Evolutionary genomics of symbiotic fungi

Rajashekar, Balaji

2007

Document Version:
Publisher's PDF, also known as Version of record

Evolutionary genomics of symbiotic fungi

Balaji Rajashekar

Doctoral Thesis, 2008

Lund University

Thesis for the degree of Doctor of Philosophy

Thesis Advisor : **Dr. Anders Tunlid**

Faculty Opponent : **Dr. David Liberles**, Department of Molecular Biology,
University of Wyoming, USA

To be presented, with the permission of the Faculty of Science of Lund University, for public criticism at Blå Hallen, Ekologihuset, Sölvegatan 37, Lund on Tuesday, the **28th March 2008, at 10:00**.

A doctoral thesis at a university in Sweden is produced either as a monograph or as a collection of papers. In the latter case, the introductory part constitutes the formal thesis, which summarizes the accompanying papers. These have either already been published or are manuscripts at various stages (in press, submitted or in manuscript).

Department of Ecology, Microbial Ecology,
Faculty of Science, Lund University,
SE-223 62 Lund, Sweden

ISBN 978-91-7105-267-4
Pages 128pp
Printed in Sweden by Tryckeriet i E-huset, Lund, 2007

To

Dad-Rajashekar, Mom-Sowbhagya

&

Wife-Bhanushri

Abstract

Ectomycorrhizae is a mutualistic association between roots of woody plants and a diverse range of soil fungi. The fungi exchange soil derived mineral nutrients for photosynthetic sugars from the host plant. The mycorrhizal symbioses are commonly found in all forest ecosystems and have a major ecological and economical importance. I have used comparative genomics, DNA microarrays and computational approaches to gain insights into the evolution of the ectomycorrhizal symbiosis in two fungi *Laccaria bicolor* (Basidiomycetes; Agaricales) and *Paxillus involutus* (Basidiomycetes; Boletales).

L. bicolor is the first symbiotic fungus to have its genome sequence determined. The genome assembly contains 65 million base pairs with about ~20,000 predicted protein-encoding genes. Here, I report the analysis of *L. bicolor* genome and its comparison with the genomes of four other basidiomycetes including the saprotrophic species *Coprinopsis cinerea* and *Phanerochaete chrysosporium*, the human pathogen *Cryptococcus neoformans* and the plant pathogen *Ustilago maydis*. The compared genomes cover about 550 million years of evolution. A total of 58,030 protein sequences from these five basidiomycetes were clustered into 7352 protein families. The evolution of protein families were analysed for accelerated rates of gain and loss along specific branches of a phylogenetic tree using a stochastic birth and death model. Analysis of the genome sequence of *L. bicolor* in comparison to other analysed basidiomycetes revealed large genome size, large number of protein families, larger size of protein families, many lineage specific and expanded families, and large number of recent duplicates. The evolution of two large and expanded protein families in *L. bicolor* having significant homology to protein kinases and Ras GTPases superfamilies were analysed in more detail. The analyses showed these families to contain many paralogs that have arisen through recent duplication events. The comparative analyses of gene families showed that the evolution of symbiosis in *L. bicolor* has been associated with the expansion of large multigene families. The functions of many of these families are unknown but many of them are differentially expressed during symbiosis.

In the second part of my thesis, I have analysed duplicated and rapidly evolving genes that could be associated with symbiotic adaptations in the ectomycorrhizal fungus *P. involutus*. Strains of *P. involutus* forming ectomycorrhiza showing various degree of host-specificity were analysed by comparative genomic hybridizations using a cDNA microarray representing 1076 putative unique genes. Approximately 17% of the genes investigated on the array were detected as rapidly and presumably non-neutrally evolving within *Paxillus*. Among these genes, there were several hydrophobins. Hydrophobins are small, secreted hydrophobic cell surface proteins having several roles in growth and development of fungi. The evolutionary mechanisms responsible for generating sequence and expression divergence among members of the hydrophobin multigene family in *P. involutus* were examined in more detail.

List of papers

This thesis is based on the following publications, which will be referred to in the text by their Roman numerals. The papers are appended at the end of the thesis.

I **The Genome of *Laccaria bicolor* Provides Insights into Mycorrhizal Symbiosis**.
F Martin, A Aerts, D Ahrén, A Brun, F Duchaussoy, J Gibon, A Kohler, E Lindquist, V Pereda, A Salamov, HJ Shapiro, J Wuyts, D Blaudez, M Buée, P Brokstein, B Canbäck, D Cohen, PE Courty, PM Coutinho, EGJ Danchin, C Delaruelle, JC Detter, A Deveau, S DiFazio, S Duplessis, L Fraissinet-Tachet, E Lucic, P Frey-Klett, C Fourrey, I Feussner, G Gay, J Grimwood, PJ Hoegger, P Jain, S Kilaru, J Labbé, YC Lin, V Legué, F Le Tacon, R Marmeisse, D Melayah, B Montanini, M Muratet, U Nehls, H Niculita-Hirzel, MP Oudot-Le Secq, M Peter, H Quesneville, **B Rajashekar,** M Reich, N Rouhier, J Schmutz, T Yin, M Chalot, B Henrissat, U Kües, S Lucas, Y Van de Peer, G Podila, A Polle, PJ Pukkila, PM Richardson, P Rouzé, IR Sanders, JE Stajich, A Tunlid, G Tuskan & IV Grigoriev.
Nature, March 6, 452(7183): 88-92 (2008). PMID: 18322534.

II **Expansion of Protein Families in the Symbiotic Fungus *Laccaria bicolor***.
Balaji Rajashekar, Annegret Kohler, Jason E. Stajich, Yao-Cheng Lin, Pierre Rouzé, Francis Martin, Anders Tunlid and Dag Ahrén.
(*Manuscript*).

III **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.
Molecular Ecology 15(2): 535-550 (2006). PMID: 16448419.

IV **Evolution of Nucleotide Sequences and Expression Patterns of Hydrophobin Genes in the Ectomycorrhizal Fungus *Paxillus involutus***.
Balaji Rajashekar, Peter Samson, Tomas Johansson & Anders Tunlid.
New Phytologist 174(2): 399-411 (2007). PMID: 17388902.

During my Ph.D period, I have also co-authored the following publications,

i **Paralysis of Nematodes : Shifts in Transcriptome of the Nematode-Trapping Fungus *Monacrosporium haptotylum* during Infection of *Caenorhabditis elegans***.
Csaba Fekete, Margareta Tholander, **Balaji Rajashekar,** Dag Ahrén, Eva Friman, Tomas Johansson and Anders Tunlid.
Environmental Microbiology, 10(2): 363-375 (2008). PMID: 18028414.

ii **Comparison of Gene Expression in Trap Cells and Vegetative Hyphae of the Nematophagous Fungus *Monacrosporium haptotylu*m**.
Dag Ahrén, Margareta Tholander, Csaba Fekete, **Balaji Rajashekar,** Eva Friman, Tomas Johansson and Anders Tunlid.
Microbiology 151(3): 789-803 (2005). PMID: 15758225.

iii **Divergence in Gene Expression Related to Variation in Host Specificity of an Ectomycorrhizal Fungus**.
Antoine Le Quéré, Andres Schützendübel, **Balaji Rajashekar,** Björn Canbäck, Jenny Hedh, Susanne Erland, Tomas Johansson and Anders Tunlid.
Molecular Ecology 13(12): 3809-3819 (2004). PMID: 15548293.

iv **Low Genetic Diversity among Isolates of the Nematode-Trapping Fungus *Duddingtonia flagrans*: Evidence for Recent Worldwide Dispersion from a Single Common Ancestor**.
Dag Ahrén, Margaret Faedo, **Balaji Rajashekar** & Anders Tunlid.
Mycological Research 108(10): 1205-1214 (2004). PMID: 15535071.

v **GalaxieEST: Addressing EST Identity through Automated Phylogenetic Analysis**.
R. Henrik Nilsson, **Balaji Rajashekar,** Karl-Henrik Larsson and Björn M. Ursing.
BMC Bioinformatics 5: 87 (2004). PMID: 15236648.

PMID refers to the PubMed Unique Identifier assigned to each PubMed citation.

Abbreviations

AM	Arbuscular mycorrhizae
bp	base pair
CAFE	Computational Analysis of gene Family Evolution
cDNA	complementary Deoxyribonucleic acid
CDS	The portion of a gene or an mRNA that codes for a protein
DNA	Deoxyribonucleic acid
EM	Ectomycorrhizae
EST	Expressed Sequence Tag
JGI	Joint Genome Institute
Mbp	Million base pair
mRNA	Messenger Ribonucleic acid

Contents

1. Introduction

Fungi comprise a highly versatile group of eukaryotic heterotrophic organisms that have diverse ecological and economical roles. They are mainly saprotrophs. They degrade dead organic matter and contribute to nutrient cycling in terrestrial ecosystem. A second group of fungi are pathogens which cause a large number of plant and animal diseases. In humans, fungal infection can range from common skin infection to life threatening respiratory infections. Plant diseases caused by fungal pathogens can lead to severe losses in agricultural crops. Common plant diseases caused by fungi include powdery mildews, rusts, smuts, damping off and rots of stem, leaf and roots. The enzymes secreted by fungi can damage wooden furniture, cotton fabrics, paints used on buildings and drugs. These and several other damages caused by the fungi can be estimated to several billions of dollars annually.

A third group of fungi are symbionts, and they form mutualistic relationships with plants and other organisms. Mycorrhizae is the symbiosis formed between fungi and plants (Smith & Read, 1997). This association is thought to have played an important role in the successful colonisation of land by plants about 400 million years ago (Pirozynski & Malloch, 1975; Pirozynski & Dalpe, 1989; Remy *et al.*, 1994). In the mycorrhizal association, the fungal partner benefits by utilizing photosynthetic sugars from the host plant and in return the plant receives essential mineral nutrients from the fungus. Other benefits for the plant in these associations are increased growth, increased photosynthetic rate, resistance to stress and drought, tolerance to adverse conditions like heavy-metal contamination and also reduced attack from root pathogens (Smith & Read, 1997).

It has been estimated that there are around 1.5 million species of fungi, and only 5% (~75,000) of these have been described (Deacon, 2006). These species belong to five phyla including Chrytridiomycota, Zygomycota, Glomeromycota, Ascomycota and Basidiomycota.

