Molecular Ecology (2005) doi: 10.1111/j.1365-294X.2005.02796.x

Screening for rapidly evolving genes in the ectomycorrhizal fungus *Paxillus involutus* using cDNA microarrays

ANTOINE LE QUÉRÉ,*‡¶ KASPER ASTRUP ERIKSEN,†¶ BALAJI RAJASHEKAR,* ANDRES SCHÜTZENDÜBEL,*§ BJÖRN CANBÄCK,* TOMAS JOHANSSON* and ANDERS TUNLID*
**Department of Microbial Ecology, Lund University, Ecology Building, SE-223 62 Lund, Sweden, †Complex System Division, Department of Theoretical Physics, Lund University, Sölvegatan 14A, SE-223 62 Lund, Sweden*

Abstract

We have examined the variations in gene content and sequence divergence that could be associated with symbiotic adaptations in the ectomycorrhizal fungus *Paxillus involutus* and the closely related species *Paxillus filamentosus*. Strains with various abilities to form mycorrhizae were analysed by comparative genomic hybridizations using a cDNA microarray containing 1076 putative unique genes of *P. involutus*. To screen for genes diverging at an enhanced and presumably non-neutral rate, we implemented a simple rate test using information from both the variations in hybridizations signal and data on sequence divergence of the arrayed genes relative to the genome of *Coprinus cinereus*. *C. cinereus* is a free-living saprophyte and is the closest evolutionary relative to *P. involutus* that has been fully sequenced. Approximately 17% of the genes investigated were detected as rapidly diverging within *Paxillus*. Furthermore, 6% of the genes varied in copy numbers between the analysed strains. Genome rearrangements associated with this variation including duplications and deletions may also play a role in adaptive evolution. The cohort of divergent and duplicated genes showed an over-representation of either orphans, genes whose products are located at membranes, or genes encoding for components of stress/defence reactions. Some of the identified genomic changes may be associated with the variation in host specificity of ectomycorrhizal fungi. The proposed procedure could be generally applicable to screen for rapidly evolving genes in closely related strains or species where at least one has been sequenced or characterized by expressed sequence tag analysis.

Keywords: cDNA microarray, comparative genomic hybridization, ECM, gene duplications, *Paxillus involutus*

Received 15 June 2005; revision received 13 September 2005; accepted 4 October 2005

Introduction

Comparative analysis of genome sequence data is an important tool to reveal genomic variations that may be related to phenotypic adaptations to specific environments. By comparing sequences encoding alternative phenotypes, it is possible to reconstruct the genomic pattern of change associated with the shift in phenotype. The observed pattern is then contrasted with what is expected in the absence of natural selection; that is with the expectation of the neutral theory of molecular evolution (Hughes 1999). A fundamental concept of this theory is the molecular clock, which predicts that as long as a protein's function remains unaltered, the protein's rate of evolution is approximately constant within different phylogenetic lineages (Kimura & Ota 1974). From this principle, it follows that a detection of change in the rate of evolution of a protein may reveal functional changes associated with adaptive changes in phenotypes.

Another principle governing molecular evolution is that gene duplication followed by functional diversification is the most important mechanism generating new genes and

Correspondence: Anders Tunlid, Fax: +46-46-222 4158; E-mail: anders.tunlid@mbioekol.lu.se.
‡Present address: Université de Genève, Département de Biologie Végétale, LBMPS — Science III, 30 Quai Ernest-Ansermet, CH-1211 Geneve 4, Switzerland.
§Present address: Institute of Forest Botany, Department of Forest Botany and Tree Physiology, Georg-August-Universität, Büsgenweg 2, DE-37077 Göttingen, Germany.
¶These authors contributed equally to this work.

Table 1 Fungal strains used in comparative genomic hybridizations (CGH)

Strain	Site and mycorrhizal host	Origin	Hybridizations*	References
P. involutus ATCC 200175 (reference strain†)	Isolated close to birch trees. Forms ECM with birch, pine, spruce and poplar (in the laboratory)	Scotland	16	Chalot *et al*. (1996)
P. involutus Pi01SE	Isolated from a pine forest	Sweden	2	S. Erland (unpublished)
P. involutus Pi08BE	Forms ECM with pine, spruce and poplar (laboratory)	Belgium	2	Blaudez *et al*. (1998)
P. involutus Maj	Isolated close to poplar trees. Forms ECM with poplar and birch (laboratory)	France	5	Gafur *et al*. (2004)
P. involutus Nau	Isolated close to oak trees. Does not form ECM with poplar and birch, but with oak (laboratory)	France	5	Gafur *et al*. (2004)
P. filamentosus Pf01De	Isolated close to alder trees	Germany	2	Jarosch & Bresinsky (1999)

*cf. Fig. 1.
†The microarray was constructed using PCR-amplified cDNA derived from this strain.

new biochemical functions (Ohno 1970; Hughes 1999). This prediction has been confirmed by recent analyses of genome sequence data. Thus many genes are members of large gene families and duplicated genes arise at very high rates. Following duplications, new genes usually evolve with rapid changes in their sequences and structures. However, the vast majority of gene duplicates are silenced within millions of years (Lynch & Conery 2000; Long *et al*. 2003). Furthermore, gene duplications and deletions are thought to play a major role in adaptations to various growth conditions including resource-limited environments (Dunham *et al*. 2002), pathogenesis and symbiosis (Ochman & Moran 2001). Accordingly, identification of divergent and duplicated genes is of major interest when studying genome evolution.

Comparisons of closely related strains or species are particularly informative for identifying adaptive evolution because they hold constant all variables shared by congeners (Harvey & Pagel 1991). However, complete genome sequence data are rarely available for closely related eukaryotes. As an alternative, microarray-based comparative genomic hybridization (array-CGH) can be used as a method for screening the presence of conserved and divergent genes (Dunham *et al*. 2002; Porwollik *et al*. 2002; Hinchliffe *et al*. 2003; Edwards-Ingram *et al*. 2004). In addition, array-CGH can be used to assess gene duplication and deletions at single-gene resolution in closely related species or strains of organisms (Hughes *et al*. 2000; Dunham *et al*. 2002; Pollack *et al*. 2002).

In this study, we have compared the gene content and patterns of large-scale genome variations in strains of the ectomycorrhizae (ECM) fungus *Paxillus involutus* using array-CGH. ECM are formed by mutualistic interactions between fungi and the roots of woody plants. The fungal partner obtains photosynthetic sugars from the host plant while in return the plant receives mineral nutrients from the fungus (Smith & Read 1997). Phylogenetic analysis has shown that the ancestors of the ECM homobasidiomycetes were free-living saprophytes and that mycorrhizal symbionts have evolved repeatedly from saprophytic precursors (Hibbett *et al*. 2000). *P. involutus* belongs to the suborder Boletineae of the homobasidiomycetes, which is one of the clades of ECM fungi identified by Hibbett *et al*. (2000). *P. involutus* is widely distributed over the Northern Hemisphere. The species has a wide host spectrum, and forms ECM with a large number of coniferous and deciduous trees (Table 1).

The cDNA microarray used in this study contained cDNA reporters representing 1076 putative unique genes in *P. involutus* and were derived from a collection of expressed sequence tag (EST) clones (Johansson *et al*. 2004). The array has previously been used for examining divergence in gene expression associated with variation in host specificity in strains of *P. involutus* (Le Quéré *et al*. 2004). To screen for genes diverging at an enhanced and presumably non-neutral rate, we implemented a simple rate test using information from both the variations in hybridizations signal and sequence divergence to genes in the genome of the homobasidiomycete *Coprinus cinereus*. *C. cinereus* is a free-living saprophyte and is the closest evolutionary relative to *P. involutus* that has been fully sequenced. *C. cinereus* is a member of the suborder Agaricinae,

which has been identified as a sister group to the boletoid lineage of homobasidiomycetes (Hibbett *et al*. 1997). We then asked whether the genes evolving at an enhanced rate typically encode proteins belonging to certain functional classes. Furthermore, we were able to unambiguously detect genes that vary in copy number within different lineages of *P. involutus*. Interestingly, the host specificity differs between strains within these lineages. Accordingly, the identified genomic changes may be associated with the variation in host specificity of ECM fungi.

Materials and methods

Fungal strains, growth conditions and DNA preparations

Five *Paxillus involutus* strains and one *Paxillus filamentosus* strain (Table 1) were grown on cellophane-covered agar plates (Brun *et al*. 1995). After 7–10 days of incubation at room temperature in the dark, the mycelium was transferred to the surface of Gamborg B-5 basal liquid medium (Sigma-Aldrich Sweden AB) (pH 5.0) supplemented with glucose (2.5 g/L) and incubated for 7–14 days. *Paxillus* DNA was prepared as described previously (Le Quéré *et al*. 2002), except that the ultracentrifugation steps were omitted. The DNA was treated with RNase A (Promega) and sonicated to generate fragments ranging between 200 and 2000 bp in size. The DNA samples were purified by phenol–chloroform extraction and with the QIAquick PCR purification kit (QIAGEN).

Microarrays and genomic hybridizations

In this study, two different batches of cDNA microarrays (Prints 1 and 2) were used. Both arrays were printed with reporters obtained from a nonredundant set of EST clones, either originating from the *P. involutus* ATCC 200175 strain (henceforth abbreviated ATCC) or from birch (*Betula pendula*) (Johansson *et al*. 2004). Each reporter was replicated in at least quadruplicates on the array. A full description of the Prints 1 and 2 array designs are available from the EMBL-EBI ArrayExpress database (www.ebi.ac.uk/arrayexpress) (Accession nos A-MEXD-184 and A-MEXP-92, respectively). From the entire set of available reporters, the plant EST reporters, 9 EST reporters without any putative origin, and 4 out of 39 reporters corresponding to fungal genomic fragments (polymerase chain reaction (PCR) products from various parts of a 33-kb genomic region of *P. involutus* contained within a cosmid) (Le Quéré *et al*. 2002) (Table S1, Supplementary material) were excluded from this investigation. This provided a uniset of 1120 reporters including 1076 EST-derived reporters, 35 cosmid-derived reporters, 1 blank and 8 heterologous and commercial control reporters [ArrayControl, Ambion (Europe) Ltd]. DNA corresponding to the heterologous control reporters were spiked in known amounts into the hybridization extracts prior to the labelling process (Table S2, Supplementary material).

