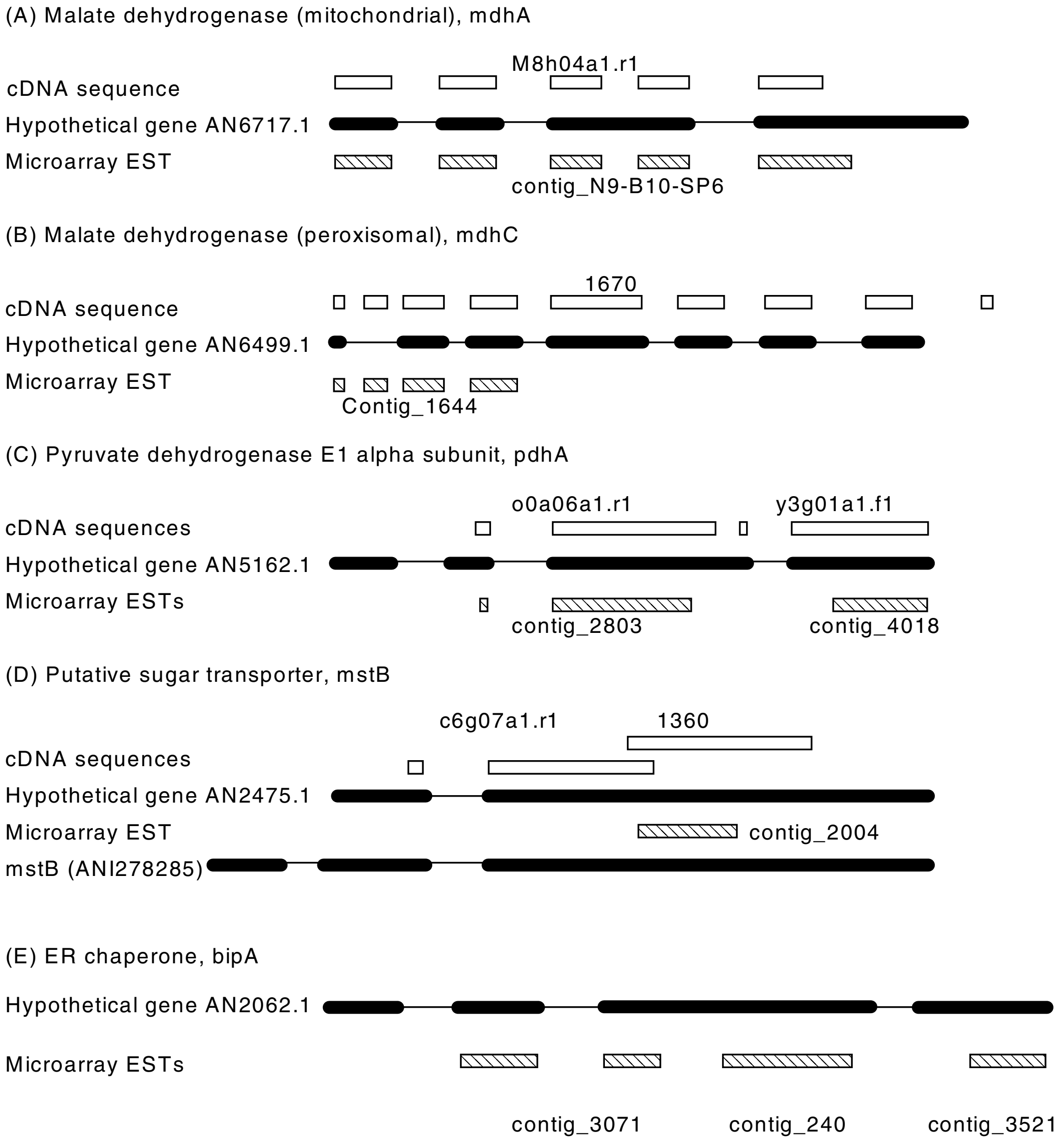
(A) Malate dehydrogenase (mitochondrial), mdhA
M8h04a1.r1
cDNA sequence
Hypothetical gene AN6717.1
Microarray EST
contig_N9-B10-SP6
(B) Malate dehydrogenase (peroxisomal), mdhC
1670
cDNA sequence
Hypothetical gene AN6499.1
Microarray EST
Contig_1644
(C) Pyruvate dehydrogenase E1 alpha subunit, pdhA
o0a06a1.r1
y3g01a1.f1
cDNA sequences
Hypothetical gene AN5162.1
Microarray ESTs
contig_2803
contig_4018
(D) Putative sugar transporter, mstB
c6g07a1.r1
1360
cDNA sequences
Hypothetical gene AN2475.1
Microarray EST
contig_2004
mstB (ANI278285)
(E) ER chaperone, bipA
Hypothetical gene AN2062.1
Microarray ESTs
contig_3071
contig_240
contig_3521

Fig. 2. *Aspergillus nidulans* deduced genes. ■, exons; —, introns; □, cDNA sequences; ▧, ESTs. Parallel blocks represent alignment of homologous sequences. cDNA sequences can be found at the University of Oklahoma cosmid and cDNA sequencing database (http://www.genome.ou.edu/fungal.html).

