 DOI: 10.1111/J.1364-3703.2005.00317.X

Comparative genomic analysis of phytopathogenic fungi using expressed sequence tag (EST) collections

DARREN M. SOANES AND NICHOLAS J. TALBOT*

School of Biosciences, University of Exeter, Washington Singer Laboratories, Perry Road, Exeter EX4 4QG, UK

SUMMARY

We describe the analysis of 57 727 unique expressed sequence tags (ESTs) from 15 species of phytopathogenic and three species of saprophytic fungi. This resource is held within the COGEME phytopathogen EST database (http://cogeme.ex.ac.uk/). Comparative analysis was performed to investigate the differences between pathogenic and free-living fungi based on a substantial collection of expressed gene sequences and available, completed fungal genome sequences. We report that the expressed gene inventories of pathogenic fungi were not significantly more similar to each other than to those of free-living filamentous fungi. As expected, however, filamentous fungi as a group share more sequences in common than with the free-living yeast species *Saccharomyces cerevisiae*. Interestingly, ESTs of the obligate biotrophic fungus *Blumeria graminis* f. sp. *hordei* were more dissimilar to those of all other fungal species assessed, having a lower number of sequences in common with filamentous ascomycetes studied to date and also possessing a larger proportion of unisequences of unknown function. Our analysis of ESTs in the COGEME database enabled identification of a set of functional groups of genes that are more highly represented in the genomes of pathogenic fungi than non-pathogenic species.

INTRODUCTION

Fungal genome analysis has progressed rapidly since completion of the genome of the budding yeast *Saccharomyces cerevisiae* (Goffeau *et al*., 1996). The next two genomes to be sequenced were those of the fission yeast *Schizosaccharomyces pombe* (Wood *et al*., 2002) and the pink bread mould *Neurospora crassa* (Galagan *et al*., 2003). More recently, the first basidiomycete genome sequence of the white rot fungus *Phanerochaete chrysosporium* was completed (Martinez *et al*., 2004) and the sequences of two pathogenic fungi, the rice-blast fungus *Magnaporthe grisea* (Dean *et al*., 2005) and the human pathogen *Cryptococcus neoformans* (Loftus *et al*., 2005), were reported. Draft genome sequences are also available for a number of other fungal species including the phytopathogens *Gibberella zeae* (*Fusarium graminearum*) and *Ustilago maydis*, as well as the model saprophyte *Emericella* (*Aspergillus*) *nidulans*.

Since its first description, the yeast genome has proved to be an invaluable resource for researchers studying a range of fungal species including phytopathogens. Many genes vital for the ability of a fungus to infect the host plant, for example, were first identified by homology to genes originally identified to encode components of signalling pathways in *S. cerevisiae*. These include genes encoding mitogen-activated protein (MAP) kinases (see review by Xu, 2000), cyclic AMP-dependent protein kinases (DeZwaan *et al*., 1999; Mitchell and Dean, 1995), adenylate cyclases (Choi and Dean, 1997) and G-protein-coupled receptors (Liu and Dean, 1997).

As more fungal genomes are sequenced, comparative genome analysis becomes an increasingly powerful tool to investigate the evolutionary relationships between related fungal species. The genomes of four closely related species of hemiascomycete yeast have, for instance, recently been compared (Dujon *et al*., 2004) and by classifying predicted proteins into gene families, the extent of genome redundancy and conservation of synteny between the five species of yeast was assessed, pinpointing major events in the evolutionary history of the hemiascomycetes. These events include a whole-genome duplication in the common ancestor of *S. cerevisiae* and *Candida glabrata* that occurred after the divergence from the *Kluyveromyces lactis* lineage (Dujon *et al*., 2004).

Comparison of a large collection of expressed sequence tags (ESTs) from *Neurospora crassa* to the *S. cerevisiae* genome has also provided a large amount of useful information about the differences between the genomes of the free-living yeast and the morphologically more complex filamentous fungus, including novel genes in the *Neurospora* genome and the loss of specific genes in the *S. cerevisiae* lineage (Braun *et al*., 2000).

Molecular phylogeny has shown that pathogenic fungi are found in many taxonomic groups, which suggests that these lifestyles have evolved repeatedly within the kingdom fungi (Tunlid and Talbot, 2002). Comparison of sequence data from phytopathogenic

*Correspondence: Tel.: +44 (0)139 2264673; Fax: +44 (0)139 2264668; E-mail: n.j.talbot@exeter.ac.uk

© 2006 BLACKWELL PUBLISHING LTD

fungi with those of closely related non-pathogenic species should therefore provide information about genetic factors that may be vital for pathogenesis, or at least consistently conserved in species with the capacity to cause fungal disease. In this study, we have analysed and compared EST data from a wide range of phytopathogenic and non-pathogenic fungi and the available fungal genome sequences. We show that there are significant differences in the expressed gene inventories between the genomes of pathogens and non-pathogens, with groups of genes encoding products involved in lipid metabolism, for instance, being more highly represented in expressed gene collections from pathogenic species. Our results also suggest that the genome of the obligate biotrophic fungus *Blumeria graminis* has significant differences to those of other phytopathogens, providing evidence that biotrophic specialization has involved distinct patterns of gene loss and gene conservation.

RESULTS

The COGEME phytopathogenic fungi and oomycete database contains 57 727 annotated unique cDNA sequences (unisequences) from 15 species of phytopathogen and three species of saprophytic fungi (Soanes *et al*., 2002). Non-redundant unisequence datasets in the database were created from raw ESTs using cluster assembly (Huang and Madan, 1999). Unisequence datasets do not translate directly into a set of unique genes because partial 5′ end sequences may be derived from different regions of the gene transcript and 3′ and 5′ end sequences from the same clone may not overlap. This means that one gene may be represented by more than one unisequence (Skinner *et al*., 2001). As well as having unisequence datasets from a large number of species of phytopathogen there are also draft genome sequences and sets of predicted proteins available for three species of phytopathogen [*Magnaporthe grisea* (Dean *et al*., 2005), *Gibberella zeae* (anamorph: *Fusarium graminearum*) and *Ustilago maydis*, two species of saprophytic ascomycete *Emericella nidulans* (anamorph: *Aspergillus nidulans*) and *Neurospora crassa* (Galagan *et al*., 2003)], the white-rot fungus *Phanerochaete chrysosporium* (Martinez *et al*., 2004) and the budding yeast *Saccharomyces cerevisiae* (Goffeau *et al*., 1996). This comparative wealth of putative genomic information has enabled us to compare unisequence EST datasets with gene inventories from a range of fungi—both ascomycetes and basidiomycetes, pathogens and saprophytes. For this study we consistently used datasets that contained at least 2000 unisequences to provide statistically valid results. The genomes of the filamentous fungi sequenced to date putatively contain between 6500 and 12 000 predicted open-reading frames based on current annotation and so the unisequence datasets used in this study are likely to represent greater than 10% of the total genes in the genome of any of the organisms from which they have been sequenced. The species used in this study are listed in Table 1.