2. Objectives

The main objectives of this thesis have been to examine the genomic mechanisms that could be associated with:

(i) The evolution of ectomycorrhizal symbiosis in fungi

(ii) The variation in host preferences of ectomycorrhizal fungi

Basically, there are three compatible mechanisms that could account for the emergence of novel phenotypes and generation of phenotypic variations; namely variations in gene content, quantitative differences in gene expression, and structural differences in gene products (Ochman & Moran, 2001; Wray, 2007; Long *et al.*, 2003; Hughes, 1999).

In the thesis, I have mainly been studying how variations in gene contents, i.e. the gain and loss of genes, are associated with the evolution of ectomycorrhize. The major mechanism for gaining novel genes in eukaryotes is by duplications of genes or large chromosomal regions (Ohno, 1970). Duplications occur frequently in many genomes. Some of the new duplicates can evolve to obtain new functions, but a majority are silenced within a few million of years (Lynch & Conery, 2000). New functions can arise by regulatory mutations affecting the expression patterns, or by mutations in the coding regions affecting the structure of the gene products. The process of differential gain and loss of genes will result in gene families that share sequence and functional homology but differ in gene numbers.

Below, I will give a background to the ectomycorrhizal symbiosis, and how comparative genomics, coupled with evolutionary analysis, can provide insights into the genomic mechanisms associated with the emergence and pruning of the symbiotic lifestyle in fungi.

3. Background

3.1. Evolution of ectomycorrhizae

The earliest fossil record of mycorrhizae dates back to 400 million years ago (Pirozynski & Malloch, 1975; Pirozynski & Dalpe, 1989; Remy *et al.*, 1994). These associations are similar to the present *arbuscular mycorrhizae* (AM). AM associations are the most common and wide-spread type of mycorrhizae. About 80% of all higher land plant species, and about 90% of all plant families are capable of forming AM symbiosis (Brachmann & Parniske, 2006). The AM fungi are found in the Glomeromycota phylum, which diverged as a monophyletic group from the same common ancestor as the Ascomycota and Basidiomycota (Schüßler *et al.*, 2001). Fungi forming AM mycorrhizae are obligate symbionts. During infection, the fungal hyphae within the roots spread by forming linear or coiled hyphae, and then penetrate the root cortex and cell walls to form branching tree-like structures referred to as arbuscules (Smith & Read, 1997).

During the subsequent history of plants, other types of mycorrhizae evolved. The most common type is ectomycorrhizae (EM). There are ~10,000 fungal and 8,000 plant species capable of forming EM. The EM fungi primarily belong to Basidiomycota (95%) with some Ascomycota (4.8%) and few Zygomycota. About thirty families of plants form EM including pine, oaks, dipterocarps and eucalypts. Members of these and other EM forming species dominate many boreal and temperate forest ecosystems (Taylor & Alexander, 2005). The earliest fossil records of EM dates back to 50 million years ago (Lepage *et al.*, 1997). However, based on data on the evolutionary history of fungi and host species, it can be assumed that the EM symbiosis were extant well before this date, and even before 135 million years ago (Alexander, 2006).

EM fungi are facultative symbionts, they can grow both as biotrophs and saprotrophs. The saprotrophic ability differs between EM species. Some grow slowly in pure cultures and have a limited ability to utilize more complex organic material. Others can degrade and release nutrients from complex organic matter like litter and pollen (Read & Perez-Moreno, 2003). During infection of the plant roots, a specialized tissue is formed where the exchange of nutrients and carbon occurs. The tissue consists of a mantle, which develops from the fungal hyphae

surrounding the root, and a Hartig net, which is formed by the hyphae penetrating between the outer cells of the root. The carbon derived from the host supports the growth of an extensive fungal mycelium into the soil. The external mycelium assimilates and transports nutrients, mainly nitrogen, back to the plant host (Smith & Read, 1997).

The evolution of EM symbiosis has been studied in basidiomycetes using phylogenetic analysis (Bruns *et al.*, 2002; Hibbett *et al.*, 2000). These analyses have shown that mycorrhizal symbionts have evolved repeatedly from saprotrophic precursors. In the study by Hibbett *et al.* (2000) it was shown that EM homobasidiomycetes occur in six independent clades (a group of species that share a common ancestor). Furthermore, they observed occasional occurrence of reversals to free-living, saprotrophic lifestyles within otherwise mycorrhizal clades of fungi. Based on these observations, the authors suggested that EMs are unstable, evolutionary dynamic associations. Bruns and Shefferson (2004) showed that the loss of the EM habit proposed by Hibbett *et al.* is depending on the model used for the phylogentic analyses. They also argued that such losses are most unlikely because the regained free-living saprotroph will have major difficulties in competing with the more specialized saprotrophs (Bruns & Shefferson, 2004). The multiple, independent evolutionary origins of mycorrhizal basidiomycetes from saprotrophic ancestors have been confirmed in several recent studies (James *et al.*, 2006; Matheny *et al.*, 2006).

Many of the EM basidiomycetes have broad host specificity and can form mycorrhizal associations with several different tree species (Smith & Read, 1997). However, there is a considerable variation in the degree of host preferences exhibited by EM fungi, and they include both generalist and specialist species (Trappe, 1962). Studies have also shown that the variation of host preferences within species can be as large as the variation between species (Smith & Read, 1997; Cairney, 1999).

I have studied two different species of EM basidiomycetes,
Laccaria bicolor and *Paxillus involutus*. Both species belong to the Agaricomycotina. This clade contains about 20,000 described species, which constitutes to 68% of all known basidiomycetes (Hibbett, 2006). *L. bicolor* and *P. involutus* belongs to two different orders within the

Agaricomycotina, the Agaricales and the Boletales, respectively. These two orders form together with the Atheliales a monophyletic group, named the Agaricomycotidae (Hibbett, 2006). A recent multilocus phylogeny have shown that Agaricales consists of six major clades and 30 families (Matheny *et al.*, 2006). The EM habit appears to have evolved at least 11 times within the Agaricales, and there are no known reversals. *L. bicolor* belongs to the family Hydnangiaceae of the Agaricoid clade. This clade also contains a number of families including saprotrophs. The saprotroph family closest to Hydnangiaceae is the Psathyrellaceae which include *Coprinopsis cinerea* (Matheny *et al.*, 2006). The Boletales contains approximately a thousand described species and a majority of them are thought to form EM (Binder & Hibbett, 2006). Other species belong to the brown-rot fungi and their growth habit is thought to be an ancestral state for Boletales (Binder & Hibbett, 2006).

L. bicolor has been found associated with numerous plant species. This is the first symbiotic fungus genome to be sequenced with the collaborations of the Department of Energy's Joint Genome Institute (JGI) and *Laccaria* genome consortium (Paper I). This species is well studied due to its rapid development and establishment of EM with plant roots under laboratory conditions. Importantly, they are frequently used in forest nurseries to inoculate plant seedlings to improve plant growth. Comparative analysis of the *L. bicolor* genome with the genomes of other sequenced basidiomycetes, including the saprophytes *C. cinerea* (Agaricales, Agaricomycotina) and *Phanerochaete chrysosporium* (Corticales, Agaricomycotina), the human pathogen *Cryptococcus neoformans* (Tremellales, Agaricomycotina), and the plant pathogen *Ustilago maydis* (Ustilaginales, Ustilagomycotina) provided insights into the gain and loss of genes in *L. bicolor* (Paper II) (Table 1 and Figure 1).

Paxillus involutus is commonly found in Northern hemisphere and forms EM with large number of coniferous and deciduous tree species. *P. involutus* in association with various host plants including birch, pine and spruce, has been a popular model for studying the ecology and physiology of EM fungi (Wallander & Söderström, 1999). Of particular interest for my work is the fact that strains of *P. involutus* differs extensively in host preferences (Laiho, 1970; Gafur *et al.*, 2004; Le Quéré *et al.*, 2004).

Table 1. Genome characteristics of *Laccaria bicolor* and other basidiomycetes [a]

Genome characteristics	***Laccaria bicolor*** [b]	***Coprinopsis cinerea*** [c]	***Phanerochaete chrysosporium*** [d]	***Cryptococcus neoformans*** [e]	***Ustilago maydis*** [f]
Strain	S238N-H82	Okayama7#130	RP78	H99	521
Sequencing institution	JGI	Broad	JGI	Broad	Broad
Draft release	Version 1.0	Release 1	Version 2.0	Assembly 1, Gene set 3.0	Release 2
Genome assembly size (Mbp)	64.9	36.2-37.5	35.1	19.5	19.7
GC content (%)	46.6	51.6	53.2	48.2	54.0
Protein coding genes	20,614	13,544	10,048	7302	6522
CDS < 300 bp	2191	838	163	313	58
CDS > 3000 bp	961	977	633	638	1034
Gene density (gene / bp)	1 / 3147	1 / 2676	1 / 3498	1 / 2666	1 /3018
Average gene length (bp)	1533	1679	1667	1828	1935
Average CDS length (bp)	1134	1352	1366	1502	1840
Average exon length	210	251	232	253	1051
Average intron length	93	75	117	66	127

[a] This table is from the Supplementary material of Paper I.

[b] http://jgi.doe.gov/laccaria

[c] http://www.broad.mit.edu/annotation/genome/coprinus_cinereus/

[d] http://genome.jgi-psf.org/Phchr1/

[e] http://www.broad.mit.edu/annotation/genome/cryptococcus_neoformans/

[f] http://www.broad.mit.edu/annotation/genome/ustilago_maydis/

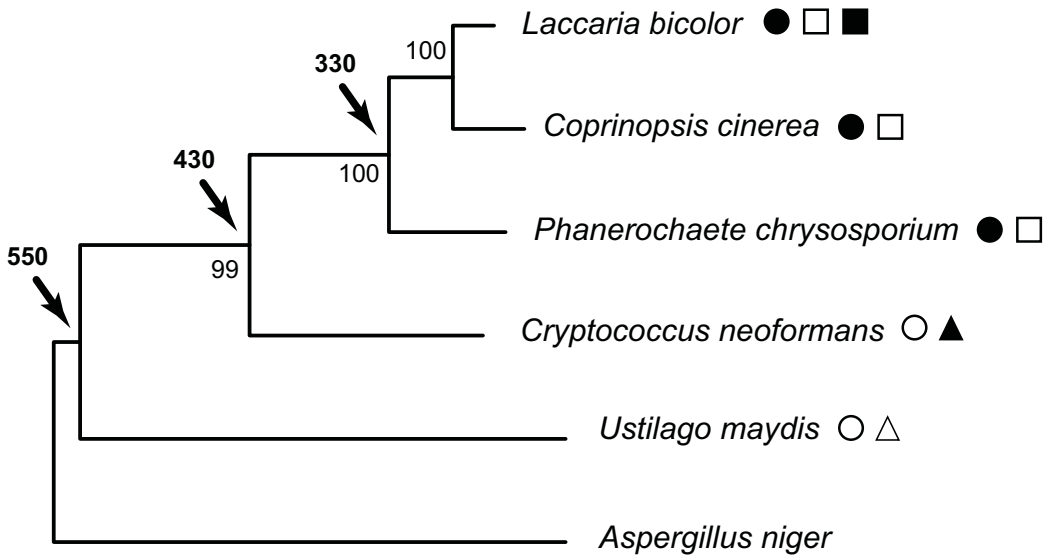


Figure 1. Phylogeny of basidiomycetes with sequenced genomes.
The Neighbor Joining tree (NJ) was constructed with 1000 bootstrap replications from 407,000 pairwise comparisons using a genome phylogeny approach (Kunin *et al.*, 2005). The branch lengths are proportional to the number of substitutions calculated from 18S rDNA sequences using Maximum Parsimony. The time of divergence of clades (Berbee, 2001), indicated in millions of years by arrows, was overlayed. The symbols correspond to filamentous fungi (●), yeast (○), mycorrhiza (■), plant pathogen (△), animal pathogen (▲), saprophyte (□). The saprotrophic filamentosus *A. niger* was used as outgroup. This figure is from Supplementary material of Paper I.