The DNA samples (hybridization extracts) were labelled with either Cy3 or Cy5 (CyScribe Post-Labelling Kit, Amersham BioSciences) and purified using the QIAquick PCR Purification Kit (QIAGEN). The samples were eluted in 50 μL nuclease-free water and 20 μg of poly(dA)$_{80}$-poly(dT)$_{80}$ was added. Before hybridization, the extracts were evaporated and resuspended in 7.5 μL nuclease-free water. They were then heated to 95 °C for 2 min and incubated at 75 °C for 45 min. Finally, one volume (i.e. 7.5 μL) of microarray hybridization buffer (CyScribe Post-Labelling Kit, Amersham Biosciences) and two volumes (i.e. 15 μL) of formamide were added, mixed and briefly centrifuged before being used for hybridization against the cDNA arrays. Prehybridization of the microarray slides was performed in 50% formamide, 5 × SSC (1 × SSC is 0.15 M NaCl and 0.015 M sodium acetate) and 0.1% SDS at 42 °C for 45 min. The slides were then washed with distilled water, then with isopropanol, and finally dried by centrifugation. The slides were hybridized at 42 °C overnight using a CMT hybridization chamber (Corning Glass). They were then washed twice with 2 × SSC and 0.1% SDS (42 °C), once with 0.1 × SSC and 0.1% SDS (20 °C), three times with 0.1 × SSC (20 °C), and finally with 0.01 × SSC (20 °C). After drying by centrifugation, the slides were placed in a dry and dark chamber until scanning. Altogether, 16 microarray slides and 32 hybridization extracts were used in this study (Fig. 1). Fluorescence intensities were measured using an Axon 4000A laser scanner and converted into digital values using GENEPIX PRO software (3.0.6.89) (Axon Laboratories). Data images were inspected manually and low-quality spots were excluded from further analysis.

Analysis of hybridization intensities by clustering

The mean background fluorescence was calculated for each slide. After local background correction for each spot, the reporters yielding intensities below twice the background were excluded and the fluorescence of the remaining reporters was multiplied by a correction factor to give a common channel mean of 5000 fluorescence units for each slide. The discarded reporters which had yielded intensities below twice the background were then reintroduced, applying the calculated correction factor. After the normalization step, the mean hybridization intensity was calculated for each fungal strain, for each unique reporter and for each of the two batches of arrays (Prints 1 and 2). The $\log_2$-transformed values were then centred by subtracting a fixed value of 12.29, corresponding to the $\log_2$ of 5000 (fixed intensity used to normalize the data). We then calculated the ratio ($\log_2$) of hybridization

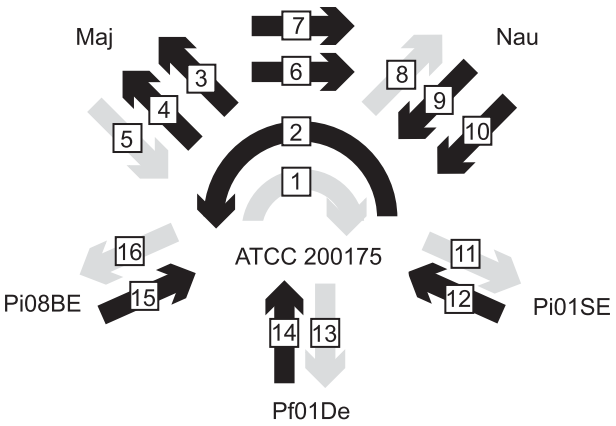

Fig. 1 Experimental design for the CGH analysis of five strains of *Paxillus involutus* (ATCC, Pi01SE, Pi08BE, Maj and Nau) and one strain of *Paxillus filamentosus* (Pf01De). Each arrow indicates paired dual-label microarray hybridizations of reference (ATCC) and test DNA, except for hybridizations 1 and 2 representing self-self hybridizations of the ATCC strain, and hybridizations 6 and 7 which represent direct comparisons of the strains Maj and Nau. The heads and tails mark the DNA sample labelled with Cy5 and Cy3, respectively. Black arrows indicate hybridizations using microarrays from Print 1 and grey arrows indicate hybridizations using Print 2 (cf. Fig. 2). The hybridization identities (1–16) are used for the organization of the data in the EBI-EMBL ArrayExpress database (www.ebi.ac.uk/arrayexpress; Accession no. E-MEXP-437)

intensities between the test and the reference strains for all the experiments performed. The $\log_2$ ratios of hybridization intensities for all replicated spots within an experiment were calculated and averaged. Data from replicated hybridizations using the same batch of array were averaged. The final data sets were entered into the program CLUSTER 3.0 (version 1.22) (http://bonsai.ims.utokyo.ac.jp/~mdehoon/software/cluster). The ratios ($\log_2$) of the hybridization intensities were clustered into 12 groups using the k-means method. The results (cf. Fig. 2) were displayed using Java TREEVIEW (version 1.0.3) (http://sourceforge.net/projects/jtreeview).

Analysis of hybridization intensities by mixed-model ANOVA

The $\log_2$-transformed hybridization intensities (h_{gad}) for the 1120 reporters were subjected to a normalization model of the form $h_{gad} = \mu + A_a + D_d + (A \times D)_{ad} + r_{gad}$, where μ is the sample mean, A_a is the effect of the ath array (a = 1–16), D_d is the effect of the dth dye (Cy3 or Cy5) $(A \times D)_{ad}$ is the array–dye interaction (channel effect), and r_{gad} is the residual. Subsequently the residuals were fitted by gene-specific models of the form: $r_{adbps} = S_s + \mu + A_a + D_d + B_b + P_p + (B \times P)_{bp} + \varepsilon_{adbps}$, where S_s is the sth strain (ATCC, Pi08BE, Pi01SE, Maj, Nau, and Pf01De), B_b is the bth batch of prints (Print 1 or 2), P_p is the pth pin used to print the reporter on the array (typically, two different pins were used to print quadruplicated reporters) and $(B \times P)_{bp}$ is the interaction between pin and batch. In the gene models, which were fitted using PROC MIXED in SAS/STAT software version 8 (SAS Institute), the A, D, P, B, $B \times P$ effects are random. The output contained an estimate of the $\log_2$ hybridization signal (S_s) and an estimate of the associated standard error. We arrived at this mixed model by assuming that the estimates of the $\log_2$ fold change of the very closely related strains Nau and Maj ($S_{\mathrm{Nau}} - S_{\mathrm{Maj}}$) is close to zero. Most reporters had a hybridization value for the ATCC strain, S_{ATCC}, close to zero (data not shown). However, there is a second group of reporters with S_{ATCC} below –3 (i.e. showing eight times lower intensity than an average reporter), probably due to poor print quality. Consequently, we disregarded all reporters with S_{ATCC} below –3, leaving 1052 reporters (1009 EST-derived, 35 cosmid-derived, and 8 heterologous control reporters). All the remaining 1052 reporters yielded an estimate of the $\log_2$ hybridization signal (S_s) for all strains.

Analysis of divergent genes using the EPLP procedure

The closest homologues for all arrayed *Paxillus* reporters (genes) were identified in the genome of *Coprinus cinereus* (www.broad.mit.edu) using the TBLASTX search tool (Altschul *et al*. 1990). For any given reporter, we retrieved a cohort of reporters with a similar degree of conservation. The cohort contained 50 reporters having the closest, but lower TBLASTX bit score and 50 reporters with the closest but higher bit score values. Subsequently, a Gaussian curve was fitted to the main peak of the distribution of the $\log_2$ fold changes of the hybridization signals of the conserved genes. The Gaussian fit was chosen such that it has the same height as the $\log_2$ fold change distribution and such that it equals the $\log_2$ fold change distribution. The estimated probability of local presence (EPLP) was defined as the ratio of the fitted Gaussian curve to the observed $\log_2$ fold change distribution. The $\log_2$ fold change value used as a cut-off to discriminate between conserved and divergent genes was identified as the fold change value closest to the mean of the Gaussian curve where the EPLP value was 0.05 (cf. Fig. 6).

Functional classifications of genes

For the 1076 putative genes represented on the array, a homology search was carried out using the TBLASTX algorithm (Altschul *et al*. 1990) with a threshold value of 14 for extending hits and an *E*-value threshold of 1e-10 against the UniProt sequence database (Apweiler *et al*. 2004). Gene Ontology (GO) (Ashburner *et al*. 2000) and InterPro