Having a large number of unisequences for nine species of pathogen and two species of saprophyte enabled us to perform

Table 1 Unisequence datasets used in this study*.

Organism	No. of unisequences†	Tissues from which cDNA libraries were constructed	Reference
Blumeria graminis	3241	conidia, appressoria, epithytic mycelium	Thomas *et al*. (2001)
Botryotinia fuckeliana	2888	nitrogen-starved mycelium	
Cryphonectria parasitica	2182	mycelium, wild-type and hypovirus infected	Dawe *et al*. (2003)
Fusarium sporotrichioides	3432	Tri 10 over-expressed cDNA library	Peplow *et al*. (2003)
Gibberella zeae	4104	mature perithecia and mycelium grown under a variey of conditions (nitrogen starvation, carbon starvation, grown on wheat, grown in conditions to optimize trichothecene production, stressed, grown on minimal media, grown on complex plant substrate, grown on V8 plates)	Trail *et al*. (2003)
Magnaporthe grisea	8809	unstarved mycelium, carbon-starved mycelium, nitrogen-starved mycelium, mycelium grown in medium containing extract from rice cell walls, conidia and appressoria, as well as germinated conidia from pmk1-mutant and a mixed mated culture	Rauyaree *et al*. (2001); Ebbole *et al*. (2004)
Mycosphaerella graminicola	2906	mycelium grown in liquid culture (one with ammonium ions as a source of nitrogen and the other grown in wheat leaf extract) and one from fungal infected plant material	Keon *et al*. (2000)
Phytophthora sojae	5727	zoospores, mycelium, soybean hypoctyl infected with *P. sojae*	Qutob *et al*. (2000)
Ustilago maydis	4275	germinating teliospores and diploid mycelium	Nugent *et al*. (2004)
Emericella nidulans	4800	vegatative mycelium, cleistothecium and early sexual structures	
Neurospora crassa	5138	conidia, perithecia and mycelium grown in complete medium, under nitrogen starvation and mycelium harvested in the morning and evening	Zhu *et al*. (2001)

*The COGEME phytopathogen EST database also contains unisequence datasets for *Aspergillus niger*, *Cladosporium fulvum*, *Colletotrichum trifolii*, *Leptosphaeria maculans*, *Phytophthora infestans* (Kamoun *et al*., 1999), *Sclerotinia sclerotiorum* (Li *et al*., 2004) and *Verticillium dahliae* (Neumann & Dobinson, 2003).
†Does not include unisequences that represent ribosomal RNA or contaminating host-plant, bacterial and vector sequence.

Table 2 Number of unisequences from each unisequence set that have homology to predicted proteins from completed fungal genomes.*

Unisequence sets	Fungal genomes						
	G. zeae	*M. grisea*	*N. crassa*	*A. nidulans*	*U. maydis*	*S. cerevisiae*	*P. chrysosporium*
Blumeria graminis	1204 (37.15)	1165 (35.95)	1190 (36.72)	1125 (34.71)	718 (22.15)	653 (20.15)	737 (22.74)
Botryotinia fuckeliana	1645 (56.96)	1558 (53.95)	1613 (55.85)	1513 (52.39)	1029 (35.63)	965 (33.41)	1052 (36.43)
Cryphonectria parasitica	1463 (67.05)	1410 (64.62)	1403 (64.30)	1289 (59.07)	878 (40.24)	732 (33.55)	883 (40.47)
Fusarium sporotrichioides	2325 (67.74)	1506 (43.88)	1566 (45.63)	1380 (40.21)	832 (24.24)	738 (21.50)	848 (24.71)
Gibberella zeae	3279 (79.90)	2363 (57.58)	2400 (58.48)	2137 (52.07)	1429 (34.82)	1267 (30.87)	1454 (35.43)
Magnaporthe grisea	4427 (50.26)	6034 (68.50)	4366 (49.56)	3722 (42.25)	2265 (25.71)	2002 (22.73)	2319 (26.33)
Mycosphaerella graminicola	1713 (58.95)	1608 (55.33)	1657 (57.02)	1720 (59.19)	1167 (40.16)	1009 (34.72)	1171 (40.30)
Phytophthora sojae	1524 (26.61)	1416 (24.72)	1489 (26.00)	1504 (26.26)	1447 (25.27)	1291 (22.54)	1501 (26.21)
Ustilago maydis	1888 (44.16)	1780 (41.64)	1855 (43.39)	1881 (44.00)	3599 (84.19)	1470 (34.39)	1933 (45.22)
Emericella nidulans	2080 (43.33)	1910 (39.79)	1984 (41.33)	3461 (72.10)	1205 (25.10)	1073 (22.35)	1211 (25.23)
Neurospora crassa	2314 (45.04)	2197 (42.76)	3596 (69.99)	1978 (38.50)	1154 (22.46)	1015 (19.75)	1184 (23.04)

*Comparisons made using the BLASTX algorithm (Altschul *et al*., 1990). Homology is defined as an *e*-value of less than or equal to10^{-10}. The percentage of total unisequences is shown in parentheses.