Two further *A. nidulans* ESTs that were up-regulated during the glucose up-shift experiment had BLASTx matches to pyruvate dehydrogenase E1 alpha subunit from various species. The EST sequences were found closely situated in the genome but, despite finding matching partial cDNA sequences (Fig. 2A), it was not possible to elucidate the full amino-acid sequence due to the low conservation of sequence at the 5′ end, which made determining the exact position of the start codon unreliable. Release of the *A. nidulans* gene list revealed that the two EST sequences are perfect sequence matches to hypothetical protein AN5162.1, predicted by the automated annotation as, 'autocalled dehydrogenase E1 component gene', which supports the putative identity. In addition, one of the EST sequences and both of the cDNA sequences flank introns 2 and 3 predicted for the hypothetical gene (Fig. 2C), demonstrating the value of using ESTs and cDNAs to elucidate the coding sequence.

Most previous investigations into the secretory pathway of filamentous fungi have been conducted using *Aspergillus niger* (Conesa *et al.* 2001), which is used for the industrial production of native and recombinant enzymes. We identified equivalent sequences for 20 known secretion-related genes using the Whitehead genome sequence and sequence alignments (clustalW) to the appropriate *A. niger* genes. PCR primers were designed based on these sequences and used to amplify exclusively part of the exon regions of the *A. nidulans* genes. The PCR products were added to our *A. nidulans* microarrays (Sims *et al.* 2004). Transcriptome data, obtained using the modified array (in preparation) was in accordance with previous work on *A. niger* (Punt *et al.* 1998, Wang & Ward 2000, Ngiam *et al.* 2000), showing that four secretion-related genes (*bipA*, *pdiA*, *prpA* and *tigA*) are up-regulated in a recombinant strain producing the mammalian protein chymosin as compared to a parental wild-type strain (Table 1). Three of these genes were represented on the array by sequences generated by PCR, using primers designed based on the deduced sequences of protein disulphide isomerases, *pdiA*, *prpA* and *tigA*. In addition, 3 ESTs with putative BLAST matches to the ER chaperone, BipA (Fig. 2E) were up-regulated. These findings support the assignment of the hypothetical genes as orthologs of the *A. niger* genes. There is 100 % concordance between the sequences of *bipA*, *pdiA* and *tigA* predicted by automated annotation and by sequence similarity alignments. However, exon 2 of the hypothetical gene for *prpA* (AN0248.1) has 15 additional nucleotides (five additional in-frame amino acids), than would be predicted by sequence alignments to the *A. niger* ortholog, no matching cDNA or EST sequences were found in the databases to confirm the coding sequence.

The examples presented here demonstrate how sequence comparisons and alignments of EST or cDNA sequences with a whole genome sequence can be used in conjunction with transcriptome data to confirm identities generated by BLAST matches and provide confident annotation for both the microarray and the genome. The initial gene annotation of *Saccharomyces cerevisiae* (Goffeau *et al.* 1996, Mewes *et al.* 1997) was performed without recourse to EST sequence data. However, <5% of this yeast's genes contain introns, and intron-containing genes usually have a single intron at the start of the coding sequence, often interrupting the initiator codon. Although the filamentous ascomycetes are close relatives of *S. cerevisiae*, their genes are far more complex, often containing multiple introns (Kupfer *et al.* 1997). Moreover, studies on a number of filamentous fungi have revealed the presence of in-frame introns in some genes (Birch *et al.* 1995). The greater complexity of gene structure in *Aspergillus* and other filamentous fungi demands that independent data on gene expression and function be used to inform and refine the automatic annotation.

The approaches described here are applicable to all species and we encourage individual groups to utilise methods such as these to improve the annotation of microarrays and genomes. We would also emphasise the importance of community-wide efforts to provide experimental evidence for functional assignments in order to reinforce automated annotations and provide reliable identities for fungal genes.

ACKNOWLEDGEMENTS

Genencor International Inc. are thanked for funding for M. E. G. and for a CASE award to A. H. S. who was additionally supported by a BBSRC studentship. This work was carried out at the Transcriptome Resource Facility of the COGEME consortium, which is part of the Investigating Gene Function Initiative of the BBSRC. We would like to thank Rolf Prade for the kind gift of the ESTs.