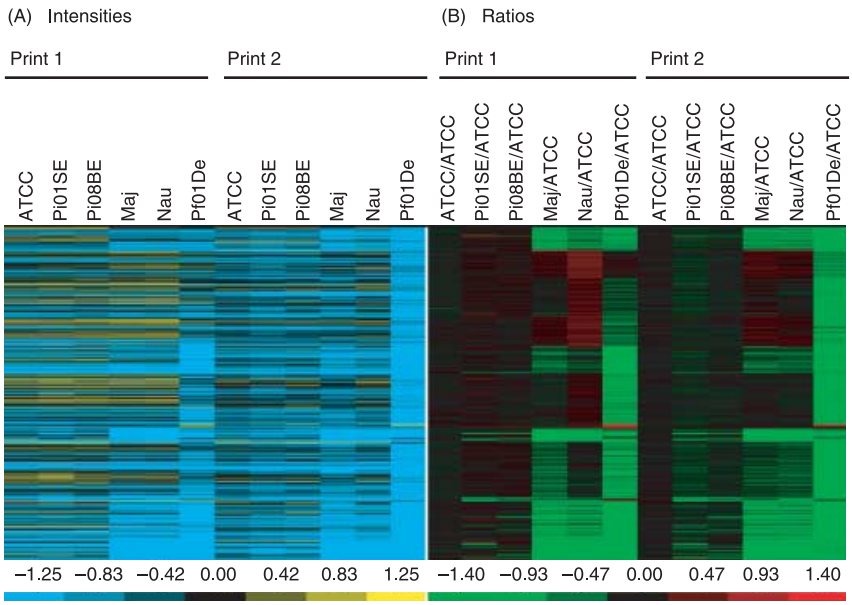

Fig. 2 CGH analyses of five strains of *Paxillus involutus* (ATCC, Pi01SE, Pi08BE, Maj and Nau) and one strain of *Paxillus filamentosus* (Pf01De). The array was printed using PCR-amplified cDNA derived from a collection of EST clones obtained from the ATCC strain (reference) and contained 1076 unique fungal reporters. By filtering out reporters that did not give hybridization signals for all the strains analysed, a set of 1009 EST-derived reporters was retained. (A) Normalized $\log_2$ transformed and centred hybridization intensities from two different batches of microarrays (Prints 1 and 2), where the scale from blue to yellow indicates low and high signal intensity, respectively. The two batches differ slightly in the arrangement of reporters and in the fabrication process (cf. Materials and methods section). (B) Ratios ($\log_2$) of hybridization intensities between the different strains and the reference ATCC strain where the scale from green to red represent variable or duplicated genes, respectively. The ratio between ATCC and ATCC (identical DNA preparation) reflects the technical variation in hybridizations. Note that although the hybridization intensity varies between the batches of arrays the ratios of hybridization intensities are similar for the two prints. The order of genes in panels (A) and (B) is identical and is based *k*-means clustering of the $\log_2$ ratios of the hybridization intensities.

annotations (Mulder *et al*. 2005) were inferred by retrieving information from the UniProt entry corresponding to the best TBLASTX hit. Using the full GO (www.geneontology.org), all the classified genes were mapped to all their parent terms in the yeast GO Slim (www.geneontology.org/GO.slims). The procedure for assigning *P* values to all the links in the GO Slim ontology in Fig. 9 is described in Supplementary material (Table S5). The sizes of the identified InterPro families in the basidiomycete *Phanerochaete chrysosporium* was retrieved from Martinez *et al*. (2004). A Wilcoxon rank test (Wilcoxon 1945) was used for analysing whether the divergent genes of *P. involutus* were over-represented among large gene families of *Phanerochaete chrysosporium* (Table S6, Supplementary material).

DNA sequencing

A selection of 17 EST clones were completely sequenced in both directions by using a pTriplEx2-specific universal forward primer P104 (5′-GGGAAGCGCGCCATTGTGTT-3′), a reverse primer T23V (5′-T_{23}V-3′, V = A, G or C), and template-specific primers. Based on the cDNA sequence information, primers were designed to amplify parts of the corresponding genomic regions in the various strains of *P. involutus* and *P. filamentosus* by PCR (Table 2). DNA sequencing was performed using a BigDye Terminator Cycle Sequencing Kit (Applied Biosystems) and a 3100 Genetic Analyser Sequencer (Applied Biosystems). Sequence assembly and analysis were performed using the program SEQUENCHER (Genes Code Corp.) and BIOEDIT (www.mbio.ncsu.edu/BioEdit/bioedit.html).

Validation of the EPLP procedure

Nucleotide sequence information for the genes amplified in the various *Paxillus* strains were translated and the protein alignments for each genes were made using CLUSTAL W (Thompson *et al*. 1994). Then the protein alignment was used as a template to align the corresponding nucleotide sequences. Homologous gene sequences from *C. cinereus* were identified using the EST2GENOME software as part of the EMBOSS package (Rice *et al*. 2000). These were added to the alignment (profile alignment). With the software MODELTEST, likelihood parameters were estimated (Posada & Crandall 1998). One of them is the gamma-distribution rate which accommodates for among-site

Table 2 Loci analysed in various strains of *Paxillus involutus*

	cDNA characterization (strain ATCC 200175)					Genetic variation[a]							
Gene	EST Accession no.	cDNA Accession no.	CDS (bp)	Protein (aa)	GenBank homologue	Strains analysed	Accession no.	Length (bp)	Introns (bp)	Exons (bp)	Part of cDNA (%)	Minimum identity (%)	ECM[b]
actA	CD274972	AY585923	1125	375	β-actin (*S. commune*) (Q9Y702)	6[c]	AY585949, 6027–31[g]	725	243	482	37	98	(+)[i]
β-tubA	CD274169	AY585924	1341	447	β-tubulin (*S. bovinus*) (CAD48933)	6[c]	AY585948, 6022–26[g]	568–571	252–255	316	20	95	
calA	CD272071	AY585925	447	149	Calmodulin (*P. cornucopiae*) (P11120)	5[d]	AY586017–21[g]	848–855	398–405	450	62	98	(+)[i]
cipC1	CD274659	AY585926	324	108	CipC (*E. nidulans*) (CAC87272)	5[d]	AY586008–12[h]	508–518	258–262	250–256	57	89	(+)[h,i]
cipC2	CD273165	AY585927	351	117	CipC (*E. nidulans*) (CAC87272)	6[c]	AY585947, 6003–7[h]	469–476	239–244	232–235	43	88	(+)[h,i]
cchA	CD273262	AY585928	210	70	Cu chaperone (*T. versicolor*) (AAN75572)	4[e]	AY586013–16[h]	286–287	130–131	156	40	90	(+)[h]
ppiA	CD270666	AY585929	492	164	Cyclophilin (*P. ostreatus*) (CAD10797)	6[c]	AY585941, 60–64[g]	495	0	495	80	96	(+)[i]
gpiA	CD274569	AY585932	1656	552	Glc-6-P isomerase (*A. bisporus*) (CAC87889)	6[c]	AY585946, 5998–6002[g]	1582–1586	98–102	1484	84	95	
lecA	CD275976	AY585930	429	143	Lectin (*X. chrysenteron*) (AAL73235)	5[d]	AY585973–77[h]	387–393	27–33	360	61	92	(+)[h,i]
gstA	CD273997	AY585931	636	212	Glutathione S-transferase (*N. fowleri*) (AAB01781)	5[d]	AY585993–97[h]	753–759	244–250	509	63	93	(+)[h,i]
hetC1	CD276279	AY585933	609	203	het-c2 (*P. anserina*) (S59950)	6[c]	AY585945, 88–92[g]	818–841	309–332	509	62	91	(+)[i]
hxtA	CD269657	AY585934	1569	523	Hexose transporter (*A. fumigatus*) (XP_747255)	5[d]	AY585978–82[g]	1925–1929	392–396	1533	84	95	(+)[i]
hspA	CD273217	AY585935	468	156	Small HSP (*L. bicolor*) (AAM78595)	6[c]	AY585944, 83–87[g]	552–562	108–112	444–450	62	88	
micA	CD272672	AY585936	897	299	Mitochondrial carrier (*C. neoformans*) (XP_569715)	4[f]	AY585943, 70–72[g]	1094–1108	322–336	772	68	93	(+)[i]
ndkA	CD271614	AY585937	456	152	NDP kinase (*N. crassa*) (XP_323542)	6[c]	AY585942, 65–69[g]	653	225	428	72	98	(+)[i]
ptrA	CD275864	AY585938	555	185	Pi transporter pho88 (*C. neoformans*) (XP_569621)	6[c]	AY585940, 55–59[g]	816–824	343–351	473	69	94	(+)[j]
rabA	CD273415	AY585939	633	211	Small GTPase (*C. neoformans*) (EAL20817)	5[d]	AY585950–4[h]	794–801	307–314	487	62	94	(+)[h]

[a]After design of primers based on EST sequence information from *P. involutus* ATCC 200175 (Johansson *et al*. 2004), genomic DNA fragments from the various strains were PCR-amplified, cloned and analysed by DNA sequencing.
[b](+) indicates that the gene is regulated in ECM tissue.
[c]Strains ATCC 200175, Pi01SE, Pi08BE, Maj, Nau and Pf01De.
[d]Strains ATCC 200175, Pi01SE, Pi08BE, Maj and Nau.
[e]Strains ATCC 200175, Pi08BE, Maj and Nau.
[f]Strains ATCC 200175, Pi01SE, Pi08BE and Pf01De.
[g]This study.
[h]Le Quéré *et al*. (2004).
[i]Le Quéré *et al*. (2005).
[j]Johansson *et al*. (2004).

 Molecular Ecology, 10.1111/j.1365-294X.2005.02796.x

variation. Maximum-likelihood trees were constructed and branch lengths were estimated with the aid of the PAUP software (Swofford 1998).

Results and discussion

Initial analysis of fluorescence intensities

Genomic DNA from five strains of *Paxillus involutus* and one strain of the closely related species *Paxillus filamentosus* were isolated and pairwise compared by hybridization on arrays containing cDNA reporters originating from the ATTC strain of *P. involutus* (reference strain) (Table 1, Fig. 1). Two different batches of microarray prints were used in these experiments. At first sight, the hybridization signals varied considerably between the two prints. However, the batch effect was only marginal when comparing the ratios of hybridization intensities (Fig. 2). When the $\log_2$ ratios of the hybridization signals were clustered, the ATCC, Pi01SE and Pi08BE strains were clustered into one group, the Maj and Nau strains into another group, and Pf01De at some distance from these two groups (Fig. 2). This partitioning is in agreement with a phylogeny based on ITS sequences, which position the ATCC, Pi01SE and Pi08BE strains into the so-called Forest clade, the Maj and Nau strains into the Park clade of *P. involutus*, whereas Pf01De of *P. filamentosus* falls outside these two clades (Le Quéré *et al*. 2004).