a large-scale comparison between the unisequence datasets and the predicted proteins of seven fungal genomes (Table 2). The BLASTX algorithm (Altschul *et al*., 1990) was used to identify the number of unisequences for each organism that had at least one homologue in the set of predicted proteins for each of seven species of fungi. A stringent maximum *e*-value of 10^{-10} was used to define homology between unisequences and predicted proteins. Homology with an *e*-value of less than 10^{-10} is considered to be highly significant (Skinner *et al*., 2001). One of the principal aims of studies of pathogenic fungi is the identification and characterization of pathogenicity determinants, which we define here as genes necessary for a pathogen to infect and colonize a host plant (Oliver and Osbourn, 1995). Some of these factors may be unique to pathogens and we have used this as an initial, simple strategy to analyse the EST collection. Analysis of the results in Table 2 suggests that the unisequence dataset for each species of phytopathogenic fungus does not have a significantly higher number of sequences in common with the genomes of *Magnaporthe grisea* or *Gibberella zeae* than with genomes of the ascomycete non-pathogens *Neurospora crassa* and *Emericella nidulans* ($P < 0.01$) (Table 3). The only exception is that *M. grisea* does have a significantly higher number of sequences in common with *G. zeae* compared with the sequences of the two saprophytic ascomycetes. Some caution must be taken with interpretation of our analysis, however. *Magnaporthe grisea*, for example, has a similar number of sequences in common with the predicted gene set of *G. zeae* and *Neurospora crassa* (all three species are in the taxonomic class Sordariomycetes), but a lower number of sequences in common with the genome of the more distantly related *Emericella nidulans* (class Eurotiomycetes). The results also indicate that *Neurospora crassa* has a significantly higher number of sequences in common with *M. grisea* and *G. zeae* than the more distantly related saprophyte *E. nidulans*, which is simply consistent with established taxonomic relations between the species. Interestingly, the results also show that only 68.5% of *M. grisea* unisequences and 70% of *N. crassa* unisequences have homologues in the genomes of the same species. This is mainly due to the draft nature of the genomes used in which there are numerous gaps in the genomic sequence and also a large number of genes that have probably not been identified based on initial automated annotation.

Table 3 Average proportion of unisequences that have homologues in pathogenic and saprophytic ascomycetes.

Organism	Pathogens	Saprophytes	*P* value¶
Blumeria graminis	0.365474	0.357143	0.4851
Botryotinia fuckeliana	0.554536	0.541205	0.3087
Cryphonectria parasitica	0.658341	0.616865	0.0043
*Fusarium sporotrichioides**	0.438811	0.429196	0.4216
*Gibberella zeae**	0.57578	0.552753	0.0354
Magnaporthe grisea†	0.502554	0.459076	< 0.0001
Mycosphaerella graminicola	0.571404	0.581039	0.4573
Phytophthora sojae	0.256679	0.261306	0.5719
Ustilago maydis	0.429006	0.436959	0.458
Emericella nidulans‡	0.415625	0.413333	0.8197
Neurospora crassa§	0.438984	0.384975	< 0.0001

*Hits against genome of *Gibberella zeae* not included in analysis.
†Hits against genome of *Magnaporthe grisea* not included in analysis.
‡Hits against genome of *Emericella nidulans* not included in analysis.
§Hits against genome of *Neurospora crassa* not included in analysis.
¶*P*-value represents the probability that differences between the proportion of hits against pathogenic and saprophytic genomes is due to real genomic differences rather than random sampling error.

The results do, however, highlight a striking difference between the barley powdery mildew fungus *Blumeria graminis* and the other fungal pathogens, which may be of greater significance and

is not likely to be due to taxonomic relationship. Taxonomically *B. graminis* is a filamentous ascomycete (subphylum Pezizomycotina) in the class Leotiomycetes (the same class as *Botryotinia fuckeliana*). The other fungal pathogens in this study are all filamentous ascomycetes, apart from *Ustilago maydis* which is a basidiomycete. The unisequence dataset representing *B. graminis* has a significantly lower number ($P < 0.01$) of sequences in common with the predicted gene inventories of the four species of filamentous ascomycete fungi than any of the unisequence datasets of the other fungal pathogens investigated. Even though *B. graminis* is an ascomycete it has a significantly lower number ($P < 0.01$) of sequences in common with the four species of filamentous ascomycete with sequenced genomes than shown by the basidiomycete *U. maydis*. This difference may be due to the lifestyle of *B. graminis*, which is the only obligate biotrophic species among the pathogens represented in this study. It has long been speculated that the more intimate relationship between *B. gramins* and the host plant may require novel genes that are not required for the life cycle of necrotrophic or hemibiotrophic fungi. Another consistent observation is that all ascomycete fungi in the study have significantly more sequences in common with the genomes of the filamentous basidiomycetes *U. maydis* and *Phanerochaete chrysosporium* than the genome of the ascomycete yeast *Saccharomyces cerevisiae*. This observation suggests that genes specific to filamentous fungal species may have been lost from genomes of the free-living yeast subsequent to their divergence.

To investigate gene conservation we examined genes present in the genomes of filamentous fungi but not present in *S. cerevisiae*. Table 4 shows unisequences in each dataset that have homologues in the genomes of all five species of filamentous fungi studied, but not in the genome of *S. cerevisiae*. Genes conserved in the species of filamentous fungi represent between 2.5 and 5% of the total unisequences for each species of pathogen. The full list of unisequences from each species of pathogen that are conserved in the genomes of filamentous fungi can be found online at the COGEME website (http://cogeme.ex.ac.uk/filamentous.html). The gene lists revealed by this analysis suggest there are differences even in basic metabolism between *S. cerevisiae* and the filamentous fungi. For example, β-oxidation of fatty acids occurs solely in peroxisomes in *S. cerevisiae*, with the first step in this pathway being catalysed by a fatty-acyl CoA oxidase (Hiltunen *et al*., 2003). By contrast, it has been shown that the β-oxidation of fatty acids occurs in catalase-free microbodies in *N. crassa* and the first step in the pathway is catalysed by a fatty-acyl CoA dehydrogenase (Thieringer and Kunau, 1991). Consistent with this experimental analysis, our informatics results show that fatty-acyl CoA dehydrogenase genes are present in the genomes of all the pathogens studied and are conserved among filamentous fungal genomes but not present in the genome of *S. cerevisiae*. There are also a significant number of unisequences that are conserved in filamentous fungal genomes that represent genes encoding enzymes involved in the breakdown of plant cell-wall polysaccharides, such as acetyl xylan esterases, β-glucosidases, endo-β-1,4-D-xylanase and mixed-linked glucanases (de Vries and Visser, 2001). This is consistent with the fact that the filamentous fungal genomes under study are phytopathogenic or saprophytic species that derive nutrition from living or dead plant tissue, respectively.

Table 4 Number of unisequences that have homologues in the genomes of *M. grisea*, *G. zeae*, *U. maydis*, *E. nidulans* and *N. crassa* but no homologue in the genome of *S. cerevisiae*.*

Organism	No. of unisequences	% total unisequences
Blumeria graminis	84	2.59
Botryotinia fuckeliana	106	3.67
Cryphonectria parasitica	108	4.95
Fusarium sporotrichioides	96	2.80
Gibberella zeae	196	4.78
Magnaporthe grisea	277	3.14
Mycosphaerella graminicola	145	4.99
Phytophthora sojae	78	1.37
Ustilago maydis	202	4.73

*Homologues identified using BLASTX, *e*-value $\leq 10^{-10}$.