The *P. involutus* genome is predicted to contain ~23 Mbp and 7700 protein coding genes (Le Quéré *et al.*, 2002). The group of Microbial Ecology at Lund University has generated a large number of Expressed Sequence Tags (ESTs) (Johansson *et al.*, 2004). (ESTs are partial sequences of expressed genes derived from random sequences of cDNA libraries). Based on a non-redundant set of EST sequences, cDNA microarrays were constructed. The arrays have been used to examine the transcriptional responses in *P. involutus* during the infection and development of EM with birch seedlings (Johansson *et al.*, 2004; Wright *et al.*, 2005). The array was also used for comparing the genomes of *P. involutus* strains that differ in host preferences (Paper III). Among the genes that were rapidly evolving were several encoding for hydrophobins.

Hydrophobins are small, secreted proteins that are unique to fungi and play important roles in development and host interactions (Kershaw & Talbot, 1998; Wösten, 2001). In the last paper, the hydrophobin gene family of *P. involutus* was characterized in detail (Paper IV). Evolutionary mechanisms generating sequence and expression divergence among members in the family were examined.

3.2. Genomes and the evolution of phenotypic diversity

Gene and genome structure varies extensively between organisms both within species and across the tree of life. There are four major evolutionary forces that are responsible for this variation - natural selection, mutations, recombination and genetic drift (Lynch, 2007). Natural selection, the differential survival of individuals in a population will lead to changes in gene frequencies. Phenotypic characters that are favoured by the process of selection are called adaptations. The other three forces affecting the variation of genes and genomes are non-adaptive, thus they are not related to the fitness of the organism. Mutations, including gene duplications and other genetic changes, are the ultimate source of variation on which natural selection acts to modify or produce novel characters. Recombination affects the patterns of variation within and among chromosomes. Genetic drift is the random fluctuations of gene frequencies occurring in populations.

During the last four decades, it has been shown that a majority of mutations are neutral or nearly "neutral" and only a minority of them will affect the function of proteins (Kimura, 1983; Ohta, 1992; Nei, 2007). The relative contribution of selection and the random processes of genetic drift and mutations for generating adaptations and phenotypic complexity have been a matter of controversy. Some evolutionary biologists have stated that neutral mutations is of limited interest for understanding phenotypic evolution because they do not affect the function of proteins and phenotypic characters (Futuyma, 2005). Other biologists have argued that considerable portion of phenotypic evolution is caused by neutral or nearly neutral mutations (Nei, 1987). Lynch and co-workers have recently proposed that the non-adaptive forces have played a major role in generating complex genomic architectures, novel gene structures, and regulatory pathways that drive patterns of gene expression, as well as increase in gene content

in eukaryotes. Many of these modifications have emerged passively in response to long-term population changes. These changes provided the substrate for the secondary evolution of phenotypic adaptations by natural selection (Lynch & Conery, 2003; Lynch, 2006; Lynch, 2007).

Comparative genomics have become an important tool for examining genomic variations that may be related to the emergence of novel phenotypes and adaptations to specific environments. By comparing genome sequences encoding alternative phenotypes it is possible to reconstruct some of the genomic changes that have accompanied the change in the phenotype. The observed pattern is then contrasted with what is expected in the absence of natural selection; that is due to random processes. The remaining variation potentially indicates the action of natural selection.

There are several methods for testing the action of natural selection on nucleotide sequences. For example, the neutral theory of molecular evolution predicts that as long as a protein's function is unaltered, the proteins rate of evolution is constant. Accordingly, a detection of change in the rate of evolution of a protein might reveal changes due to selection (Hughes, 1999). This was the basis for detecting rapidly and non-randomly evolving genes in Paper III. In Paper IV, selection on nucleotide sequences was examined in hydrophobin encoding sequences by comparing the rates of nonsynonymous (d_N) and synonymous (d_S) substitutions per site. A d_N/d_S ratio significantly larger than one is indicative of positive selection, and a d_N/d_S ratio smaller than 1 of purifying selection (Hughes, 1999).

In contrast to the well-accepted methods for detecting adaptations in sequences, there are no such methods for testing the action of selection on gene family sizes. Recently, a tool called CAFE (Computational Analysis of gene Family Evolution) was developed for the statistical analysis of the evolution of the size of gene families (De Bie *et al.*, 2006). The method uses a stochastic birth and death process to identify unexpectedly large size changes in expansions or contractions over a phylogeny. I have used this method to examine the mode of gene family evolution in basidiomycetes (Paper II). The CAFE tool has previously been used to analyse gene family evolution

in *Saccharomyces* species (Hahn *et al.*, 2005) and mammalian species (Demuth *et al.*, 2006; Gibbs *et al.*, 2007; Sawatzki & Cooper, 2007).

Comparisons of closely related strains or species are particularly informative for identifying genomic mechanisms of adaptive evolution because they hold constant to all variables shared by congeners (Harvey & Pagel, 1991). However, complete genome sequence data are seldom available for closely related strains and species. As an alternative, comparative genomic hybridization analysis on DNA microarrays can be used to screen for gene gains and losses, and genes with divergent sequences (Dunham *et al.*, 2002; Kim *et al.*, 2002). In Paper III, a cDNA microarray was used to screen for variation in gene contents and rapidly evolving genes among strains of *P. involutus.*

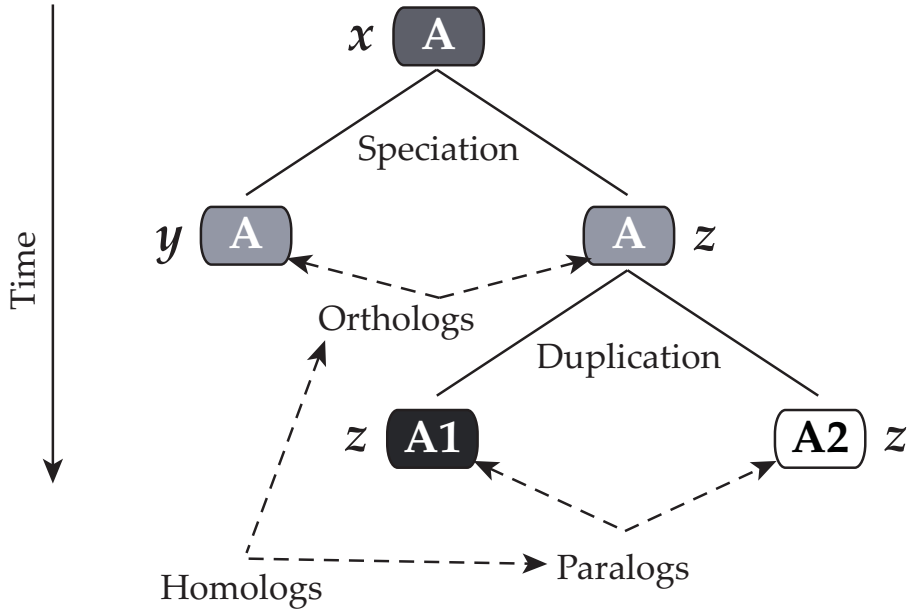


Figure 2. A diagrammatic representation of homologs, orthologs and paralogs. Gene relationships after a speciation event (in ancestor *x*) creates new species *y* and *z*. Gene **A** is an ortholog between species *y* and *z*. Gene duplications of gene **A** in species *z* creates paralogs **A1** and **A2**. Gene **A** in species *y* now has two orthologs **A1** and **A2** in species *z*. (*Homologs* are genes sharing a common origin. *Homologous* genes are of two types - *Orthologs* are genes originating from a single ancestral gene in the common ancestor and separated by speciation event, and *Paralogs* are genes evolved by a duplication event).

3.3. Gene duplications

Duplications of genes or large chromosomal regions in combinations with mutations that causes functional divergence of the duplicates are thought to have played an important role in the generation of novel phenotypes deep in evolutionary trees as well as the diversification of more closely related species (Ohno, 1970; Hurles, 2004).

Analysis of genome sequences have shown that gene duplications arise at very high rates (Lynch & Conery, 2000). Duplication events are detected in a genome as very similar homologs formed before speciation (orthologs) or after speciation (paralogs) (Fitch, 1970; Koonin, 2005) (Figure 2). These two subclasses were first defined by Fitch (Fitch, 1970) to describe the distinct evolutionary relationships between genes. The orthologs are assumed to have similar functions, whereas paralogs can have diverged functions. Classifying and identifying orthologs and paralogs in complete genome sequences have been one of the central problems in genome comparisons. They have become increasingly important to describe the history of speciation and gene duplication events. Some of the mechanisms by which genes become duplicated include tandem duplications, segmental duplications of large chromosomal regions and whole genome duplications (Hurles, 2004).

After duplication, the fate of duplicate genes could be one of the following: (a) functional redundancy, the duplicated copy retains the same function of the parent gene, this is often beneficial to produce extra amounts of proteins or RNA (Zhang, 2003);
(b) pseudogenization (nonfunctionalization), one of the duplicate accumulate deleterious mutations and become a pseudogene, which is either unexpressed or become functionless (Lynch & Conery, 2000);
(c) subfunctionalization, both duplicates experience degenerative mutations and divide the ancestral function between them (Force *et al.*, 1999; Lynch & Force, 2000); and
(d) neofunctionalization, the duplicate develops a novel function while the other copy always retains the ancestral function (Lynch *et al.*, 2001).
The role of these mechanisms for preserving duplicates of the hydrophobin genes are discussed in Paper IV.