Normalizations of hybridization signals

The hybridization data were further analysed using a mixed-model analysis of variance (ANOVA) as implemented in SAS (Wolfinger *et al*. 2001). In SAS, like most other procedures for analysing microarray data, the hybridization signal intensities are converted to $\log_2$ scale and each data point is normalized by subtraction of the array mean $\log_2$ ratio value in order to centre the distribution on zero. In Fig. 3, the distributions of the resulting $\log_2$ fold changes in the hybridization signals for the sample strains relative to the reference strain of *P. involutus* have been plotted. As expected, the distributions of the fold changes comparing strains within the Forest clade (ATCC/Pi08BE and ATCC/Pi01SE) were almost identical and the main peaks were centred on zero. Similarly, the distribution comparing the two Park strains (Maj and Nau) was centred on zero (Fig. 3, inserted panel). In contrast, when comparing the ATCC strain with the phylogenetically more distant strains within the Park clade as well as *P. filamentosus* (ATCC/Maj, ATCC/Nau and ATCC/Pf01De), the main peaks of the distribution plots were shifted to the right (upwards). The shift is due to the presence of a large tail of divergent genes with weak hybridizations signals.

In CGH experiments, the main peak should primarily consist of conserved genes. One way for the identification

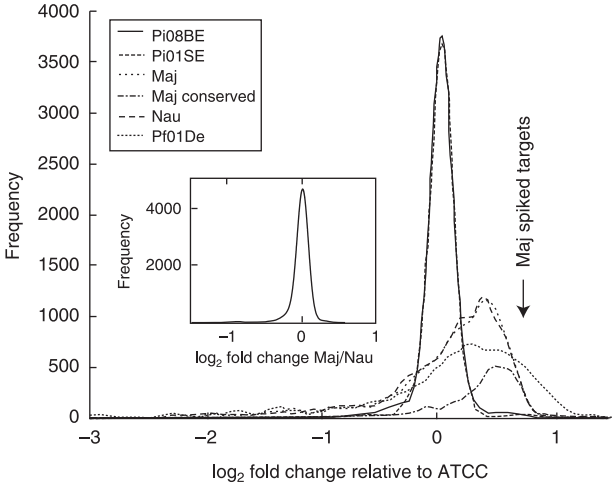

Fig. 3 Distribution of $\log_2$ fold changes in hybridization signals for various *Paxillus* strains relative to the reference strain ATCC. 'Maj conserved' show the distribution of fold changes for a conserved subset of genes in the strain Maj. Those were identified by a TBLASTX search including all the *Paxillus involutus* reporters present on the array against the three genome sequences of *Saccharomyces cerevisiae* (GenBank Accession nos. NC_001133 to NC_001148 and NC_001224), *Schizosaccharomyces pombe* (NC_003421, NC_003423, NC_003424, and NC_001326) and *Eremothecium gossypii* (NC_005782 to NC_005789). By using a TBLASTX cut-off score of 75 we defined 276 out of 1076 reporters to represent a cohort of conserved genes. The arrow 'Maj spiked targets' is the average $\log_2$ fold change for eight heterologous reporters. These were printed on the arrays and the corresponding targets were spiked at known amounts into the sample DNA prior to the labelling procedure (Table S2, Supplementary material).

of the correct main peak position is to plot the distribution for the most conserved genes within the various strains. An alternative method for determining the correct position of conserved genes within the fold-change distribution curve is by using hybridization ratios for a set of heterologous reporters (in this study, a total of 8). Both methods indicate that the main peak of the two Park strains (Maj and Nau) should be centred at a $\log_2$ fold change value relative to the ATCC of 0.7, whereas that of the Pf01De strain at 1.0 (Fig. 3). Consequently, the $\log_2$ fold change values reported below for the ATCC/Maj and ATCC/Nau comparisons have been shifted by –0.7, whereas the fold change values for the ATCC/Pf01De comparison by –1.0.

Sequence divergence and gene copy numbers

In CGH experiments, the variation in hybridization signals depends on several factors, including sequence divergence between the sample and reference DNA, and differences in gene copy number (Wu *et al*. 2001; Hinchliffe *et al*. 2003). To investigate how sequence divergence affected the hybridization signals in our study, 17 loci were sequenced

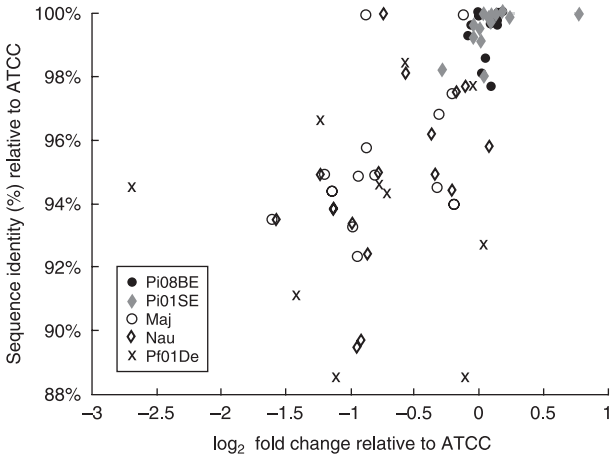

Fig. 4 Relationship between sequence identity and fold changes in hybridization signals for 17 loci in various *Paxillus* strains relative to the ATCC reference strain. The sequence identity in exon regions for each gene (cf. Table 2) is plotted against its corresponding $\log_2$ fold change in hybridization signals for pairwise comparisons between the reference and the five other *Paxillus* strains (Pi08BE, Pi01SE, Maj, Nau and Pf01De).

in the different strains of *Paxillus* (Table 2). Depending on the position of the PCR primers, the regions analysed covered 20–84% of the predicted exons. Considering data from all genes and all pairwise comparisons, a weak correlation ($r^2 = 0.30$, $P < 0.02$) was found between sequence identity and $\log_2$ fold changes in hybridization signals (Fig. 4).

Further insight into the source of variation was achieved by analysing the data of strains from a single clade and those from two different clades separately. Analysis of the strains of the Forest clade showed that all genes except *ptrA* displayed at least 97.7% sequence identity, and their $\log_2$ fold changes in hybridization signals varied between –0.3 to +0.3. The $\log_2$ fold change value for *ptrA* when comparing ATCC/Pi01SE was significantly outside this narrow range (0.78, corresponding to an antilog fold change value of 1.7). Considering the fact that the sequence identity for *ptrA* was 100%, we propose that this gene has been duplicated in Pi01SE relative to the ATCC strain. Similar values were obtained for *ptrA* when comparing data from the Maj and Nau strains within the Park clade (sequence identity 100%, antilog fold change value of 1.9) which suggest that this gene is also duplicated in Maj relative to the ATCC strain.

The variations in sequence identity and hybridization signals were considerably larger when contrasting the ATCC strain with strains from other clades. When comparing data for the ATCC strain and the two Park strains, one outlier was identified, namely the *ppiA* gene. The $\log_2$ fold changes in hybridization signals for *ppiA* when comparing ATCC/Maj and ATCC/Nau were in both cases –0.7, a value that could be expected for genes having a sequence identity in the range 92–96%. However, the sequence identity for *ppiA* was found to be 100% and all information considered we classify *ppiA* as putatively duplicated. Excluding *ppiA* from the analysis, the linear correlation coefficients of the relationship between sequence identities and $\log_2$ fold changes for all genes included in the pairwise comparisons of ATCC/Maj and ATCC/Nau were 0.62 ($P < 0.03$) and 0.64 ($P < 0.02$), respectively. In *P. filamentosus*, only 10 of the 17 loci could be investigated, presumably due to the low sequence identity between the primers used and the corresponding genes in Pf01De. Furthermore, the linear correlation coefficients of the relationship between sequence identities and $\log_2$ fold changes for the PCR-amplified genes in the pairwise comparisons ATCC/Pf01De were not significant.

The above analyses suggest that when comparing closely related strains within the Forest or Park clades, respectively, the fold changes in CGH hybridization signals will primarily be associated with differences in gene copy numbers and not sequence divergence. In contrast, when comparing strains from different clades, the variation in hybridization signals can be related to both sequence divergence and copy number differences. To distinguish between these two processes, information on both hybridization signal intensities and sequence divergence are needed.

Identification of locally divergent genes using the EPLP approach

To screen for genes diverging at an enhanced rate within the lineage of *P. involutus*, we developed a simple rate test using information from both variation in hybridization signals and sequence similarities to the basidiomycete *Coprinus cinereus* (Fig. 5). Basically, the hybridization signals for any given gene is compared to the signals of genes displaying a similar degree of sequence similarity to *C. cinereus*. An algorithm that depends on the shape of the signal-ratio distribution curve for this cohort of genes provides an estimate (EPLP) of the degree of divergence (Fig. 6).

Using a EPLP cut-off value of 0.05, the numbers of locally divergent genes identified in the pairwise comparisons between strains from the three lineages of *Paxillus*, the ATCC/Park clade (average Maj and Nau) (Fig. 7A), the ATCC/*P. filamentosus* (Fig. 7B) and the Park clade/*P. filamentosus* (not shown) were 106, 64 and 102, respectively. In total, we identified a cohort of 177 genes that were locally divergent according to this procedure in at least one of the three comparisons made (Table S3, Supplementary material).