The vast majority of pathogenicity factors that have been identified so far have homologues in non-pathogenic fungi, although this is likely to be a reflection of the methods used to identify them (Gold *et al*., 2001). In many cases the sequences of well-conserved genes involved in signal transduction in yeast have been used to identify homologues in pathogenic species, for example, many of which play an important role in pathogenicity (Xu and Hamer, 1996). It has been suggested that there may be a set of pathogenicity factors that are present in the genomes of many species of pathogen but not in the genomes of non-pathogens (Yoder and Turgeon, 2001); however, no systematic study has set out to identify them. Conserved pathogen-specific genes have been identified by looking for unisequences that have homologues in the genomes of three species of pathogen but no homologue in the genomes of two species of saprophyte (Table 5). Due to the draft nature of the fungal genomes used, some the sets of predicted proteins may not be complete as a result of misannotation. Therefore, to check that these unisequences were indeed unique to pathogens, both sets of predicted proteins and genomic DNA were searched for homology. Surprisingly, only 19 pathogen-specific genes were identified in this way. These include genes that putatively encode a phosphatidylinositol 4-kinase, a 26S proteasome regulatory subunit, glycerophosphoryl diester phosphodiesterase, two putative ion-transporting ATPases, an alcohol dehydrogenase, DNA photolyase, a paraoxonase/arylesterase, an aminotransferase, a methyltransferase, as well as genes of unknown function. The full list of pathogen-specific unisequences can be obtained online from the COGEME website (http://cogeme.ex.ac.uk/cgi-bin/path.pl).

Table 5 Unisequences that have homology to predicted proteins from the genomes of *M. grisea*, *G. zeae* and *U. maydis* but no homology to predicted proteins from the genomes of *N. crassa* and *A. nidulans*.*

Unisequence ID	Organism	Function
Bg13901768	*Blumeria graminis*	phosphatidylinositol 4-kinase
BfCon[0449]	*Botryotinia fuckeliana*	alcohol dehydrogenase
Fs14662471	*Fusarium sporotrichioides*	Na^+/K^+-ATPase
Fs14666787	*Fusarium sporotrichioides*	paraoxonase/arylesterase
Fs14668177	*Fusarium sporotrichioides*	Na^+/K^+-ATPase alpha subunit
Gz31370824	*Gibberella zeae*	26S proteasome regulatory subunit 9
Mag24773065	*Magnaporthe grisea*	aminotransferase
Mag30399373	*Magnaporthe grisea*	H^+/K^+-ATPase alpha subunit
Mag30402756	*Magnaporthe grisea*	unknown
MagCon[1003]	*Magnaporthe grisea*	methyltransferase
MagCon[11360]	*Magnaporthe grisea*	unknown
MagCon[3955]	*Magnaporthe grisea*	unknown
MagCon[6133]	*Magnaporthe grisea*	unknown
mga0389f	*Mycosphaerella graminicola*	phosphatidylinositol 4-kinase
mgc09d06f	*Mycosphaerella graminicola*	unknown
Pi10230944	*Phytophthora infestans*	cryptochrome/DNA photolyase
UmCon[0629]	*Ustilago maydis*	unknown
UmCon[0817]	*Ustilago maydis*	26S proteasome regulatory subunit 9
UmCon[2356]	*Ustilago maydis*	glycerophosphoryl diester phosphodiesterase

*Homologues identified using BLASTX, *e*-value $\leq 10^{-10}$.

Table 6 Functional categories that are significantly more highly represented in the unisequence sets from ascomycete pathogens as compared with ascomycete non-pathogens ($P < 0.01$).

Functional category	Pathogens*	Non-pathogens*	*P*-value†
phosphate metabolism	0.20	0.09	0.0088
lipid metabolism	2.49	1.86	0.0001
protein synthesis	3.89	3.19	0.0009
cell rescue, defence, death and ageing	2.41	1.64	< 0.0001
cellular organization	1.29	0.70	< 0.0001
transposon, IS and plasmid	0.42	0.11	< 0.0001

*Percentage of total unisequences allocated to each functional category.
†Value representing the probability that the difference in the proportion of unisequences in each functional category is caused by random sampling error.

Functional categories

EST unisequences were allocated to functional categories based on homology with known proteins using a scheme adapted from that produced by MIPS (Munich Information Centre for Protein Sequences) (Mewes *et al*., 1997). The number of unisequences in each functional category was calculated for each organism (Fig. 1). A list of unisequences allocated to each functional classification group for each organism can be retrieved from the COGEME database (http://cogeme.ex.ac.uk/advsearch.html). The proportion of unisequences allocated to each functional category was compared between ascomycete pathogenic fungi and ascomycete non-pathogenic fungi to identify those functional groups of genes significantly ($P < 0.01$) more highly represented in pathogenic than non-pathogenic species (Table 6). Differences in the number of transposon-encoding unisequences can be attributed to the fact that transposable elements are common in the genomes of *M. grisea* and *B. graminis* but not in the genomes of the non-pathogens examined (Daboussi and Capy, 2003). Some of the other functional classification groups were found to be more highly represented in the pathogenic species of ascomycete when compared with the non-pathogenic species, and to contain genes that may potentially be important in pathogenesis. Examples include genes encoding products involved in basic lipid metabolism. Here, the lipid metabolism functional category group also contains genes encoding enzymes involved in synthesis of steroids and terpenes, such as the trichothecene toxin synthesized by *Fusarium sporotrichioides* and *Gibberella zeae* (Sweeney and Dobson, 1999). The cell rescue, defence, death and ageing functional category also contains genes involved in stress response, detoxification, and the degradation of exogenous proteins, esters and polysaccharides. Many of these products may play roles in pathogenicity, for example, being involved in the degradation of plant cell-wall components or defence against active oxygen species and toxins produced by the host plant (Idnurm and Howlett, 2001).