Duplication events will result in the evolution of multigene families (Prince & Pickett, 2002). Such families contain groups of homologous protein sequences that are evolutionarily related which are created by successive gene duplication and speciation events, and have descended from a common ancestor. Several methods have been developed to cluster protein sequences into families (Liu & Rost, 2003; Krause *et al.*, 2005; Kelil *et al.*, 2007). TRIBEMCL (Enright *et al.*, 2002), based on a Markov clustering approach and BlastClust (ftp://ftp.ncbi.nih.gov/blast/), which uses score-based single-linkage clustering are some of the widely used algorithms for clustering protein sequences. TRIBEMCL is a fast algorithm which can be applied for large datatsets and has been reported to handle multidomain proteins and partial sequences. We applied this algorithm to generate protein families from five basidiomycete protein sequences. The protein families were constructed based on the significant pair-wise sequence similarities (Paper II).

The function of a gene has two distinct components, what its product does and the circumstances under which that product is expressed (Wray, 2007). The first component is modified by mutations in the coding parts of the DNA where as the second component is affected by mutations in regions of the DNA that regulates transcription. Transcription is regulated by sequences in so-called *cis*-regulatory regions. Typically such segments are located adjacent to the 5′ region of the transcriptional start site, but they can also be found at the 3′region of the gene, as well as in introns or further away from the gene they regulate. There are number of evidences that *cis*-regulatory mutations underlie a number of ecologically significant changes in phenotypic differences in morphology, physiology and behaviour (Wray, 2007). I have examined variation in expression levels of gene duplicates using data obtained from DNA microarray experiments (Paper II and IV).

3.4. Phylogenetic analysis

Phylogenetic methods are important in comparative genomics for revealing the evolutionary history of organisms, genes and genomes, the events and mechanisms of gene duplications, identifying homologs and for providing information for functional classifications of genes and proteins. A majority of phylogenetic studies are based on the analysis of molecular data. Several methods have been developed for constructing and evaluating trees obtained from such data (Nei, 1996). Statistical tests are of importance for understanding if the generated tree hierarchy is just by a chance or instead of a genealogical process. The methods that I have used in this thesis are maximum likelihood, neighbor-joining and split network.

In genome comparisons it is of major importance to provide a comprehensive functional classification for all proteins. An important tool for providing such classification is to cluster proteins to families. The protein families will contain conserved proteins having a similar function. In many protein families, members could have undergone numerous gene duplication events leading to functional diversification and the formation of sub-families. Occasionally, large families or superfamilies will arise consisting of a many sub-families in hierarchical clusters. I have analyzed two such superfamilies protein kinases and ras GTPases using several methods like sequence clustering, graph algorithms, tools for visualizations, identifications and annotations of functional domains. More detailed phylogenetic analysis was done to identify clusters of orthologs and lineage specific paralogs (Paper II).

In Paper IV, phylogenetic analysis was used to understand evolutionary mechanisms generating sequence and expression divergence of hydrophobins in *P. involutus*. Gain and loss of hydrophobin genes were demonstrated and variations in expression levels were related to the evolutionary history of the duplicates.

4. Summary of thesis

My thesis consists of two parts. In the first part, I used comparative genomics to reveal changes in the genomes that have accompanied the evolution of symbiosis in *Laccaria bicolor*. In the second part, I have analysed the genomic diversity in strains of the symbiotic fungi *Paxillus involutus* that differ in host preferences.

Laccaria bicolor

The draft sequence of the ectomycorrhizal fungus *L. bicolor* was publicly released in July 2006. The sequencing was a joint collaboration between JGI and the *Laccaria* Genome Consortium. A total of 68 collaborators from different laboratories have made contribution to this project, with Francis Martin from INRA, Nancy being the main coordinator. I, together with other scientists at the Department of Microbial Ecology in Lund, have made contributions to this project in sequence analysis; genome phylogeny; gene duplications; age distribution of duplicates; construction, comparisons and analysis of protein families; comparison of protein family sizes; evolution of protein family sizes and domain phylogeny.

A major highlight from Paper I is the finding that *L. bicolor* has the largest fungal genome sequenced yet. The genome size was estimated to be 65 Mbp and the assembly contains 20,614 predicted protein-encoding genes. In comparison to other fungal genomes *L. bicolor* contained large number of transposable elements; large percentage of proteins belonging to multigene families; large number of protein families (Figure 3); large expansions in family sizes in lineage specific protein families. Many transcripts were present in expanded and lineage specific multigene families and were being up-regulated in symbiotic and fruiting body tissues, suggesting a role in tissue differentiation. The *L. bicolor* genome contained a battery of genes encoding small secreted proteins (SSPs) with unknown function. Some of them were expressed during infection of host plant roots. A few of the SSPs displayed sequence similarity to so-called effector proteins of biotrophic plant pathogens. Despite the fact that the *L. bicolor* genome contains many genes encoding for hydrolytic enzymes such as proteases and lipases, it was lacking carbohydrate active enzymes involved in degrading plant cell walls.

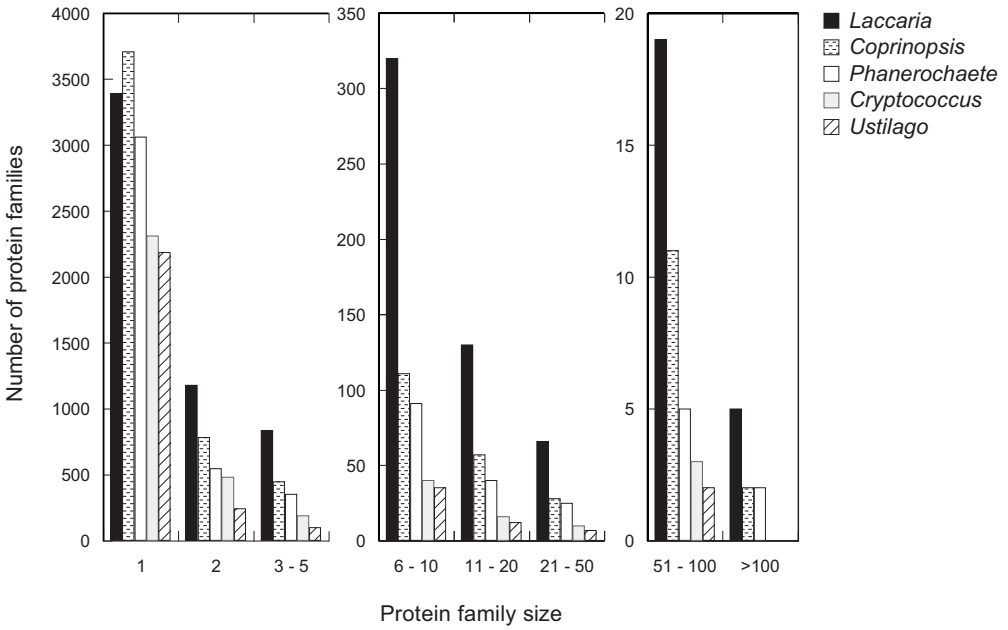


Figure 3. Comparison of protein family sizes in basidiomycetes.
Protein sequences predicted from the genome sequences of *Laccaria bicolor*, *Coprinopsis cinerea*, *Phanerochaete chrysosporium*, *Cryptococcus neoformans* and *Ustilago maydis* were clustered into families using the TRIBEMCL algorithm (Enright *et al.*, 2002). In total 7352 protein families (containing at least two sequences) were identified.

In Paper II the expansion of gene families in *L. bicolor* was analysed in detail by comparing gene content and protein families in the five basidiomycetes that have been sequenced – the symbiont *L. bicolor*, the saprophytes *C. cinerea* and *P. chrysosporium*, the animal pathogen *C. neoformans,* and the plant pathogen *U. maydis* (Figure 4). In comparison to the other basidiomycetes, *L. bicolor* contained a high number of gene duplicates. A low level of sequence divergence between many of these duplicates suggests that they are of recent origin. A large fraction of these young duplicates were found in 55 large gene families containing more than 25 members. A majority of these families did not contain any known protein domain and the functions of them are not known. However, analysis of microarray data showed that they contain members that are differentially expressed in mycorrhizal root tips as compared to fruiting bodies and mycelia. Two of the large families displayed sequence similarities to proteins involved in signal transduction including protein kinases and Ras GTPases. Based on phylogenetic analysis it was shown that both families contain clusters of paralogs that have arisen in the *L. bicolor* lineage following the split from the saprophyte *C. cinerea*.