The EPLP algorithm is similar to the algorithm used within the so-called GACK procedure for analysing array-CGH data (Kim *et al*. 2002). Like other more traditional methods to analyse CGH array data (Porwollik *et al*. 2002; Hinchliffe *et al*. 2003), GACK categorizes genes as being

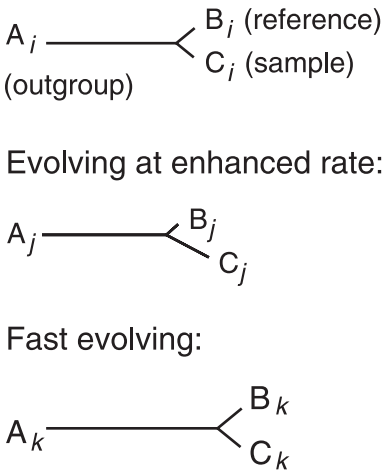

Fig. 5 The EPLP screen for identifying locally divergent genes. (A) represents an organism with a fully sequenced genome located outside the clade of strains analysed by CGH. (B) here is the reference strain, which is also the strain used for the construction of the microarray. (C) is a closely related sample strain analysed by CGH hybridizations. In our experiments (A) was the basidiomycete *Coprinus cinereus*, which is the closest evolutionary relative to *Paxillus involutus* that has been fully sequenced. (B) was the ATCC strain of *P. involutus* and (C) was other strains of *P. involutus* or *Paxillus filamentous*. The distance of genes between (A) and (B) is estimated by using the TBLASTX algorithm. The divergence of genes between (B) and (C) is measured by CGH analysis. In the screen, the observed CGH fold changes for any given gene is compared with the fold changes from a cohort of genes displaying a similar distance to the outgroup. This cohort contains approximately 100 genes having a similar TBLASTX bit score as the analysed gene. An algorithm that depends on the shape of the signal-ratio distribution curve for this cohort of genes provides an estimate (the estimated probability of local presence, EPLP) of the degree of divergence (Fig. 6). Shown is the hypothetical evolutionary relationship between three genes *i*, *j* and *k*. Sequence comparisons between (A) and (B) show that *i* and *j* have the same overall evolutionary rate, while *k* evolves faster. Analysis of the CGH hybridization signals between (B) and (C) indicates that *j* and *k* are more divergent than *i*. Note that the position of the internal node along the branch separating strains B and C from A is not known, as the distance between A and C has not been determined. The EPLP procedure will select *j* but not *k* as evolving at an enhanced rate. According to the neutral theory of molecular evolution, a shift in the rate of evolution may indicate an alteration in the selection pressure on the genes.

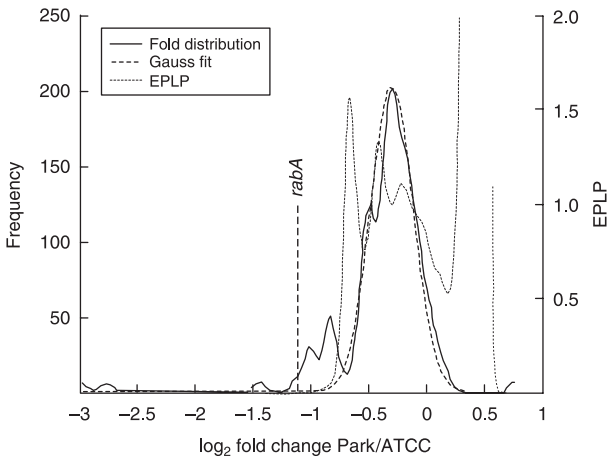

Fig. 6 Application of the EPLP algorithm. The method is illustrated here for the *rabA* gene (cf. Table 2), which has a TBLASTX score of 105 against *Coprinus cinereus*. The solid line shows the distribution of $\log_2$ fold change in hybridization signals between Park and ATCC for 102 genes with a similar TBLASTX score against *C. cinereus* as for the *rabA* gene. The 102 genes had a TBLASTX score from 95 to 117. This is the smallest interval containing at least 50 genes with a smaller TBLASTX score than *rabA*, and also genes with a larger TBLASTX score. The dashed line shows a normal curve fitted to the main peak of the fold distribution curve of the 102 genes. This normal curve represents the expected local $\log_2$ fold distribution of genes which are conserved to the same degree as the *rabA* gene. The estimated probability of local presence (EPLP), marked as a dotted line, shows the expected distribution density (dashed line) divided by the real density of genes (solid line). The *rabA* gene has a $\log_2$ fold change Park/ATCC of –1.14 and is located in the tail of the divergent genes. The EPLP value for *rabA* is 0.001, which is smaller than the cut-off value of 0.05 used to separate conserved and variable genes. Thus *rabA* was classified as a locally divergent gene.

variable or conserved solely on basis of the hybridization signal. In GACK the signal-ratio distribution curve is analysed for all reporters and one EPP (estimated probability of presence) function is calculated that is fixed for all genes. Using an EPP cut-off value of 0.05, we identified a cohort of 195 genes that were divergent according to the GACK procedure. One hundred forty-one of these were also identified as divergent using the EPLP method. However, 36 out of the 177 genes classified as divergent using the EPLP procedure were not identified as divergent using the GACK procedure. These included mainly genes displaying medium to high sequence similarities to *C. cinereus* (Fig. 7).

The EPLP methods' ability to predict genes evolving at an enhanced rate was validated using information from the 17 sequenced loci (cf. Table 2). The sequences were analysed using a procedure that is based on the comparison of phylogenetic branch lengths of orthologous proteins from three species (Jordan *et al*. 2001). Adopted to our data set, the length of the branch separating the ATCC strain and *C. cinereus* ($dist_{ATCC,\ C.\ cinereus}$) was compared with the sum of the branch separating the ATCC and the sample strain ($dist_{ATCC,\ sample}$). Consistent with the rate-constancy prediction of neutral evolution, the ($dist_{ATCC,\ C.\ cinereus}$)/($dist_{ATCC,\ sample}$) ratio should be approximately constant. Accelerated evolution should be manifested by a low ratio. The sequence- and the EPLP-based measures of accelerated evolution were indeed related (Fig. 8). The EPLP method

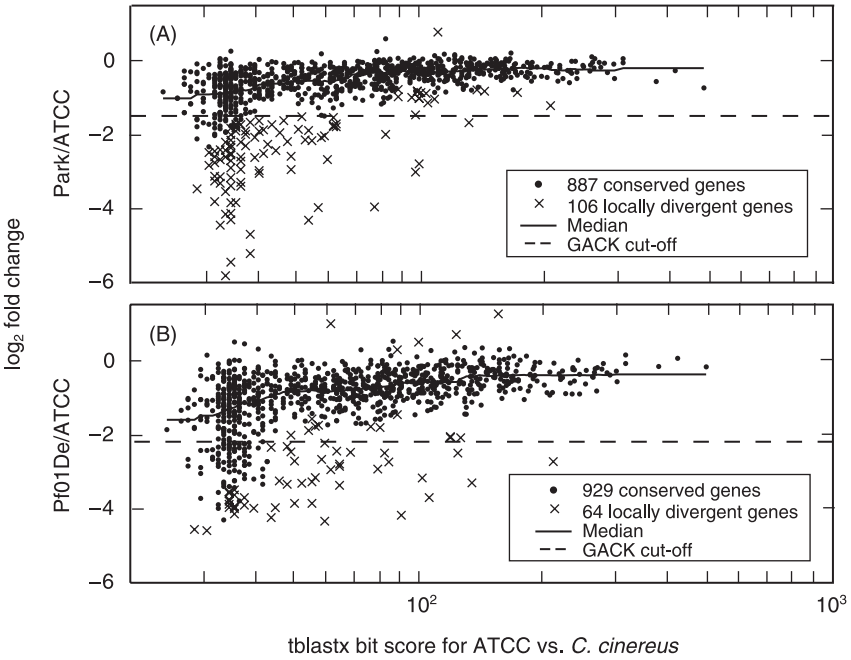

Fig. 7 Locally divergent genes identified in *Paxillus involutus* using the EPLP procedure. The scatter plots show the $\log_2$ fold changes in hybridization signals between (A) the reference strain ATCC and Park, and (B) the ATCC and the *Paxillus filamentosus* strain Pf01De, vs. the sequence similarity (TBLASTX bit score) for each reporter to homologous genes in the genome of *Coprinus cinereus*. The Park $\log_2$ fold change was calculated as the average $\log_2$ fold change for Maj and Nau. The median $\log_2$ fold change (solid line) was estimated in a window using 101 reporters showing similar TBLASTX scores against *C. cinereus*. The locally divergent genes were identified by contrasting the fold values of the clones to that of 101 genes with similar TBLASTX scores (cf. Fig. 6). The dashed lines 'GACK cut-off' show the $\log_2$ fold ratio value discriminating between divergent (below the line) and conserved genes according to the GACK procedure (Kim *et al*. 2002).

when comparing the ATCC and Maj strains predicts that two genes (*rabA* and *gpiA*) were evolving at an enhanced rate. The genes *rabA* and *gpiA* were also among those having the lowest $(\text{dist}_{\text{ATCC, } C.\ cinereus})/(\text{dist}_{\text{ATCC, samples}})$ ratio. Note that *rabA* and *gpiA* were not identified as being divergent using the GACK procedure. The EPLP procedure appeared to fail to detect two genes with low branch lengths ratios, namely *gstA* and *calA*. For *gstA* the normal distribution is not appropriate (Fig. S2, Supplementary material). The *calA* gene is one of the most conserved genes comparing *P. involutus* and *C. cinereus*. Thus only a few mutations in this gene can affect the ratios of the branch lengths significantly. Such small changes could presumably not be detected by CGH analysis. The above analysis indicates that the EPLP procedure could be used for screening genes diverging at an enhanced rate. However, it should be possible to optimize the algorithm further given a larger set of genes with an a priori known sequence history.

Orphans

The genomes of fungi and other organisms contain a significant portion of genes that exhibit no significant similarity to protein sequences present in databases (Tunlid & Talbot 2002). Such orphans may represent genes whose phylogenetic distribution is restricted to certain evolutionary lineages. Orphan genes might also represent genes that rapidly diverge between closely related strains or species.

From a total of 1076 *Paxillus* gene representatives on the array, 382 showed a TBLASTX score below 45 when compared to *C. cinereus*. These orphans varied in hybridization signals and represented both conserved (large $\log_2$ fold change values) and variable genes (small $\log_2$ fold change values) (Fig. 7). Since the EPLP procedure cannot be used for analysing genes in cohorts displaying low similarity scores to genes in other organisms, other methods are needed to discriminate between conserved and variable orphans. In the Supplementary material, we show that the distribution of fold changes for the well-conserved genes can be used to estimate an upper number of conserved orphans (Fig. S1, Supplementary material). In total, 52 orphan genes might be just as conserved between the Park and Forest clades as the genes showing a TBLASTX score against *C. cinereus* of 100 or above (for Pf01De/ATCC, the corresponding number was 44). Conserved orphan genes with very low divergence rates have also been identified in *Drosophila* (Domazet-Loso & Tautz 2003). The authors proposed that these slowly evolving orphan genes might represent genes that have evolved to perform lineage-specific functions.