REFERENCES

Ayoubi, P., Jin, X., Leite, S., *et al.* (2002) PipeOnline 2.0: automated EST processing and functional data sorting. *Nucleic Acids Research* **30**: 4761–4769.

Birch, P. R., Sims, P. F. G. & Broda, P. (1995) Substrate-dependent differential splicing of introns in the regions encoding the cellulose binding domains of two exocellobiohydrolase I-like genes in *Phanerochaete chrysosporium. Applied and Environmental Microbiology* **61**: 3741–3744.

Conesa, A., Punt, P. J., van Luijk, N. & van den Hondel, C. A. M. J. J. (2001) The secretion pathway in filamentous fungi: a biotechnological view. *Fungal Genetics and Biology* **33**: 155–171.

Corpet, F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Research* **16**: 10881–10890.

Cullen, D., Gray, G. L., Wilson, L. J., *et al.* (1987) Controlled expression and secretion of bovine chymosin in *Aspergillus nidulans. Bio-technology* **5**: 369–376.

Denning, D. W., Anderson, M. J., Turner, G., Latge, J.-P. & Bennett, J. W. (2002) Sequencing the *Aspergillus fumigatus* genome. *Lancet Infectious Diseases* **2**: 251–253.

Emtage, J. S., Angal, S., Doel, M. T., *et al.* (1983) Synthesis of calf prochymosin (prorennin) in *Escherichia coli. Proceedings of the National Academy of Sciences, USA* **80**: 3671–3675.

Goffeau, A., Barrell, B. G., Bussey, H., *et al.* (1996) Life with 6000 genes. *Science* **274**: 546, 563–567.

Hegde, P., Qi, R., Abernathy, K., *et al.* (2000) A concise guide to cDNA microarray analysis. *Biotechniques* **29**: 548–554, 556.

Kupfer, D. M., Reece, C. A., Clifton, S. W., Roe, B. A. & Prade, R. A. (1997) Multicellular ascomycetous fungal genomes contain more than 8000 genes. *Fungal Genetics and Biology* **21**: 364–372.

Martinelli, S. D. (1994) *Aspergillus* as an experimental organism. In *Aspergillus: 50 years on* (S. D. Martinelli & J. R. Kinghorn, eds): 33–58. Elsevier, Amsterdam.

McAlister-Henn, L., Steffan, J. S., Minard, K. I. & Anderson, S. L. (1995) Expression and function of a mislocalized form of peroxisomal malate dehydrogenase (MDH3) in yeast. *Journal of Biological Chemistry* **270**: 21220–21225.

Mewes, H. W., Albermann, K., Bahr, M., *et al.* (1997) Overview of the yeast genome. *Nature* **387** (Suppl.): 7–65.

Ngiam, C., Jeenes, D. J., Punt, P. J., van den Hondel, C. A. & Archer, D. B. (2000) Characterization of a foldase, protein disulfide isomerase A, in the protein secretory pathway of *Aspergillus niger. Applied and Environmental Microbiology* **66**: 775–782.

Oliver, S. G. (1996) From DNA sequence to biological function. *Nature* **379**: 597–600.

Punt, P. J., van Gemeren, I. A., Drint-Kuijvenhoven, J., *et al.* (1998) Analysis of the role of the gene bipA, encoding the major endoplasmic reticulum chaperone protein in the secretion of homologous and heterologous proteins in black Aspergilli. *Applied Microbiology and Biotechnology* **50**: 447–454.

Rustom, I. Y. S. (1997) Aflatoxin in food and feed: occurrence, legislation and inactivation by physical methods. *Food Chemistry* **59**: 57–67.

Sims, A. H., Robson, G. D., Hoyle, D. C., *et al.* (2004) Use of expressed sequence tag (EST) analysis and cDNA microarrays of the filamentous fungus *Aspergillus nidulans. Fungal Genetics and Biology* **41**: 199–212.

Steffan, J. S. & McAlister-Henn, L. (1992) Isolation and characterization of the yeast gene encoding the MDH3 isozyme of malate dehydrogenase. *Journal of Biological Chemistry* **267**: 24708–24715.

Wang, H. & Ward, M. (2000) Molecular characterization of a PDI-related gene prpA in *Aspergillus niger var. awamori. Current Genetics* **37**: 57–64.

Ward, M., Wilson, L. J., Kodama, K. H., *et al.* (1990) Improved production of chymosin in *Aspergillus* by expression as a glucoamylase-chymosin fusion. *Biotechnology* **8**: 435–440.

Corresponding Editor: P. Hooley