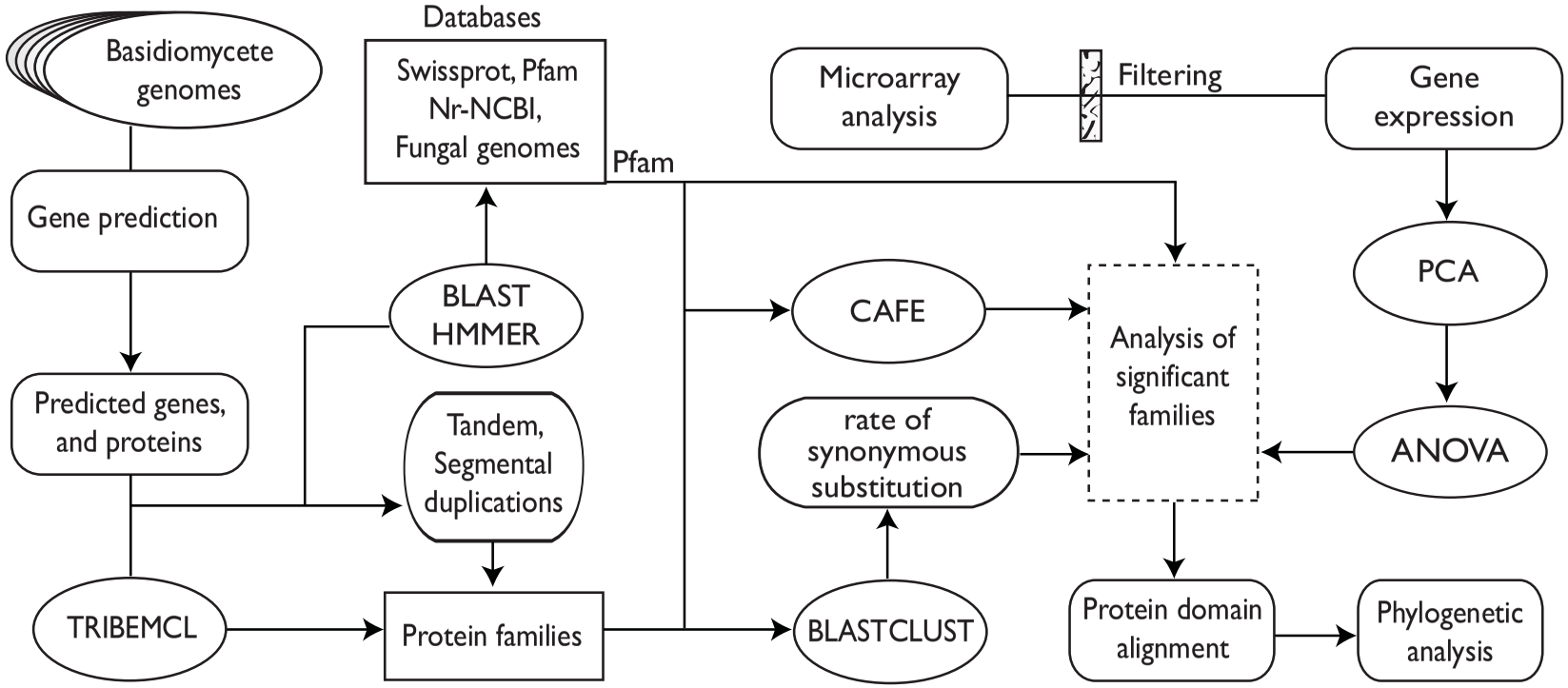


Figure 4. Schematic representation of the various steps involved in comparative analyses in basidiomycete genomes.
The bioinformatics tools, molecular databases, statistical software's used in the analysis are listed in Appendix A. The comparative analyses have been discussed in Paper II.

Paxillus involutus

A major research area at the Department of Microbial Ecology in Lund during the recent years has been to study the molecular background and evolution of the interaction between the EM fungus *P. involutus* and birch (*Betula pendula*). Due to a joint effort by several scientists, post docs and Ph.D students a large number of ESTs have been generated and cDNA microarrays have been constructed and used for global transcriptional analyses and comparative genomic hybridizations.

One of the areas of research has been to identify the genomic mechanisms underlying the variation in host preferences observed between strains of *P. involutus*. In Paper III, a screen for identifying rapidly evolving genes in different strains of this fungus by cDNA microarray hybridizations was developed. The array used contained reporters for 1076 putative unique genes in *P. involutus*. About 17% of these genes were identified as rapidly diverging and 6% of these varied in copy numbers between the strains analysed. Most of the duplicates were orphans, membrane proteins, or related to stress/defence reactions.

Among the genes detected as rapidly and presumably non-neutrally evolving within *Paxillus*, there were several encoding hydrophobins. Hydrophobins are small, secreted hydrophobic cell surface proteins having several roles in growth and development of fungi. I together with another Ph.D student, Peter Samson who did the sequencing of hydrophobin genes in *Paxillus* strains and species, have examined the evolutionary mechanisms responsible for generating sequence and expression divergence among members of the hydrophobin multigene family in *P. involutus* (Paper IV). Sequences were analysed using phylogenetic methods and expression levels were inferred using data from microarray experiments. Seven hydrophobin genes and one pseudogene were identified among the *P. involutus* strain; the young (recent) duplicates of hydrophobin genes showed positive selection; no correlation between sequence divergence and expression divergence; three hydrophobin genes showed shift in expression levels and finally it was shown that the hydrophobin genes evolve by birth-and-death model.

5. Future perspectives

In many organisms, non-random codon usage in synonymous codons has been observed. The biased codon usage is considered to be a result of natural selection or mutational biases. The correlations of codon bias with tRNA abundance, gene expression level, and guanine-cytosine content at synonymous codons are taken as evidences for natural selection or mutational biases. Highly expressed genes are commonly biased in codon usage. Hydrophobin genes are among the genes in *P. involutus* and *L. bicolor* that have the most biased codon usage. The bias is not correlated to gene expression levels and alternative explanations need to be investigated.

An alternative strategy to the comparative analysis of gene families performed in Paper I and II for identifying proteins that distinguish symbiotic fungi from saprophytic and parasitic fungi is phylogenetic profiling. This approach involves the study of the occurrence of proteins or protein families across fully sequenced genomes. The method creates a phylogenetic profile of all proteins across organisms, to identify presence or absence of homologs, to measure evolutionary distance, and to identify proteins within the same pathway or protein complex. By comparing the phylogenetic profiles to phenotypic profiles it is possible to infer the proteins partially responsible for establishing a phenotype. Presently, the *L. bicolor* proteins can be compared with more than 50 available fungal genomes.

Phylogenetic analyses have shown that the EM growth habit has evolved repeatedly, in many lineages of the Basidiomycetes. Recently, JGI has funded sequencing of the genome of *P. involutus* (coordinated jointly by A. Tunlid and F. Martin). *P. involutus* and *L. bicolor* belong to two different evolutionary lineages of Basidiomycetes, Boletales and Agaricales. Their genomes appear to be very different in size. However, their morphological responses in association with the roots of host plant are similar. Comparative genome analysis of the changes among species within the Boletales and Agaricales will provide insights into both common and unique mechanisms underlying the evolution of mutualism in these two lineages of Basidiomycetes.

6. Acknowledgements

I would like to thank everyone who have helped and encouraged me in different stages of my personal life and in my research career.

Ecology, I would like to sincerely thank my supervisor *Anders Tunlid* for inviting me to *Sweden* and for giving me an opportunity to be a part of his research group. It has been great pleasure to have you as my supervisor. I admire your sharp scientific thinking, ideas and critical evaluation. Thank you for your inspiration and for giving me many challenging opportunities. I am grateful to you, for all your help during my research, thesis writing and stay in *Sweden*. Thanks to *Tomas* for helping me to settle down on my first day in Lund, for your guidance in Microarray analysis and for correcting my thesis and manuscript. I am grateful to my roommates, *Björn* for helping me in shell scripting and phylogeny; and my second roommate *Dag*, you have extended your help without asking, you came out with some amazing ideas and were never short of suggestions, thanks for all your critical inputs and for a great collaboration during the *Laccaria* project; it has been mainly because of you both I could improve my Bioinformatics skills. I would like to thank all my current and past colleagues who have contributed in my work and for all the good times at the department. *Fredrik* & *Louise*, I will definitely miss talking with you both, you've been very caring and good friends; *Peter*, for your collaboration in *hydrophobin* project and for your company in travel and other activities; *Antoine*, for inspiring me in the initial stages of my research; *Csaba*, for the discussions on sequencing; *Erland*, thanks for helping me in statistics. *Eva*, *Derek*, *Jenny*, *Johannes*, *Edith*, *Lars-Ola*, *Margareta*, *Bengt*, *Håken*, *Pål-Axel*, *Susanne* – for extending help when ever I came looking for you. *Annette*, without you I could not have done some of the administration work. I thank you all for everything and I will surely miss all of you. I would also like to thank all the course organizers, librarians and other fellow researchers in the ecology building.

Co-authors, I would like to express my deepest gratitude to all my co-authors involved with me in several research projects. I would like to specially thank *Francis Martin* for the *Laccaria* genome project and seeing it through to the Nature.

Research School, special thanks to *Anders Blomberg* for allowing me to be a part of Research school of Genomics and Bioinformatics, and providing a platform to network with other students, the open days and summer schools had been very educative, helpful and worthy.

Sponsors, I would like to thank the following who have provided financial assistance during my Ph.D studies - the Swedish Research Council and the Research School in Genomics & Bioinformatics, *Göteborg* for funding my research; the ISMB/ECCB and NBN for travel expenses; and Kungliga Fysiografiska Sällskapet for funding my laptop.

Thesis, *Johan Cedervall***,** my printer thank you for being very helpful during my thesis printing and cover design.

Lund, I am grateful to all my friends in *Lund* for all the good times and for the parties, picnics and especially cricket. *Srini*, *Viji* you both have been a family for us in *Lund*. *Eva* for taking care of us and providing help when ever we had problems in the apartment. I will fondly remember you all forever.

Teachers, I would also like to acknowledge all my teachers, researchers and professors at school (KECS and KPHS, Bangalore), college (St. Joseph's, Bangalore), Bioinformatics (Pune) and IISc (Bangalore).

Friends, all these years these long distances have not broken our friendship, thank you all for being such great friends.

Family, I am grateful to all my family members for being very supportive and encouraging. My *mom (Sowbhagya)* and *dad* (*Rajashekar)*, what ever I have achieved is because of your support and love, I hope your sacrifices have been fruitful. My *sister* (*Suma)* & *brother-in-law* (*Muarlidhar)*, you both have been very caring and supportive as always, my *nephew* (*Tannay)* who has given me joy and his energy has always reminded me of my childhood. My *grandparents*, I miss you. My *guru*, I wish you had seen this day, thank you for helping me to understand the truthful realities of life. How can I forget my *wife (Bhanushri)*, you've been very supportive, patient, encouraging, and because of you my research and stay have been extremely successful and happy.

Balaji Rajashekar

21 November 2007, Lund

7. References

Alexander IJ. **2006.** Ectomycorrhizas--out of Africa? *New Phytol.* **172:** 589-591.

Berbee ML. **2001.** The phylogeny of plant and animal pathogens in the Ascomycota. *Physiological and Molecular Plant Pathology* **59:** 165-187.

Binder M, Hibbett DS. **2006.** Molecular systematics and biological diversification of Boletales. *Mycologia* **98:** 971-981.

Brachmann A, Parniske M. **2006.** The most widespread symbiosis on earth. *Plos Biology* **4:** 1111-1112.

Bruns TD, Bidartondo MI, Taylor DL. **2002.** Host specificity in ectomycorrhizal communities: What do the exceptions tell us? *Integrative and Comparative Biology* **42:** 352-359.

Bruns TD, Shefferson RP. **2004.** Evolutionary studies of ectomycorrhizal fungi: recent advances and future directions. *Canadian Journal of Botany-Revue Canadienne de Botanique* **82:** 1122-1132.

Cairney JWG. **1999.** Intraspecific physiological variation: implications for understanding functional diversity in ectomycorrhizal fungi. *Mycorrhiza* **9:** 125-135.