Functional classification of divergent genes

The locally divergent genes showed an under-representation of genes predicted to be involved in protein biosynthesis and those encoding structural molecules (Fig. 9A). Both of these categories included ribosomal proteins. In contrast, the cohort of locally divergent genes was over-represented by orphans, proteins predicted to be located at membranes and those involved in transport and lipid metabolism. The predicted membrane proteins

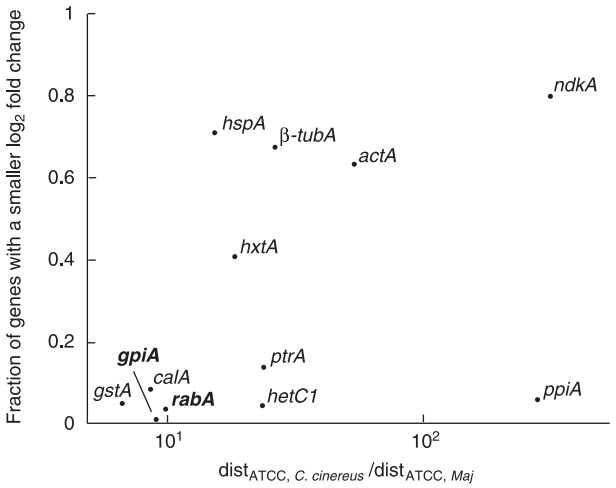

Fig. 8 Validation of the EPLP procedure. Plotted is the relationship between an EPLP estimate of divergence vs. a sequence-based measures of accelerated evolution. Both methods analyse how fast genes are evolving between Maj (from the Park clade) and ATCC (Forest clade) strains of *Paxillus involutus* as compared with the rate between *P. involutus* and *Coprinus cinereus*. Data are shown for 12 genes sequenced in *P. involutus* (cf. Table 2) and for which homologues were identified in *C. cinereus*. The *y*-axis value indicates the relative position of the gene in the $\log_2$ fold change distribution curve of genes displaying a similar degree of sequence similarity to *C. cinereus* (cf. Fig. 6). Specifically, it is the percentile of genes with a smaller fold change value than the analysed gene. The sequence-based measure (*x*-axis) compare the length of the branch separating the ATCC strain and *C. cinereus* ($dist_{ATCC, C. cinereus}$) with the sum of the branch separating the ATCC and the sample strains ($dist_{ATCC, Maj}$). The branch lengths were calculated using the evolutionary distances between the genes of the three organisms as described in Materials and methods. Accelerated evolution should be manifested by a low ($dist_{ATCC, C. cinereus}$)/($dist_{ATCC, Maj}$) ratio.

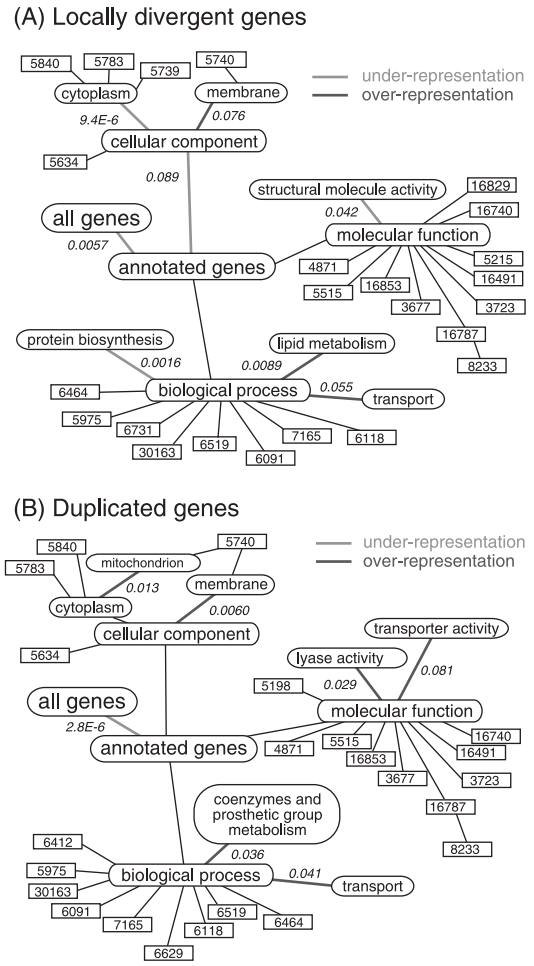

Fig. 9 Functional annotations of (A) locally divergent and (B) duplicated genes. Based on sequence similarities, the *Paxillus* genes were annotated into GO categories organized as molecular function, biological process and cellular component (Ashburner *et al*. 2000). The relationships between GO categories at different levels of specialization (parents and child terms) are displayed as directed acyclic graphs (DAGs). In the figure 'all genes' represents the top-level parent, and more specialized terms are connected by lines. A statistical test was developed to compare the GO distribution for the variable and duplicated genes with the distribution observed for the entire set of arrayed reporters. Briefly, we tested whether the frequencies of genes in a pair of a parent and a child term among the locally divergent or duplicated genes were significantly different from the frequencies observed in the complete set of arrayed genes (Table S5, Supplementary material). A thick line indicates a parent–child pair in which the child term is either significantly ($P < 0.05$) over-represented (blue) or under-represented (red). Descriptions of the GO terms can be found in Table S5 (Supplementary material).

displayed significant sequence similarities to several larger families of transport proteins (Table 3): the mitochondrial carrier proteins (Vozza *et al*. 2004), the drug-resistant subfamily of the major facilitator superfamily (Goffeau *et al*. 1997), the MAPEG (membrane-associated proteins in eicosanoid and gluthathione metabolism) superfamily (Jakobsson *et al*. 1999), and the ELO, GNS1/SUR4 family which is involved in long-chain fatty acid synthesis (Rossler *et al*. 2003). Among the cohort of locally divergent genes were also genes displaying sequence similarities to thioredoxins, thiol peroxidases and glutathione S-transferases. These components of the thioredoxin and glutathione/glutaredoxin system are important for the regulation of the intracellular redox status and the detoxification of oxidation products generated in various defence reactions (Grant 2001). Among variable genes were also a gene showing similarities to members of the cytochrome P450 gene family, which are important in the oxidative metabolism of endogenous and xenobiotic compounds (Nelson 1999). Notably, orphans, genes whose products are located at membranes, or genes encoding for components of stress/defence reactions are also known to evolve at an accelerated rate in other organisms including bacteria and mammals (Jordan *et al*. 2001).

Table 3 Functional predictions of variable genes*

dbEST Accession no.	Uniprot Accession no.	Best hit description	Gene Ontology (GO)		
			Molecular function	Cellular compartment	Biological process
CD274586	Q9UTF7	GNS1/SUR4 family protein	Fatty acid elongase activity	Integral to membrane	Vesicle-mediated transport; sphingolipid biosynthesis; fatty acid elongation
CD275161	Q9USN4	Putative transporter C1529.01	Transporter activity	Membrane	Transport
CD275133	Q7SHF8	Predicted protein	Binding	Membrane	Transport
CD272672	O74439	Mitochondrial carrier protein	Binding	Integral to membrane	Transport
CD273558	Q9HEM5	Related to microsomal glutathione S-transferase	Transferase activity	Membrane fraction; Microsome	Lipid metabolism; Signal transduction
CD272657	P48011	DNA-directed RNA polymerases I, II, and III	DNA binding; DNA-directed RNA polymerase activity	Nucleus	Transcription
CD274951	O23676	Mago nashi protein homolog		Nucleus	Sex determination
CD273709	Q8BPF9	Casein kinase II	Protein kinase CK2 activity		Regulation of cell cycle
CD272606	Q872E3	Related to MNORI-2 protein (hypothetical protein)	Protein binding		Protein amino acid phosphorylation
CD273762	Q9Y4Y9	U6 snRNA-associated Sm-like protein LSm5	RNA binding	Nucleus	mRNA processing
CD274085	O13639	Adenosyl homocysteinase (EC 3.3.1.1)	Adenosyl homocysteinase activity	Cytoplasm	Methionine metabolism; Selenocysteine metabolism
CD276165	Q7S2L1	Hypothetical protein	Oxidoreductase activity		Metabolism
CD271655	P00440	Tyrosinase precursor (EC 1.14.18.1)	Oxidoreductase activity		Metabolism
CD274569	Q9HGZ2	Glucose-6-phosphate isomerase (EC 5.3.1.9)	Glucose-6-phosphate isomerase activity; Isomerase activity		Gluconeogenesis; Glycolysis
CD269625	Q82HX7	Putative monooxygenase	Monooxygenase activity; Disulphide oxidoreductase activity		Electron transport; aromatic compound metabolism
CD275123	Q9HFJ1	Related to n-alkane-inducible cytochrome P450	Monooxygenase activity		Electron transport
CD271249	Q9UW02	Thioredoxin (Allergen Cop c 2)	Electron transporter activity		Electron transport
CD271304	Q38879	Thioredoxin H-type 2 (TRX-H-2)	Electron transporter activity		Electron transport
CD269698	Q7NX63	Probable isovaleryl-CoA dehydrogenase	Isovaleryl-CoA dehydrogenase activity; Oxidoreductase activity		Electron transport
CD273899	Q9NL98	Peroxiredoxin (EC 1.11.1.-) (AsPrx)	Peroxidase activity		
CD274635	O14064	Bir1 protein (chromosome segregation protein …			
CD275828	O74162	Ich1	O-methyltransferase activity		
CD276314	Q7SGE9	Hypothetical protein	Alcohol dehydrogenase activity; zinc-dependent; zinc ion binding		

Table 3 *Continued*

dbEST Accession no.	Uniprot Accession no.	Best hit description	Gene Ontology (GO)		
			Molecular function	Cellular compartment	Biological process
CD271867	Q871S2	Related to acid sphingomyelinase	Hydrolase activity		
CD270379	Q25556	Glutathione S-transferase III homolog	Transferase activity		
CD270527	Q8Y0Q1	Probable glutathione S-transferase-related *trans* …	Glutathione transferase activity; Transferase activity		
CD269885	Q10344	Translationally controlled tumour protein		Cytoplasm	

*The table lists 27 genes that were identified as being locally divergent (Figs 5–7). Only genes found to be significantly different in at least two of the three pairwise comparisons and displaying a significant homology to proteins in the UniProt database (Apweiler *et al*. 2004) are listed. A full list of locally divergent genes (177 in total) can be found in the Table S3 (Supplementary material). GO annotations (Ashburner *et al*. 2000) were inferred by retrieving information from the UniProt entry. The underlined gene was identified as duplicated within the clades analysed (Table S4, Supplementary material). Genes in bold were characterized by DNA sequencing (cf. Table 2).