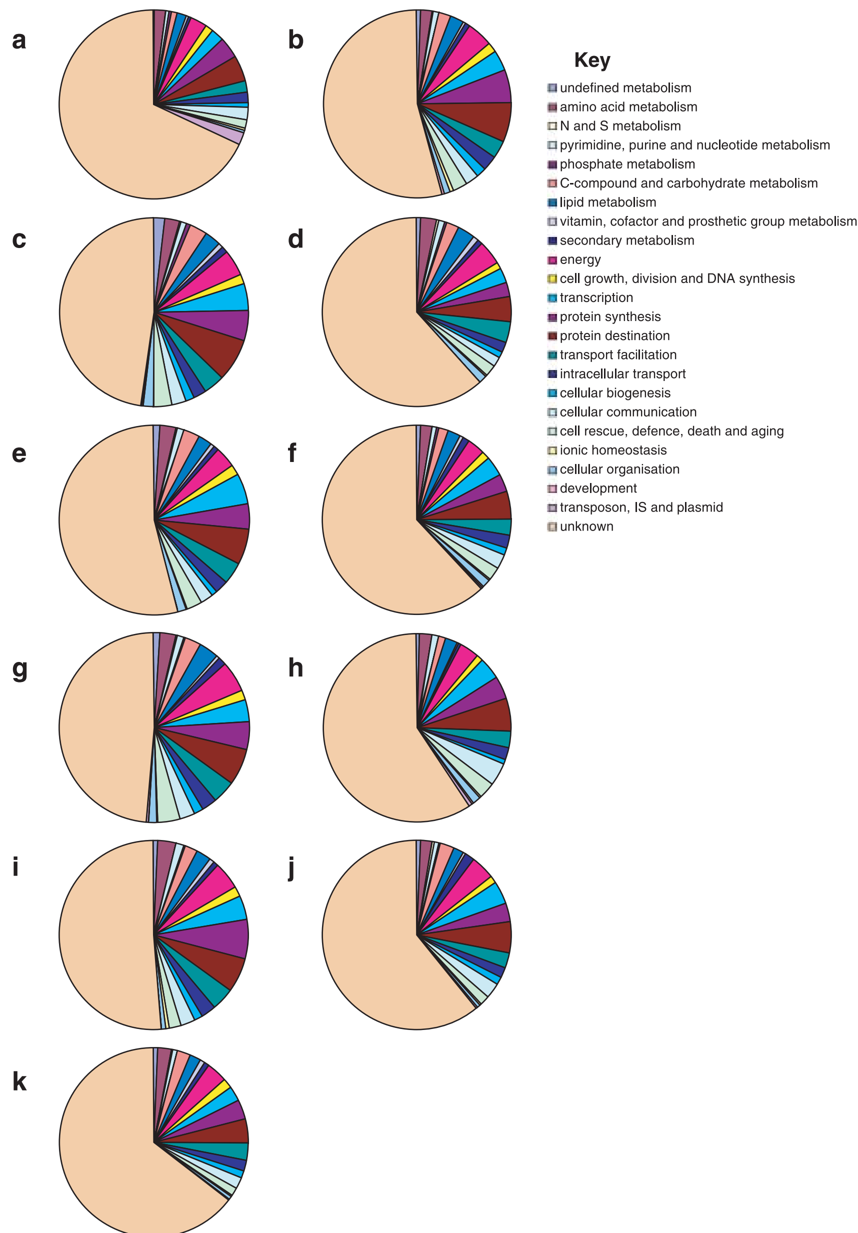

Fig. 1 A series of pie charts showing the proportion of unisequences from each species allocated to each functional category group. Species are *Blumeria graminis* (a), *Botryotinia fuckeliana* (b), *Cryphonectria parasitica* (c), *Fusarium sporotrichioides* (d), *Gibberella zeae* (e), *Magnaporthe grisea* (f), *Mycosphaerella graminicola* (g), *Phytophthora sojae* (h), *Ustilago maydis* (i), *Emericella nidulans* (j) and *Neurospora crassa* (k).

Another observation that emerged from broad-scale comparison was the proportion of unisequences that have an unknown function in each species (Table 7). These unisequences either show no homology to known sequences or only show similarity to sequences of unknown function. The obligate biotrophic species *B. graminis* has a significantly ($P < 0.01$) higher number of sequences of unknown function than other species of phytopathogenic fungus, providing further evidence of the dissimilarity of the obligate biotrophic lifestyle compared with that of the relatively closely related pathogenic species studied. The 2204 *B. graminis* unisequences labelled as unknown were studied in more detail. A total of 1622 (49.9% of all *B. graminis* unisequences) of these unisequences had no homology to sequences in the NCBI non-redundant database or to sequences in the genomes of *M. grisea*, *N. crassa*, *G. zeae*, *E. nidulans*, *S. cerevisiae* or *U. maydis*, suggesting that the genome of *B. graminis* contains a large number of completely unique genes. A group of 16 *B. graminis* genes were conserved in the genomes of ascomycete phytopathogens (with homologues to sequences in the genomes of both *M. grisea* and *G. zeae* but not the saprophytes *E. nidulans* or *N. crassa*). Moreover, 160 (7.3%) of the unknown *B. graminis* unisequences were predicted to be secreted from the cell, using the Pexfinder protocol to predict genes that encode secreted proteins using EST sequences (Torto *et al*., 2003). This is lower than

Table 7 Number of unisequences of unknown function for each species.

Organism	No. of unisequences	% total unisequences
Blumeria graminis	2204	68.0
Botryotinia fuckeliana	1573	54.5
Cryphonectria parasitica	1040	47.7
Fusarium sporotrichioides	2118	61.7
Gibberella zeae	2218	54.0
Magnaporthe grisea	5461	62.0
Mycosphaerella graminicola	1422	48.9
Phytophthora sojae	3401	59.4
Ustilago maydis	2191	51.3
Emericella nidulans	2916	60.8
Neurospora crassa	3316	64.5

the 12.2% of ESTs from the oomycete *Phytophthora infestans* that were predicted to be secreted using this protocol (Torto *et al*., 2003), but demonstrates the diversity of secreted products that appear to be encoded by an obligate pathogenic fungus. It would be interesting to study this group of unknown unisequences further and determine whether any particular protein domains are over-represented in the group.

DISCUSSION

Comparison of gene inventories, either from genomic sequence or from EST collections, enables us to address a fundamental question about what makes a pathogenic fungus different from a non-pathogenic species. In fact, there is very likely to be more than one answer to such a question because pathogenic fungi are present in such a wide range of taxonomic groups and are often closely related to particular non-pathogenic species. Pathogenicity to both plant and animal hosts is therefore likely to have evolved many times within the kingdom fungi. It can be speculated that there are three plausible, but not mutally exclusive, mechanisms that could account for the evolution of pathogenicity (Tunlid and Talbot, 2002). First, the genomes of pathogens may have acquired novel genes that enable them to infect and colonize plants. These genes may have been acquired by horizontal transfer, as demonstrated in bacteria, although there are very few possible examples where this process may have occurred among fungi or other eukaryotic taxa (Rosewich and Kistler, 2000). Alternatively, pathogenic fungi may have evolved by duplication of ancestral genes and subsequent divergence of function. Gene duplication is often postulated to be a primary function in the innovation of new genes (Ohno, 1970). If this is the case then certain gene families may contain more members in the genomes of pathogenic species compared with non-pathogenic species (Dean *et al*., 2005). Secondly, there may be genes in pathogens that are also present in non-pathogenic species but take on a different role in the former. In this case, the difference between the pathogen and non-pathogen would be likely to be at the level of regulation of a particular gene, or set of genes. There are many examples of this potential evolutionary route, mainly concerning genes involved in signal transduction such as those encoding MAP kinases (Xu and Hamer, 1996), adenylate cyclases (Choi and Dean, 1997), G-proteins (Liu and Dean, 1997) and cyclic AMP-dependent protein kinases (Mitchell and Dean, 1995). Thirdly, it may be that pathogenicity is associated with gene loss. For example, obligate intracellular parasites such as the bacterium *Mycobacterium leprae* have undergone significant gene loss during obligate association with a host (Sakharkar *et al*., 2004).