De Bie T, Cristianini N, Demuth JP, Hahn MW. **2006.** CAFE: a computational tool for the study of gene family evolution. *Bioinformatics.* **22:** 1269-1271.

Deacon JW. **2006.** Introduction: the fungi and fungal activities. In:*Fungal Biology*. Blackwell Publishing Ltd, 1.

Demuth JP, De BT, Stajich JE, Cristianini N, Hahn MW. **2006.** The evolution of mammalian gene families. *PLoS ONE.* **1:** e85.

Dunham MJ, Badrane H, Ferea T, Adams J, Brown PO, Rosenzweig F, Botstein D. **2002.** Characteristic genome rearrangements in experimental evolution of *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the United States of America* **99:** 16144-16149.

Enright AJ, Van DS, Ouzounis CA. **2002.** An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* **30:** 1575-1584.

Fitch WM. **1970.** Distinguishing homologous from analogous proteins. *Syst.Zool.* **19:** 99-113.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. **1999.** Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531-1545.

Futuyma DJ. **2005.** *Evolution*. Sinauer Associates.

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 x canescens*) differ in H_2O_2 production. *Plant Biology* **6:** 91-99.

Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK, Batzer MA, Bustamante CD, Eichler EE, Hahn MW, Hardison RC, Makova KD, Miller W, Milosavljevic A, Palermo RE, Siepel A, Sikela JM, Attaway T, Bell S, Bernard KE, Buhay CJ, Chandrabose MN, Dao M, Davis C, Delehaunty KD, Ding Y, Dinh HH, Dugan-Rocha S, Fulton LA, Gabisi RA, Garner TT, Godfrey J, Hawes AC, Hernandez J, Hines S, Holder M, Hume J, Jhangiani SN, Joshi V, Khan ZM, Kirkness EF, Cree A, Fowler RG, Lee S, Lewis LR, Li ZW, Liu YS, Moore SM, Muzny D, Nazareth LV, Ngo DN, Okwuonu GO, Pai G, Parker D, Paul HA, Pfannkoch C, Pohl CS, Rogers YH, Ruiz SJ, Sabo A, Santibanez J, Schneider BW, Smith SM, Sodergren E, Svatek AF, Utterback TR, Vattathil S, Warren W, White CS, Chinwalla AT, Feng Y, Halpern AL, Hillier LW, Huang XQ, Minx P, Nelson JO, Pepin KH, Qin X, Sutton GG, Venter E, Walenz BP, Wallis JW, Worley KC, Yang SP, Jones SM, Marra MA, Rocchi M, Schein JE, Baertsch R, Clarke L, Csuros M, Glasscock J, Harris RA, Haviak P, Jackson AR, Jiang HY, Liu Y, Messina DN, Shen YF, Song HXZ, Wylie T, Zhang L, Birney E, Han K, Konkel MK, Lee JN, Smit AFA, Ullmer B, Wang H, Xing J, Burhans R, Cheng Z, Karro JE, Ma J, Raney B, She XW, Cox MJ, Demuth JP, Dumas LJ, Han SG, Hopkins J, Karimpour-Fard A, Kim YH, Pollack JR, Vinar T, ddo-Quaye C, Degenhardt J, Denby A, Hubisz MJ, Indap A, Kosiol C, Lahn BT, Lawson HA, Marklein A, Nielsen R, Vallender EJ, Clark AG, Ferguson B, Hernandez RD, Hirani K, Kehrer-Sawatzki H, Kolb J, Patil S, Pu LL, Ren Y, Smith DG, Wheeler DA, Schenck I, Ball EV, Chen R, Cooper DN, Giardine B, Hsu F, Kent WJ, Lesk A, Nelson DL, O'Brien WE, Prufer K, Stenson PD, Wallace JC, Ke H, Liu XM, Wang P, Xiang AP, Yang F, Barber GP, Haussler D, Karolchik D, Kern AD, Kuhn RM, Smith KE, Zwieg AS. **2007.** Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316:** 222-234.

Hahn MW, De BT, Stajich JE, Nguyen C, Cristianini N. **2005.** Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res.* **15:** 1153-1160.

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

Hibbett DS. **2006.** A phylogenetic overview of the Agaricomycotina. *Mycologia* **98:** 917-925.

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

Hughes AL. **1999.** *Adaptive Evolution of Genes and Genomes*. Oxford: Oxford University Press.

Hurles M. **2004.** Gene duplication: the genomic trade in spare parts. *PLoS.Biol.* **2:** E206.

James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, Cox CJ, Celio G, Gueidan C, Fraker E, Miadlikowska J, Lumbsch HT, Rauhut A, Reeb V, Arnold AE, Amtoft A, Stajich JE, Hosaka K, Sung GH, Johnson D, O'Rourke B, Crockett M, Binder M, Curtis JM, Slot JC, Wang Z, Wilson AW, Schussler A, Longcore JE, O'Donnell K, Mozley-Standridge S, Porter D, Letcher PM, Powell MJ, Taylor JW, White MM, Griffith GW, Davies DR, Humber RA, Morton JB, Sugiyama J, Rossman AY, Rogers JD, Pfister DH, Hewitt D, Hansen K, Hambleton S, Shoemaker RA, Kohlmeyer J, Volkmann-Kohlmeyer B, Spotts RA, Serdani M, Crous PW, Hughes KW, Matsuura K, Langer E, Langer G, Untereiner WA, Lucking R, Budel B, Geiser DM, Aptroot A, Diederich P, Schmitt I, Schultz M, Yahr R, Hibbett DS, Lutzoni F, McLaughlin DJ, Spatafora JW, Vilgalys R. **2006.** Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature* **443:** 818-822.

Johansson T, Le Quéré A, Ahrén D, Söderström B, Erlandsson R, Lundeberg J, Uhlen M, Tunlid A. **2004.** Transcriptional responses of *Paxillus involutus* and *Betula pendula* during formation of ectomycorrhizal root tissue. *Mol.Plant Microbe Interact.* **17:** 202-215.

Kelil A, Wang S, Brzezinski R, Fleury A. **2007.** CLUSS: clustering of protein sequences based on a new similarity measure. *BMC Bioinformatics* **8:** 286.

Kershaw MJ, Talbot NJ. **1998.** Hydrophobins and repellents: proteins with fundamental roles in fungal morphogenesis. *Fungal Genet.Biol.* **23:** 18-33.

Kim CC, Joyce EA, Chan K, Falkow S. **2002.** Improved analytical methods for microarray-based genome-composition analysis. *Genome Biology* **3:** 65.1-65.17.

Kimura M. **1983.** *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

Koonin EV. **2005.** Orthologs, paralogs, and evolutionary genomics. *Annu.Rev.Genet.* **39:** 309-338.

Krause A, Stoye J, Vingron M. **2005.** Large scale hierarchical clustering of protein sequences. *BMC Bioinformatics* **6:** 15.

Kunin V, Ahrén D, Goldovsky L, Janssen P, Ouzounis CA. **2005.** Measuring genome conservation across taxa: divided strains and united kingdoms. *Nucleic Acids Res.* **33:** 616-621.

Laiho O. **1970.** *Paxillus involutus* as a mycorrhizal symbiont of forest trees. *Acta Forestalia Fennica* **106:** 1-72.

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, Canbäck B, Hedh J, Erland S, Johansson T, Tunlid A. **2004.** Divergence in gene expression related to variation in host specificity of an ectomycorrhizal fungus. *Mol.Ecol.* **13:** 3809-3819.

Lepage BA, Currah RS, Stockey RA, Rothwell GW. **1997.** Fossil ectomycorrhizae from the middle Eocene. *American Journal of Botany* **84:** 410-412.

Liu J, Rost B. **2003.** Domains, motifs and clusters in the protein universe. *Curr.Opin.Chem.Biol.* **7:** 5-11.

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

Lynch M. **2006.** The origins of eukaryotic gene structure. *Molecular Biology and Evolution* **23:** 450-468.

Lynch M. **2007.** The evolution of genetic networks by non-adaptive processes. *Nature Reviews Genetics* **8:** 803-813.

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

Lynch M, Conery JS. **2003.** The origins of genome complexity. *Science* **302:** 1401-1404.

Lynch M, Force A. **2000.** The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154:** 459-473.

Lynch M, O'Hely M, Walsh B, Force A. **2001.** The probability of preservation of a newly arisen gene duplicate. *Genetics* **159:** 1789-1804.

Matheny PB, Curtis JM, Hofstetter V, Aime MC, Moncalvo JM, Ge ZW, Yang ZL, Slot JC, Ammirati JF, Baroni TJ, Bougher NL, Hughes KW, Lodge DJ, Kerrigan RW, Seidl MT, Aanen DK, DeNitis M, Daniele GM, Desjardin DE, Kropp BR, Norvell LL, Parker A, Vellinga EC, Vilgalys R, Hibbett DS. **2006.** Major clades of Agaricales: a multilocus phylogenetic overview. *Mycologia* **98:** 982-995.

Nei M. **1987.** *Molecular Evolutionary Genetics*. Columbia University Press.

Nei M. **1996.** Phylogenetic analysis in molecular evolutionary genetics. *Annu.Rev.Genet.* **30:** 371-403.

Nei M. **2007.** The new mutation theory of phenotypic evolution. *Proceedings of the National Academy of Sciences of the United States of America* **104:** 12235-12242.

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*. Berlin: Springer Verlag.

Ohta T. **1992.** The Nearly Neutral Theory of Molecular Evolution. *Annual Review of Ecology and Systematics* **23:** 263-286.

Pirozynski KA, Dalpe Y. **1989.** Geological History of the Glomaceae with Particular Reference to Mycorrhizal Symbiosis. *Symbiosis* **7:** 1-36.

Pirozynski KA, Malloch DW. **1975.** The origin of land plants: a matter of mycotrophism. *Biosystems* **6:** 153-164.

Prince VE, Pickett FB. **2002.** Splitting pairs: the diverging fates of duplicated genes. *Nature Review Genetics* **3:** 827-837.

Read DJ, Perez-Moreno J. **2003.** Mycorrhizas and nutrient cycling in ecosystems - a journey towards relevance? *New Phytologist* **157:** 475-492.

Remy W, Taylor TN, Hass H, Kerp H. **1994.** Four hundred-million-year-old vesicular arbuscular mycorrhizae. *Proc.Natl.Acad.Sci.U.S.A* **91:** 11841-11843.

Sawatzki HK, Cooper DN. **2007.** Understanding the recent evolution of the human genome: Insights from human-chimpanzee genome comparisons. *Human Mutation* **28:** 99-130.