The locally divergent genes were also over-represented ($P < 0.01$) by genes being homologous to members within large gene families identified in the basidiomycete *Phanerochaete chrysosporium* (Martinez *et al*. 2004) (Table S6, Supplementary material). Several of these large gene families have been shown to be rapidly expanding when comparing lineages of more distantly related eukaryotes (Lespinet *et al*. 2002; van Nimwegen 2003). Among them are the major facilitator superfamily transporters, the cytochrome P450 family hydroxylases, and the glutathione S-transferases.

Gene duplications

To identify genomic differences that could be associated with variations in host specificities, the genomes of the closely related Park strains Maj and Nau were compared. The Maj strain forms ECM with birch and poplar, while Nau is incompatible with these trees (Gafur *et al*. 2004; Le Quéré *et al*. 2004). The $\log_2$ fold changes for Maj and Nau relative to the ATCC strain are shown in Fig. 10. Most of the reporters were scattered along the diagonal and thus had highly similar hybridization signals in Maj and Nau. However, there were 21 outliers with $\log_2$ fold changes > 0.5 or < 0.5, which indicate that the genes are found in different copy numbers in the two strains. Among these putatively duplicated genes, 14 were identified by EST-derived reporters whereas 7 by cosmid-derived reporters (Table S4, Supplementary material).

Only 2 out of the 14 duplicated genes identified by EST-derived reporters displayed significant sequence similarities to proteins in the GenBank nr protein database (Benson *et al*. 2005). One of them corresponded to the sequenced loci *ptrA* (Table 2). The *ptrA* gene translates into a polypeptide of 185 amino acid residues that shows a high sequence identity (43%) to Pho88p in *Saccharomyces cerevisiae*. Pho88p is a putative membrane protein involved in inorganic phosphate transport and regulation (Yompakdee *et al*. 1996). The second duplicated gene (corresponding to the EST clone CD274497) showed a significant similarity to a hypothetical protein in *Neurospora crassa*. The 7 cosmid-derived reporters that indicated variation in copy number between Maj and Nau all originate from one continuous 3.2-kb genomic region, in positions covering the putative ORF PiC1-11 and adjacent DNA regions (Table S1, Supplementary material). PiC1-11 displays a high sequence similarity to a WD40 repeat motif (Le Quéré *et al*. 2002). This domain acts as site for interaction with other proteins, and there are 102 proteins in *S. cerevisiae* with at least one copy of this motif (IPR001680).

We have previously shown that approximately 66 (6%) of the arrayed genes are differentially expressed in Maj and Nau following the contact with the roots of birch seedlings (Le Quéré *et al*. 2004). Notably, none of the duplicated genes identified in this study were among these differentially expressed genes. Thus, the differences in expression levels cannot be explained by differences in gene copy numbers. Most probably, the observed differences in expression levels are due to variation in promoter elements or levels of transcription factors.

An analysis was also performed to identify duplicated genes between strains of the Forest clade. In total, we identified 56 genes and one cosmid-derived fragment that were found in different copy numbers in at least one of the three pairwise comparisons made between ATCC/Pi01SE, ATCC/Pi08BE, and Pi01SE/Pi08BE. Notably, six of these duplicated genes including *ptrA* were among the genes that also varied in copy number in the comparison between Maj and Nau. Altogether, a cohort of 64 genes was identified as being duplicated in at least one of the four

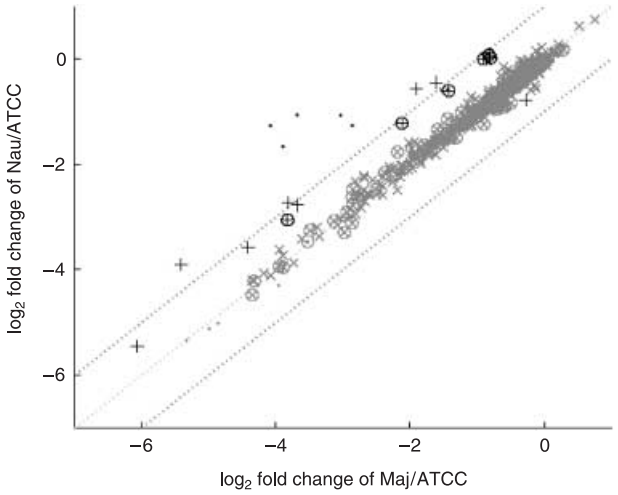

Fig. 10 Identification of genes being duplicated in the compatible strain Maj and the incompatible strain Nau (i.e. not infecting birch or poplar). The scatter plot shows the $\log_2$ fold change in hybridizations signals of Maj relative to ATCC (reference strain) vs. the $\log_2$ fold change of Nau relative to ATCC, for 1019 reporters (990 EST and 29 cosmid-derived reporters). The dotted line is the diagonal showing genes with almost identical hybridization signals in Maj and Nau. The position for each gene along the diagonal is a measure of its divergence between the two Park clade strains Maj and Nau relative to the ATCC strain of the Forest clade, with the most divergent genes scattered towards the lower left corner of the plot. The dotted lines at $y = x + 1$ and $y = x - 1$ correspond to a $\log_2$ fold change between Nau and Maj of 1 and –0.5, respectively. Genes that have been duplicated in Nau relative to Maj or vice versa are expected to scatter along these lines. Indicates nonduplicated genes (in total 976); duplicated genes (14); non duplicated cosmid-derived reporters (22); duplicated cosmid-derived reporters (7); duplicated genes in the Forest clade (49); duplicated genes in both the Forest and Park clades. All duplicated genes identified to be duplicated in the analysed *Paxillus* strains are listed in the Supplementary material (Table S4).

pairwise comparisons made between the strains within the Forest and Park clades, respectively (Table S4, Supplementary material). Notably, a large fraction of these duplicated genes (40 out of 64) were also identified as being locally divergent when the different clades of *P. involutus* and *P. filamentosus* were compared. The fraction of genes displaying no significant homology to proteins in the UniProt database was under-represented among the duplicated genes (Fig. 9B). Similarly to the pattern observed for the divergent genes, the genes predicted to be localized to membranes and involved in transport were over-represented among the duplicated genes.

Conclusions

Due to the fact that the costs for EST sequencing and the fabrication of microarrays are rapidly decreasing, we foresee that in the near future DNA microarray analysis will become a common tool for comparing genome composition in many organisms. Here we present several novel procedures for the analysis of such data. We suggest methods for the normalization of hybridization data that correct for the presence of a large number of variable genes yielding weak signals, which typically complicates CGH analyses. We have shown that the hybridization signal in the CGH experiments depends on both sequence divergence and gene copy number. When comparing closely related strains, the fold changes in CGH hybridization signals will primarily be associated with differences in gene copy number and not sequence divergence. In contrast, when comparing more distantly related strains, the variation in hybridization signals can be related to both sequence divergence and gene copy number. To distinguish between these two processes, information on both hybridization signal intensity and sequence divergence are needed.

We developed a simple rate test, the EPLP procedure, to screen for genes diverging at an enhanced rate. Such changes in the rate of evolution may indicate cases of functional diversification associated with adaptations. The comparison is made by contrasting the observed fold change in hybridization signal for each gene with the signals from a cohort of genes displaying a similar degree of sequence similarity to an outgroup organism. In principle, any organism with a fully sequenced genome can be used as an outgroup as long as it is more distantly related than the species or lineages being examined by CGH. However, the genomes should be close enough to avoid saturation of nucleotide substitution. Presently, the CGH-based procedure for detecting non-neutral evolving genes has only been validated using sequence data from a limited set of genes. In fact we observed a correlation between the CGH-based measure of non-neutrality and a standard phylogenetic analysis. The analysis also indicated that it should be possible to further optimize the algorithm given a larger set of genes with an a priori known sequence history. In any case, the 'candidates' identified by the EPLP screen should be sequenced to verify their rate of divergence and to identify possible selection mechanisms acting on the genes.

The developed procedures were used to screen for duplicated and rapidly evolving genes in strains of the ECM fungus *Paxillus involutus*. Approximately 17% of the printed genes were detected as rapidly and presumably non-neutrally evolving within *P. involutus*. Furthermore, 6% of the analysed genes varied in gene copy numbers. The cohort of divergent and duplicated genes showed an over-representation of orphans, genes whose products are located at membranes, and genes encoding for components of stress/defence reactions. Some of the identified genomic changes may be associated adaptations to the symbiotic lifestyle, including variations in host specificity

of ECM fungi. However, due to the fact that there are other, more closely related species to *Paxillus* than *Coprinus cinereues* that are nonmycorrhizal, part of the detected genomic changes might be associated with adaptations to nonsymbiotic growth.

Acknowledgements

This study was supported by grants from the Swedish Research Council. Kasper Astrup Eriksen acknowledges support from both the Danish Natural Science Research Council (grant number 21-03-0284) and the Bio+IT programme under the Øresund Science Region and Øforsk. Andres Schützendübel received financial support through a Marie Curie Fellowship. Custom microarrays were produced at the SWEGENE DNA Microarray Resource Center at the Bio Medical Center B10 in Lund, and DNA sequencing was performed at the SWEGENE Center of Genomic Ecology at the Ecology Building in Lund, supported by the Knut and Alice Wallenberg Foundation through the SWEGENE consortium. We thank Eva Friman for help with DNA sequencing and Charles Kurland for stimulating discussions.