If pathogenicity was associated with a large number of unique genes shared between pathogenic species then it would be expected that the genomes of pathogens would have more sequences in common with each other than the genomes of non-pathogenic species. Our analysis, albeit on a limited number of species, shows that this is not likely to be the case. A search of 42 559 unisequences from 15 species of phytopathogen found only 19 sequences that had homologues in the genomes of three species of plant-pathogenic fungi, but no homologue in two species of saprophytic fungi. There may therefore be few universal pathogenicity factors (if any at all) identified in such a broad-scale analysis. The first main determinant of similarity between genomes is clearly the taxonomic relationship between the species sampled. For example, the genome of *M. grisea* was more similar to the genome of *G. zeae* and *N. crassa*, both from the class Sordariomycetes, than the genome of *E. nidulans*, class Eurotiomycetes. The second determinant of similarity between the genomes is the lifestyle and growth habit of the fungus. The genomes of the filamentous ascomycetes clearly share more sequences in common with the genomes of filamentous basidiomycetes than of the free-living ascomycete yeast *S. cerevisiae*. A search of 42 559 unisequences from 15 species of phytopathogen found 1802 unisequences that had homologues in the genomes of five species of filamentous fungi (one of which was a basidiomycete), but no homologues in the genome of *S. cerevisiae*. It is likely that these represent ancestral genes that have been lost from the genome of *S. cerevisiae* and other yeast lineages because they are not necessary for the propagation in the yeast form and the relatively restricted ecological niche occupied by such species. A large-scale comparison between EST sequences of *N. crassa* and the genomic sequence of *S. cerevisiae* identified a number of genes that seem to have been lost from the yeast genome (Braun *et al*., 2000), consistent with our observations here.

The barley powdery mildew fungus *Blumeria graminis* f. sp. *hordei* has a distinct life cycle among the pathogenic fungi we investigated. It is an obligate biotroph that requires the plant host for completion of its life cycle and cannot be cultured at all *in vitro*. Obligate biotrophs internalize within host cells and establish specialized infection structures called haustoria to redirect host metabolism to meet their own needs without causing the death of the plant

cell (Panstruga, 2003). By contrast, necrotrophic fungi trigger host cell death during the course of infection. The distinct lifestyle of *B. graminis* is reflected in its gene inventory based on EST analysis, having significantly fewer sequences in common with the genomes of filamentous ascomycetes than any of the other pathogenic fungi studied (including the basidiomycete *U. maydis*). It also has a significantly greater number of unisequences of unknown function than the other phytopathogens we investigated. Nearly half of the *B. graminis* unisequences are orphans (having no homologues in the NCBI database of sequenced proteins or in the sequenced fungal genomes). When considered together, our observations suggest that the biotrophic life cycle of *B. graminis* necessitates a large number of gene products not found in necrotrophic and hemibiotrophic phytopathogens, many of which are of unknown function. Of 60 *B. graminis* genes deposited in NCBI, five have no homologues present in the NCBI non-redundant database. These include three genes up-regulated during barley infection and two genes expressed in germinating conidia (Justesen *et al*., 1996). Obligate biotrophs occur in a variety of taxonomic groups and include oomycetes (e.g. *Peronospora parasitica*), basidiomycetes (e.g. *Puccinia* ssp.) as well as ascomycetes such as *B. graminis* (Panstruga, 2003). They all form haustoria and once sequence information becomes available for other obligate biotrophs it will be fascinating to determine if they have gene functions in common that are not shared with facultative phytopathogens. Our initial analysis suggests this might be the case.

Comparison of the distribution of unisequences among functional categories between pathogens and non-pathogens revealed a number of groups of genes that seem to be more highly represented in the genomes of pathogens. These include genes involved in lipid metabolism and cell rescue, defence, death and ageing. These groups include genes that have been implicated in pathogenicity, such as those involved in fungal toxin biosynthesis, detoxification and degradation of exogenous polysaccharides (Idnurm and Howlett, 2001). Gene family expansion may have occurred in some of these groups in pathogens as compared with non-pathogens, and this is worthy of further study.

The amount of fungal sequence data from both genomic sequencing and ESTs has been growing rapidly, enabling for the first time some meaningful comparison of the partial and (almost) complete gene sets of pathogens and non-pathogens to be made. We have shown that even broad-scale analysis can reveal that the extent of diversity is likely to be very great among pathogenic species and that obligate parasitism, so recalcitrant to experimental study, may be best explored by comparative genomics.

EXPERIMENTAL PROCEDURES

Raw EST sequences were downloaded from the dbEST database at NCBI (http://www.ncbi.nlm.nih.gov/dbEST/index.html) with the following exceptions.

ESTs from *B. graminis* were obtained by sequencing a library constructed from infected plant material by Dr Sarah Gurr (University of Oxford), from an EST dataset produced by the Carlsberg Institute, Denmark (Thomas *et al*., 2001), and sequences submitted by Dr Pietro Spanu and Maike Both (Imperial College, London). ESTs from *M. graminicola* were obtained by sequencing three cDNA libraries constructed at IACR-Long Ashton, UK, by Drs John Hargreaves and John Keon (Keon *et al*., 2000). The sequences for the end rot pathogen *Botryotinia fuckeliana* (anamorph: *Botrytis cinerea*) comprised ESTs downloaded from the website of the French sequencing centre, Genoscope. Sequencher (Gene Codes Corp.) was used to cluster ESTs into unisequences using a minimum 80% sequence identity over a 60-bp region. The BLASTX algorithm (Altschul *et al*., 1990) was used to compare unisequences to the NCBI non-redundant protein database. The top 20 hits with an *e*-value of less than 10^{-5} were recorded for each unisequence. These were used to assign a putative function to each unisequence. Based on this, each unisequence was placed in a functional classification group using a scheme adapted from one produced by the MIPS (Mewes *et al*., 1997). Unisequence data were entered into a web-accessible database (http://cogeme.ex.ac.uk) as described previously (Soanes *et al*., 2002).