Schüßler A, Schwarzott D, Walker C. **2001.** A new fungal phylum, the Glomeromycota: phylogeny and evolution. *Mycological Research* **105:** 1413-1421.

Smith SE, Read DJ. **1997.** *Mycorrhizal symbiosis*. Academic Press, San Diego, CA.

Taylor A, Alexander I. **2005.** The ectomycorrhizal symbiosis: life in the real world. *Mycologist* **19:** 102-112.

Trappe JM. **1962.** Fungus Associates of Ectotrophic Mycorrhizae. *Botanical Review* **28:** 538-606.

Wallander H, Söderström B. **1999.** Interactions between *Paxillus involutus* and its hosts. In:*Ectomycorrhizal Fungi: Key Genera in Profile*. Springer Verlag, Berlin, 231-252.

Wösten HA. **2001.** Hydrophobins: multipurpose proteins. *Annu.Rev.Microbiol.* **55:** 625-646.

Wray GA. **2007.** The evolutionary significance of cis-regulatory mutations. *Nat.Rev.Genet.* **8:** 206-216.

Wright DP, Johansson T, Le Quéré A, Söderström B, Tunlid A. **2005.** Spatial patterns of gene expression in the extramatrical mycelium and mycorrhizal root tips formed by the ectomycorrhizal fungus *Paxillus involutus* in association with birch (*Betula pendula*) seedlings in soil microcosms. *New Phytol.* **167:** 579-596.

Zhang JZ. **2003.** Evolution by gene duplication: an update. *Trends in Ecology & Evolution* **18:** 292-298.

Appendix A

The following bioinformatics tools, molecular databases, statistics software's, text editors and computer programs have been used for data analyses in this thesis.

A.1 Bioinformatics tools

- ATV : for the visualization and annotating phylogenetic trees.
- BELVU : to view alignment, inferring phylogenetic trees and colour output
- BIOEDIT : is a biological sequence alignment editor
- BIOLAYOUT : Visualization of biological networks
- BLAST : finds regions of similarity between biological sequences
- CAFE : for the statistical analysis of the evolution of the size of gene families
- CLUSTAL : for multiple sequence alignment
- CODONW : multivariate analysis of codon and amino acid usage
- CRANN : to detect adaptive evolution in protein-coding genes
- DAMBE : data analysis in molecular biology and evolution
- DNASP : analysis of nucleotide polymorphism from aligned DNA sequence
- EMBOSS : an open source software analysis package for molecular biology
- HMMER : profile hidden markov model software for protein sequence analysis
- HYPERTREE : a phylogenetic tree viewer, with a hyperbolic view
- JCOLORGRID : for transforming numerical or text data into color-grids
- MATGAT : calculates similarity and identity between pairs of sequences
- MEGA : for molecular and sequence analysis
- MUSCLE : multiple sequence comparison by log-expectation
- PAML : phylogenetic analyses of molecular sequences using maximum likelihood
- PAUP : for inferring and interpreting phylogenetic tools
- PHYLIP : a free package of programs for inferring phylogenies
- PHYLO_WIN : a graphical colour interface for molecular phylogenetic inference
- SEAVIEW : a graphical multiple sequence alignment editor
- SPLITSTREE : for computing evolutionary networks from molecular sequence data
- TREEVIEW : a simple program for displaying phylogenies
- TRIBE-MCL : for clustering proteins into protein families
- QUICKTREE : rapid reconstruction of phylogenies by the Neighbor-Joining method

A.2 Programming & Databases

- AWK : a general purpose programming language for processing text-based data
- BIOPERL : a toolkit of Perl modules useful in building bioinformatics solutions
- MYSQL : an open source, multi-user SQL database management system
- PERL : a general-purpose programming language
- SED : a Unix utility for parsing text files and the programming language
- SHELL script : a simple domain-specific programming language

A.3 Bioinformatics tools in web servers

- BIOWEB Pasteur : biological software's on the web
- CBS : bioinformatics tools produced at center for biological sequence analysis
- EBI : centre for research and services in bioinformatics
- NCBI : a national resource for molecular biology information
- MATRIX2PNG : for making visualizations of microarray and other data types

A.4 Molecular databases

- BROAD institute, which is one of the largest genome sequencing centers in the world
- COGEME : consortium for the functional genomics of microbial eukaryotes
- EBI : biological databases, support to data deposition and exploitation
- JGI : provides high-quality genome sequencing data to scientific community
- NCBI : biological databases, literature Complete Genomes, Taxonomy, and others
- PARSY : *Paxillus* database for EST analysis
- Pfam : a large collection of protein families
- PHOREST : web based tool for comparative analysis of EST sequences
- SWISSPROT : a curated protein sequence database

A.5 Statistical software's

- KALEIDAGRAPH : graphing and data analysis
- MVSP : multivariate numerical analyses
- OriginLab : professional graphing and data analysis
- PCORD : multivariate statistical analysis

A.6 Text editors

- EDITPLUS : text editor, HTML editor and programmers editor for Windows
- EMACS : extensible, customizable, self-documenting, real-time display editor
- Vi : a screen-oriented text and command editor

A.7 Others

- ADOBE ILLUSTRATOR : sophisticated a vector-graphics software

* These software's were used on Microsoft Windows, Mac OS X and Linux operating system.

This manuscript has been printed in Nature (2008), March 6, 452 (7183): 88-92

The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis

F Martin[1*], A Aerts[2+], D Ahrén[3+], A Brun[1+], F Duchaussoy[1+], J Gibon[1+], A Kohler[1+], E Lindquist[2+], V Pereda[1+], A Salamov[2+], HJ Shapiro[2+], J Wuyts[1,4+], D Blaudez[1], M Buée[1], P Brokstein[2], B Canbäck[3], D Cohen[1], PE Courty[1], PM Coutinho[5], EGJ Danchin[5], C Delaruelle[1], JC Detter[2], A Deveau[1], S DiFazio[6], S Duplessis[1], L Fraissinet-Tachet[8], E Lucic[1], P Frey-Klett[1], C Fourrey[1], I Feussner[7], G Gay[8], J Grimwood[9], PJ Hoegger[10], P Jain[11], S Kilaru[10], J Labbé[1], YC Lin[4], V Legué[1], F Le Tacon[1], R Marmeisse[8], D Melayah[8], B Montanini[1], M Muratet[11], U Nehls[12], H Niculita-Hirzel[13], MP Oudot-Le Secq[1], M Peter[1,14], H Quesneville[15], B Rajashekar[3], M Reich[1,10], N Rouhier[1], J Schmutz[9], T Yin[16], M Chalot[1++], B Henrissat[5++], U Kües[10++], S Lucas[2++], Y Van de Peer[4++], G Podila[11++], A Polle[10++], PJ Pukkila[17++], PM Richardson[2++], P Rouzé[4,18++], IR Sanders[13++], JE Stajich[19++], A Tunlid[3++], G Tuskan[16++] & IV Grigoriev[2++]

(1) UMR 1136, INRA-Nancy Université, Interactions Arbres/Microorganismes, INRA-Nancy, 54280 Champenoux, France. (2) US DOE Joint Genome Institute, Walnut Creek, CA 94598. (3) Microbial Ecology, Ecology Building, Lund University, Sweden. (4) Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, B-9052 Ghent, Belgium. (5) Architecture et Fonction des Macromolécules Biologiques, UMR 6098 CNRS and Universites of Aix-Marseille I & II, Marseille, France. (6) Department of Biology, West Virginia University, Morgantown, WV 26506 USA. (7) Department for Plant Biochemistry, Albrecht-von-Haller-Institute for Plant Sciences, Georg-August-Universität Göttingen, 37077 Göttingen, Germany. (8) Université de Lyon, Université Lyon 1, UMR CNRS - USC INRA d'Ecologie Microbienne, Villeurbanne France. (9) Stanford Human Genome Center, Stanford/JGI, USA. (10) Institute of Forest Botany, Georg-August-Universität Göttingen, 37077 Göttingen, Germany. (11) Department of Biological Sciences, University of Alabama, Huntsville, USA. (12) Eberhard-Karls-Universität, Physiologische Oekologie der Pflanzen, D-72076 Tübingen, Germany. (13) Dept. of Ecology & Evolution, University of Lausanne, Biophore Building, 1015 Lausanne, Switzerland. (14) Swiss Federal Research Institute WSL, Zuercherstrasse 111, 8903 Birmensdorf, Switzerland. (15) Unité de Recherches en Génomique-Info, Tour Évry 2, 91034 Évry Cedex. (16) Environmental Science Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA. (17) Department of Biology, The University of North Carolina, Chapel Hill, NC 27599-3280, USA. (18) Laboratoire Associé de l'INRA, Ghent University, B-9052 Gent, Belgium. (19) Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720-3102, USA.

* to whom correspondence should be addressed: E-mail: fmartin@nancy.inra.fr
+ These authors contributed equally to this work as second authors.
++ These authors contributed equally to this work as senior authors.

Mycorrhizal symbioses -- the union of roots and soil fungi -- are universal in terrestrial ecosystems and may have been fundamental to land colonization by plants[1,2]. Boreal, temperate, and montane forests all depend upon ectomycorrhizae[1]. Identification of the primary factors that regulate symbiotic development and metabolic activity will therefore open the door to understanding the role of ectomycorrhizae in plant development and physiology, allowing the full ecological significance of this symbiosis to be explored. Here, we report the genome sequence of the ectomycorrhizal basidiomycete *Laccaria bicolor* (Fig. 1) and highlight gene sets involved in rhizosphere colonization and symbiosis. This 65-million-base genome assembly contains ~20,000 predicted protein-encoding genes and a very large number of transposons and repeated sequences. We detected unexpected genomic features most notably a battery of effector-type small secreted proteins (SSP) with unknown function, several of which are only expressed in symbiotic tissues. The most highly expressed SSP accumulates in the proliferating hyphae colonizing the host root. The ectomycorrhizae-specific proteins likely play a decisive role in the establishment of the symbiosis. The unexpected observation that the genome of *L. bicolor* lacks carbohydrate-active enzymes involved in degradation of plant cell walls, but maintains the ability to degrade non-plant cell walls, reveals the dual saprotrophic and biotrophic lifestyle of the mycorrhizal fungus which enables it to grow within both soil and living plant roots. The predicted gene inventory of the *L. bicolor* genome, therefore, points to previously unknown mechanisms of symbiosis operating in biotrophic mycorrhizal fungi. The availability of this genome provides an unparalleled opportunity to develop a deeper understanding of the processes by which symbionts interact with plants within their ecosystem in order to perform vital functions in the carbon and nitrogen cycles that are fundamental to sustainable plant productivity.