Supplementary material

The supplementary materials are available from http://www.blackwellpublishing.com/products/journals/suppmat/MEC/MEC2796/MEC2796sm.htm

Table S1 Regions within a 32-kb genomic fragment being duplicated among strains of *Paxillus* as determined by CGH analysis

Table S2 Heterologous control DNA for the validation of dual-label ratio analysis of microarray data

Table S3 Locally divergent genes identified according to the EPLP procedure in the strains of *Paxillus*

Table S4 Duplicated genes in the strains of *Paxillus*

Table S5 Gene Ontology (GO) annotations of locally divergent and duplicated genes

Table S6 Protein domains of arrayed *Paxillus* genes

Fig. S1 Estimation of the number of conserved orphans in the *Paxillus involutus* strains.

Fig. S2 The EPLP procedure for the genes *gstA*, *calA*, *gpiA* and *rabA* in *Paxillus involutus*.

Fig. S3 The EPLP procedure for the genes *hspA*, *hetA*, *ptrA* and *hetC1* in *Paxillus involutus*.

Fig. S4 The EPLP procedure for the genes β-*tubA*, *actA*, *ppiA* and *ndkA* in *Paxillus involutus*.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Apweiler R, Bairoch A, Wu CH *et al*. (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, **32**, D115–D119.

Ashburner M, Ball CA, Blake JA *et al*. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, **25**, 25–29.

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

Blaudez D, Chalot M, Dizengremel P, Botton B (1998) Structure and function of the ectomycorrhizal association between *Paxillus involutus and Betula pedula*. II. Metabolic changes during mycorrhiza formation. *New Phytologist*, **138**, 543–552.

Brun A, Chalot M, Finlay RD, Söderström B (1995) Structure and function of the ectomycorrhizal association between *Paxillus involutus* (Batsch) Fr. and *Betula pendula* (Roth.). I. Dynamics of mycorrhiza formation. *New Phytologist*, **129**, 487–493.

Chalot M, Brun A, Botton B, Söderström B (1996) Characterization of the general amino acid transporter from the ECM fungus *Paxillus involutus*. *Microbiology*, **142**, 1749–1756.

Domazet-Loso T, Tautz D (2003) An evolutionary analysis of orphan genes in *Drosophila*. *Genome Research*, **13**, 2213–2219.

Dunham MJ, Badrane H, Ferea T *et al*. (2002) Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences, USA*, **99**, 16144–16149.

Edwards-Ingram LC, Gent ME, Hoyle DC *et al*. (2004) Comparative genomic hybridization provides new insights into the molecular taxonomy of the *Saccharomyces* sensu stricto complex. *Genome Research*, **14**, 1043–1051.

Gafur A, Schützendübel A, Langenfeld-Heyser R, Fritz E, Polle A (2004) Compatible and incompetent *Paxillus involutus* isolates for ectomycorrhiza formation *in vitro* with poplar (*Populus* × *canescens*) differ in H_2O_2 production. *Plant Biology*, **6**, 91–99.

Goffeau A, Park J, Paulsen IT *et al*. (1997) Multidrug-resistant transport proteins in yeast: complete inventory and phylogenetic characterization of yeast open reading frames with the major facilitator superfamily. *Yeast*, **13**, 43–54.

Grant CM (2001) Role of the glutathione/glutaredoxin and thioredoxin systems in yeast growth and response to stress conditions. *Molecular Microbiology*, **39**, 533–541.

Harvey PH, Pagel MD (1991) *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford, UK.

Hibbett DS, Gilbert L-B, Donoghue MJ (2000) Evolutionary instability of ectomycorrhizal symbiosis in basidiomycetes. *Nature*, **407**, 506–508.

Hibbett DS, Pine EM, Langer E, Langer G, Donoghue MJ (1997) Evolution of gilled mushrooms and puffballs inferred from ribosomal DNA sequences. *Proceedings of the National Academy of Sciences, USA*, **94**, 12002–12006.

Hinchliffe SJ, Isherwood KE, Stabler RA *et al*. (2003) Application of DNA microarrays to study the evolutionary genomics of *Yersinia pestis* and *Yersinia pseudotuberculosis*. *Genome Research*, **13**, 2018–2029.

Hughes AL (1999) *Adaptive Evolution of Genes and Genomes*. Oxford University Press, Oxford, UK.

Hughes TR, Roberts CJ, Dai H *et al*. (2000) Widespread aneuploidy

revealed by DNA microarray expression profiling. *Nature Genetics*, **25**, 333–337.

Jakobsson PJ, Morgenstern R, Mancini J, Ford-Hutchinson A, Persson B (1999) Common structural features of *MAPEG* — a widespread superfamily of membrane associated proteins with highly divergent functions in eicosanoid and glutathione metabolism. *Protein Science*, **8**, 689–692.

Jarosch M, Bresinsky A (1999) Speciation and phylogenetic distances within *Paxillus* s. str. (Basidiomycetes, Boletales). *Plant Biology*, **1**, 701–706.

Johansson T, Le Quéré A, Ahrén D *et al.* (2004) Transcriptional responses of *Paxillus involutus* and *Betula pendula* during formation of ectomycorrhizal root tissue. *Molecular Plant-Microbe Interactions*, **17**, 202–215.

Jordan IK, Kondrashov FA, Rogozin IB *et al.* (2001) Constant relative rate of protein evolution and detection of functional diversification among bacterial, archaeal and eukaryotic proteins. *Genome Biology*, **2**, 53.1–53.9.

Kim CC, Joyce EA, Chan K, Falkow S (2002) Improved analytical methods for microarray-based genome-composition analysis. *Genome Biology*, **3**, 65.1–65.65.17.

Kimura M, Ota T (1974) On some principles governing molecular evolution. *Proceedings of the National Academy of Sciences, USA*, **71**, 2848–2852.

Le Quéré A, Johansson T, Tunlid A (2002) Size and complexity of the nuclear genome of the ectomycorrhizal fungus *Paxillus involutus*. *Fungal Genetics and Biology*, **36**, 234–241.

Le Quéré A, Schützendübel A, Rajashekar B *et al.* (2004) Divergence in gene expression related to variation in host specificity of an ectomycorrhizal fungus. *Molecular Ecology*, **13**, 3809–3819.

Le Quéré A, Wright DP, Söderström B, Tunlid A, Johansson T (2005) Global patterns of gene regulation associated with the development of ectomycorrhiza between birch (*Betula pendula* Roth.) and *Paxillus involutus* (Batsch) Fr. *Molecular Plant Microbe Interaction*, **18**, 659–673.

Lespinet O, Wolf YI, Koonin EV, Aravind L (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Research*, **12**, 1048–1059.

Long M, Betran E, Thornton K, Wang W (2003) The origin of new genes: glimpses from the young and old. *Nature Reviews: Genetics*, **4**, 865–875.

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.

Martinez D, Larrondo LF, Putnam N *et al.* (2004) Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. *Nature Biotechnology*, **22**, 695–700.

Mulder NJ, Apweiler R, Attwood TK *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Research*, **33**, D201–D205.

Nelson DR (1999) Cytochrome P450 and the individuality of species. *Archives of Biochemistry and Biophysics*, **369**, 1–10.

van Nimwegen E (2003) Scaling laws in the functional content of genomes. *Trends in Genetics*, **19**, 479–484.

Ochman H, Moran NA (2001) Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science*, **292**, 1096–1099.

Ohno S (1970) *Evolution by Gene Duplication*. Springer Verlag, Berlin, Germany.

Pollack JR, Sorlie T, Perou CM *et al.* (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences, USA*, **99**, 12963–12968.

Porwollik S, Wong RM, McClelland M (2002) Evolutionary genomics of *Salmonella*: gene acquisitions revealed by microarray analysis. *Proceedings of the National Academy of Sciences, USA*, **99**, 8956–8961.

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics*, **16**, 276–277.

Rossler H, Rieck C, Delong T, Hoja U, Schweizer E (2003) Functional differentiation and selective inactivation of multiple *Saccharomyces cerevisiae* genes involved in very-long-chain fatty acid synthesis. *Molecular Genetics and Genomics*, **269**, 290–298.

Smith SE, Read DJ (1997) *Mycorrhizal Symbiosis*, 2nd edn. Academic Press, San Diego, California.

Swofford DL (1998) *PAUP: Phylogenetic Analysis Using Parsimony (and Other Methods), Version 4*. Sinauer Associates, Sunderland, Massachusetts.

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

Tunlid A, Talbot NJ (2002) Genomics of parasitic and symbiotic fungi. *Current Opinion in Microbiology*, **5**, 513–519.

Vozza A, Blanco E, Palmieri L, Palmieri F (2004) Identification of the mitochondrial GTP/GDP transporter in *Saccharomyces cerevisiae*. *Journal of Biological Chemistry*, **279**, 20850–20857.

Wilcoxon F (1945) Individual comparisons by ranking methods. *Biometrics*, **1**, 80–83.

Wolfinger RD, Gibson G, Wolfinger ED *et al.* (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, **8**, 625–637.

Wu L, Thompson DK, Li G *et al.* (2001) Development and evaluation of functional gene arrays for detection of selected genes in the environment. *Applied and Environmental Microbiology*, **67**, 5780–5790.

Yompakdee C, Ogawa N, Harashima S, Oshima Y (1996) A putative membrane protein, Pho88p, involved in inorganic phosphate transport in *Saccharomyces cerevisiae*. *Molecular and General Genetics*, **251**, 580–590.

The paper is one in a series of ongoing studies of the functional and evolutionary genomics of the ectomycorrhizal fungus *Paxillus involutus*. The work described in this study was part of the PhD programs of Antoine Le Quéré and Balaji Rajashekar. Kasper Astrup Eriksen with a PhD in Physics (University of Copenhagen), Andres Schützendübel with a PhD in Forest Botany (Georg-August-University) and Björn Canbäck with a PhD in Molecular Biology (Uppsala University) participated in this project as postdocs. Tomas Johansson is assistant professor at Lund University. His research focused on gene expression during the development of ectomycorrhizal association. Anders Tunlid is professor of microbial ecology, Lund University.