For the purposes of this analysis unisequences representing ribosomal RNA, bacterial and host plant genes were removed from each organism's unisequence dataset. They were identified by the comparison of unisequences with rRNA-encoding genes from *Verticilium dahliae* (AF104926), the genomic sequence of *Escherichia coli* and EST datasets from host plants, respectively, using BLASTN (Altschul *et al*., 1990). Unisequences representing cloning vector genes were identified based on their top 20 hits against the NCBI non-redundant database and also removed from the unisequence datasets. The number of unisequences in each functional classification group was calculated for each organism. The proportion of total unisequences represented in each functional classification group was calculated by dividing the number of unisequences in each functional classification group by the total number of unisequences for that organism. When comparing the proportion of unisequences in a functional classification group between organisms it was assumed that the unisequence datasets were a random selection of the genome and that the number of unisequences in each functional classification group followed a normal distribution. The number of standard deviations between the observed proportions of unisequences in a functional classification group between two organisms was calculated using the following formula:

$$\frac{P_1 - P_2}{\sqrt{\frac{P_1 \times (1 - P_1)}{N_1} + \frac{P_2 \times (1 - P_2)}{N_2}}}$$

where P_1 and P_2 are the proportion of unisequences representing the functional category in question in organism 1 and 2, respectively,

and N_1 and N_2 are the total number of unisequences for organism 1 and 2, respectively. The *P* value was calculated from the normal distribution curve using the number of standard deviations between the proportions of unisequences.

Genome sequence assemblies and sets of predicted proteins for *Magnaporthe grisea*, *Gibberella zeae*, *Neurospora crassa* and *Emericella nidulans* were downloaded from the website of the Broad Institute, Cambridge, MA (http://www.broad.mit.edu/). A genome sequence assembly and set of predicted proteins for the white-rot fungus *Phanerochaete chrysosporium* was downloaded from the website of the John Doe Institute, University of California (http://genome.jgi-psf.org/whiterot1/whiterot1.download.html). A set of predicted proteins from the genome of the budding yeast *Saccharomyces cerevisiae* was downloaded from the website of the *Saccharomyces* genome project (http://www.yeastgenome.org/). Sequence comparisons were carried out using the blastall program version 2.2.5 (Altschul *et al*., 1990, 1997) downloaded from the NCBI website (ftp://ftp.ncbi.nih.gov/blast/).

ACKNOWLEDGEMENTS

This work was supported by the Biotechnology and Biological Sciences *e-fungi* project, which is part of the BBSRC e-science Initiative. We would like to thank Dr David Hoyle for help with statistics.

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. *Nucl. Acids Res.* **25**, 3389–3402.

Braun, E.L., Halpern, A.L., Nelson, M.A. and Natvig, D.O. (2000) Large-scale comparison of fungal sequence information: mechanisms of innovation in *Neurospora crassa* and gene loss in *Saccharomyces cerevisiae*. *Genome Res.* **10**, 416–430.

Choi, W. and Dean, R.A. (1997) The adenylate cyclase gene *MAC1* of *Magnaporthe grisea* controls appressorium formation and other aspects of growth and development. *Plant Cell*, **9**, 1973–1983.

Daboussi, M.J. and Capy, P. (2003) Transposable elements in filamentous fungi. *Annu. Rev. Microbiol.* **57**, 275–299.

Dawe, A.L., McMains, V.C., Panglao, M., Kasahara, S., Chen, B. and Nuss, D.L. (2003) An ordered collection of expressed sequences from *Cryphonectria parasitica* and evidence of genomic microsynteny with *Neurospora crassa* and *Magnaporthe grisea*. *Microbiology*, **149**, 2373–2384.

Dean, R.A., Talbot, N.J., Ebbole, D.J., Farman, M.L., Mitchell, T.K., 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., Rehmeyer, 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.

DeZwaan, T.M., Carroll, A.M., Valent, B. and Sweigard, J.A. (1999) *Magnaporthe grisea* pth11p is a novel plasma membrane protein that mediates appressorium differentiation in response to inductive substrate cues. *Plant Cell*, **11**, 2013–2030.

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., *et al.* (2004) Genome evolution in yeasts. *Nature*, **430**, 35–44.

Ebbole, D.J., Jin, Y., Thon, M., Pan, H., Bhattarai, E., Thomas, T. and Dean, R. (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.

Galagan, J.E., Calvo, S.E., Borkovich, K.A., Selker, E.U., Read, N.D., Jaffe, D., *et al.* (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature*, **422**, 859–868.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., *et al.* (1996) Life with 6000 genes. *Science*, **274**, 563–567.

Gold, S.E., Garcia-Pedrajas, M.D. and Martinez-Espinoza, A.D. (2001) New (and used) approaches to the study of fungal pathogenicity. *Annu. Rev. Phytopathol.* **39**, 337–365.

Hiltunen, J.K., Mursula, A.M., Rottensteiner, H., Wierenga, R.K., Kastaniotis, A.J. and Gurvitz, A. (2003) The biochemistry of peroxisomal beta-oxidation in the yeast *Saccharomyces cerevisiae*. *FEMS Microbiol. Rev.* **27**, 35–64.

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

Idnurm, A. and Howlett, B.J. (2001) Pathogenicity genes of phytopathogenic fungi. *Mol. Plant Pathol.* **2**, 241–255.

Justesen, A., Somerville, S., Christiansen, S. and Giese, H. (1996) Isolation and characterization of two novel genes expressed in germinating conidia of the obligate biotroph *Erysiphe graminis* f. sp. *hordei*. *Gene*, **170**, 131–135.

Kamoun, S., Hraber, P., Sobral, B., Nuss, D. and Govers, F. (1999) Initial assessment of gene diversity for the oomycete pathogen *Phytophthora infestans* based on expressed sequences. *Fungal Genet. Biol.* **28**, 94–106.

Keon, J., Bailey, A. and Hargreaves, J. (2000) A group of expressed cDNA sequences from the wheat fungal leaf blotch pathogen, *Mycosphaerella graminicola* (*Septoria tritici*). *Fungal Genet. Biol.* **29**, 118–133.

Li, R., Rimmer, R., Buchwaldt, L., Sharpe, A.G., Seguin-Swartz, G., Coutu, C. and Hegedus, D.D. (2004) Interaction of *Sclerotinia sclerotiorum* with a resistant *Brassica napus* cultivar: expressed sequence tag analysis identifies genes associated with fungal pathogenesis. *Fungal Genet. Biol.* **41**, 735–753.

Liu, S. and Dean, R.A. (1997) G protein α subunit genes control growth, development, and pathogenicity of *Magnaporthe grisea*. *Mol. Plant–Microbe Interact.* **10**, 1075–1086.

Loftus, B.J., Fung, E., Roncaglia, P., Rowley, D., Amedeo, P., Bruno, D., *et al.* (2005) The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. *Science*, **307**, 1321–1324.

Martinez, D., Larrondo, L.F., Putnam, N., Gelpke, M.D., Huang, K., Chapman, J., *et al.* (2004) Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. *Nat. Biotechnol.* **22**, 695–700.

Mewes, H.W., Albermann, K., Heumann, K., Liebl, S. and Pfeiffer, F. (1997) MIPS: a database for protein sequences, homology data and yeast genome information. *Nucl. Acids Res.* **25**, 28–30.

Mitchell, T.K. and Dean, R.A. (1995) The cAMP-dependent protein kinase catalytic subunit is required for appressorium formation and pathogenesis by the rice blast fungus *Magnaporthe grisea*. *Plant Cell*, **7**, 1869–1878.

Neumann, M.J. and Dobinson, K.F. (2003) Sequence tag analysis of gene expression during pathogenic growth and microsclerotia development in the vascular wilt pathogen *Verticillium dahliae*. *Fungal Genet. Biol.* **38**, 54–62.

Nugent, K.G., Choffe, K. and Saville, B.J. (2004) Gene expression during *Ustilago maydis* diploid filamentous growth: EST library creation and analyses. *Fungal Genet. Biol.* **41**, 349–360.

Ohno, S. (1970) *Evolution by Gene Duplication.* New York: Springer.

Oliver, R.P. and Osbourn, A.E. (1995) Molecular dissection of phytopathogenicity. *Microbiology*, **141**, 1–9.

Panstruga, R. (2003) Establishing compatibility between plants and obligate biotrophic pathogens. *Curr. Opin. Plant Biol.* **6**, 320–326.

Peplow, A.W., Tag, A.G., Garifullina, G.F. and Beremand, M.N. (2003) Identification of new genes positively regulated by tri10 and a regulatory network for trichothecene mycotoxin production. *Appl. Environ. Microbiol.* **69**, 2731–2736.

Qutob, D., Hraber, P.T., Sobral, B.W. and Gijzen, M. (2000) Comparative analysis of expressed sequences in *Phytophthora sojae*. *Plant Physiol.* **123**, 243–254.

Rauyaree, P., Choi, W., Fang, E., Blackmon, B. and Dean, R.A. (2001) Genes expressed during early stages of rice infection with the rice blast fungus *Magnaporthe grisea*. *Mol Plant. Pathol.* **2**, 347–354.

Rosewich, U.L. and Kistler, H.C. (2000) Role of horizontal gene transfer in the evolution of fungi. *Annu. Rev. Phytopathol.* **38**, 325–363.

Sakharkar, K.R., Dhar, P.K. and Chow, V.T. (2004) Genome reduction in prokaryotic obligatory intracellular parasites of humans: a comparative analysis. *Int. J. Syst. Evol. Microbiol.* **54**, 1937–1941.

Skinner, W., Keon, J. and Hargreaves, J. (2001) Gene information for fungal plant pathogens from expressed sequences. *Curr. Opin. Microbiol.* **4**, 381–386.

Soanes, D.M., Skinner, W., Keon, J., Hargreaves, J. and Talbot, N.J. (2002) Genomics of phytopathogenic fungi and the development of bioinformatic resources. *Mol. Plant–Microbe Interact.* **15**, 421–427.

Sweeney, M.J. and Dobson, A.D. (1999) Molecular biology of mycotoxin biosynthesis. *FEMS Microbiol. Lett.* **175**, 149–163.

Thieringer, R. and Kunau, W.H. (1991) The beta-oxidation system in catalase-free microbodies of the filamentous fungus *Neurospora crassa*. Purification of a multifunctional protein possessing 2-enoyl-CoA hydratase, L-3-hydroxyacyl-CoA dehydrogenase, and 3-hydroxyacyl-CoA epimerase activities. *J. Biol. Chem.* **266**, 13110–13117.

Thomas, S.W., Rasmussen, S.W., Glaring, M.A., Rouster, J.A., Christiansen, S.K. and Oliver, R.P. (2001) Gene identification in the obligate fungal pathogen *Blumeria graminis* by expressed sequence tag analysis. *Fungal Genet. Biol.* **33**, 195–211.

Torto, T.A., Li, S., Styer, A., Huitema, E., Testa, A., Gow, N.A., van West, P. and Kamoun, S. (2003) EST mining and functional expression assays identify extracellular effector proteins from the plant pathogen *Phytophthora*. *Genome Res.* **13**, 1675–1685.

Trail, F., Xu, J.R., San Miguel, P., Halgren, R.G. and Kistler, H.C. (2003) Analysis of expressed sequence tags from *Gibberella zeae* (anamorph *Fusarium graminearum*). *Fungal Genet. Biol.* **38**, 187–197.

Tunlid, A. and Talbot, N.J. (2002) Genomics of parasitic and symbiotic fungi. *Curr. Opin. Microbiol.* **5**, 513–519.

de Vries, R.P. and Visser, J. (2001) *Aspergillus* enzymes involved in degradation of plant cell wall polysaccharides. *Microbiol. Mol. Biol. Rev.* **65**, 497–522.

Wood, V., Gwilliam, R., Rajandream, M.A., Lyne, M., Lyne, R., Stewart, A., *et al.* (2002) The genome sequence of *Schizosaccharomyces pombe*. *Nature*, **415**, 871–880.

Xu, J.R. (2000) Map kinases in fungal pathogens. *Fungal Genet. Biol.* **31**, 137–152.

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.

Yoder, O.C. and Turgeon, B.G. (2001) Fungal genomics and pathogenicity. *Curr. Opin. Plant Biol.* **4**, 315–321.

Zhu, H., Nowrousian, M., Kupfer, D., Colot, H.V., Berrocal-Tito, G., Lai, H., *et al.* (2001) Analysis of expressed sequence tags from two starvation, time-of-day-specific libraries of *Neurospora crassa* reveals novel clock-controlled genes. *Genetics*, **157**, 1057–1065.