The 65 million base pairs genome of *Laccaria bicolor* (Maire) P.D. Orton (hereafter referred to as *Laccaria*) is the largest sequenced fungal genome published so far[3,4,5,6,7] (Supplementary Table 3). While no evidence for large scale duplications was observed within the *Laccaria* genome, tandem duplication occurred within multigene families (Supplementary Fig. S4). Transposable elements (TE) comprised a higher proportion (21%) than that identified in the other sequenced fungal genomes and may therefore account for the relatively large genome of *Laccaria* (Supplementary Table 4, Supplementary Fig. S16).

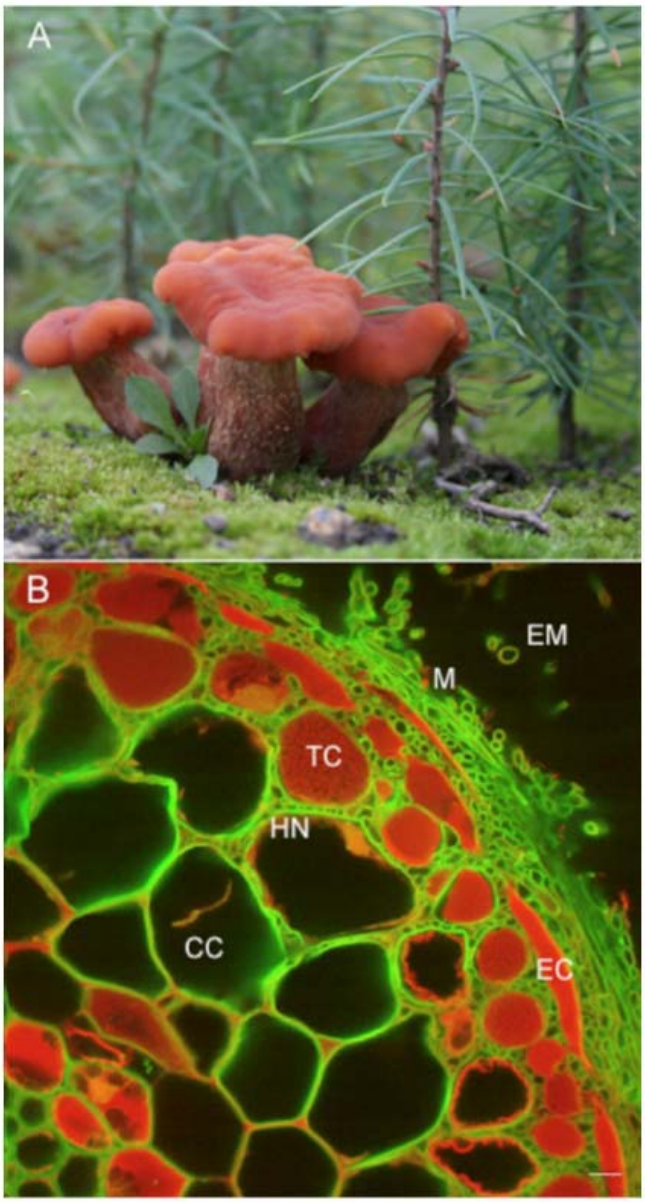


Figure 1. The ectomycorrhizal fungus *Laccaria bicolor*. A, Fruiting bodies of *L. bicolor* colonizing seedlings of Douglas fir (*Pseudotsuga menziesii*). The subterranean mycelial web has developed symbiotic ectomycorrhizal tissues on host root tips and has produced fruiting bodies above ground (Photograph courtesy of D. Vairelles, INRA-Nancy). **B**, Laser scanning confocal microscopy image of a transverse section of *Populus trichocarpa-L. bicolor* ectomycorrhizal root tips showing extramatrical mycelium (em), aggregated hyphae of the mantle sheath (m), hyphae proliferating between the epidermal (ec), tannin (tc) and cortical (cc) of the host root to form the symbiotic Hartig net (hn).
Bar = 10 μm.

Approximately 20,000 protein-coding genes were identified by combined gene predictions (Supplementary Online Material [SOM]). Expression of nearly 80% (ca. 16,114) of the predicted genes was detected in either free-living mycelium, ectomycorrhizal root tips or fruiting bodies (Supplementary Table 5) using NimbleGen custom-oligoarrays (SOM). Most genes are activated in almost all tissues, whereas other more specialized genes were only activated in some specific developmental stages, such as free-living mycelium, ectomycorrhizae or fruiting body (Supplementary Table 6).

Only 14,464 of *Laccaria* proteins (70%) showed significant similarity to documented proteins. Most homologs were found in the sequenced basidiomycetes *Phanerochaete chrysosporium*[4], *Cryptococcus neoformans*[5], *Ustilago maydis*[6], and *Coprinopsis cinerea*[7] (Supplementary Table 7). The percentage of proteins found in multigene families was related to genome size and was the largest in *Laccaria* (Supplementary Fig. S5). This was mainly due to the expansion of protein family size, but also to the larger number of protein families in *Laccaria* when compared to the other basidiomycetes (Supplementary Table 8). Expansion of protein family sizes in *Laccaria* was prominent in the lineage-specific multigene families. Striking gene family expansions occurred in those genes predicted to have roles in protein-protein interactions (e.g. WD40) and signal transduction mechanisms (Supplementary Table 8). Two new classes of Ga genes were found and may be candidates for the complex communication that must occur between the mycobiont and its host-plant during mycorrhizae establishment (Supplementary Table 9). Several transcripts coding for expanded and lineage-specific gene families were up-regulated in symbiotic and fruiting body tissues, suggesting a role in tissue differentiation (Supplementary Table 10).

In our analysis of annotated genes, and in particular paralogous gene families, we highlighted processes which may be related to the biotrophic and saprotrophic lifestyles of *Laccaria*. Twelve predicted proteins showed a similarity to known effector proteins of *Magnaporthe grisea*[8] and haustoria-expressed secreted proteins (HESP) of *Melampsora lini*[9] which are involved in pathogenesis (Supplementary Table 11). Of the 2,931 proteins predicted to be secreted by *Laccaria*, most (67%) cannot be ascribed a function and 82% of these predicted proteins are

Table 1. Changes in the expression of transcripts coding for mycorrhizae-induced cysteine-rich small secreted proteins (MISSPs)

Protein ID	Family size	Length (AA)	Transcript Concentration (FLM)	*Pseudotsuga* ECM/FLM Ratio (fold)	*Populus* ECM/FLM Ratio (fold)	Features
298595	sc	68	nd	21877	12913	MISSP7
333839	5	129	nd	7844	1931	GPI-anchored
298667	2	70	nd	1906	1407	
332226	8	181	43	847	780	CFEM domain (IPR014005)
311468	29	59	nd	191	nd	
295737	8	288	131	171	252	
334759	sc	101	nd	109	18	
395403	4	121	24	103	93	
333423	9	120	6	102	72	Gonadotropin domain (IPR0001545)
312262	4	106	85	69	53	
295625	4	199	325	66	48	
325402	8	238	310	49	74	Snake toxin-like (SSF57302)
316998	sc	56	137	29	57	
333197	3	148	266	17	8	
327918	2	154	763	13	4	Homolog in *Coprinopsis cinerea*
307956	sc	74	336	13	90	Whey acidic domain (IPR008197)
327246	sc	194	1025	10	18	Homolog in *Coprinopsis cinerea*
303550	5	98	1365	10	14	
300377	2	291	5499	10	8	
293250	sc	224	127	9	10	Homolog in *Coprinopsis cinerea*
298648	sc	64	1108	8	12	
298646	2	73	1028	7	14	
293729	3	210	3000	7	7	

Transcript profiling was performed on free-living mycelium (FLM), and *Populus* and Douglas-fir ectomycorrhizal root tips (ECM). Values are the means of technical and biological duplicates. Based on the statistical analysis, a gene was considered significantly upregulated if it met all two criteria: (1) *t*-test *P*-value < 0.001 (Cyber-T web interface, http://www.igb.uci.edu/servers/cybert/); (2) mycorrhizae *vs* control fold change ≥ 4. Cut-off values for signal intensity (50 to 100 arbitrary units), corresponding to three times the background values estimated from random 60-mer probes on the NimbleGen oligoarrays, have been subtracted from the normalized intensity values. The highest signal intensity value observed on these arrays was 65,535 arbitrary units. Abbreviations: AA, amino acids; nd, not detected; sc, single copy.

specific to *Laccaria*. Within this set, we found a large number of genes that encode cysteine-rich products with a predicted size of <300 amino acids. Of these 278 small secreted proteins (SSPs), 69% belong to multigene families, but only nine groups comprising a total of 33 SSPs co-localized in the genome (Supplementary Fig. S7). The structures of two of these clusters are shown in Supplementary Fig. S6. Other SSPs are scattered all over the genome and we found no correlation between SSPs and TE genome localization (Supplementary Fig. S7). Transcript profiling revealed that the expression of several SSP genes is specifically induced upon in the symbiotic interaction (Table 1, Supplementary Fig. S11). Five of the 20 most highly up-regulated fungal transcripts in ectomycorrhizal root tips code for SSPs (Supplementary Table 6). These mycorrhiza-induced cysteine-rich SSPs (MISSPs) belong to *Laccaria*-specific orphan gene families. Within the MISSPs, we found a family of secreted proteins with a CFEM domain (IPR014005) (Supplementary Fig. S8 & S9), as previously identified in the plant pathogenic fungi *M. grisea*[8] and *M. lini*[9] (Supplementary Table 11), and proteins with a gonadotropin- (IPR0001545) or snake toxin-like (SSF57302) domains related to the cysteine-knot domain. Expression of several SSPs were down-regulated in ectomycorrhizal root tips (Supplementary Fig. S11) suggesting a complex interplay between these secreted proteins in symbiosis interaction.

The rich assortment of MISSPs may therefore act as effector proteins to manipulate host cell signalling or suppress defence pathways during infection, as suggested for pathogenic rusts[9], smuts[6] (*U. maydis*) and *Phytophthora*[10] species. To play a role in symbiosis development, MISSPs should be expressed in *Laccaria* hyphae colonizing the root tips. To test this assertion, we determined the tissue distribution of the MISSP7 protein (ID 298595) showing the highest induction in ectomycorrhizal tips (Table 1, Supplemental Table S6). Two peptides located in the N-terminal and C-terminal parts of the mature protein were selected as antigens for the production of anti-MISSP7 immunserum. The selected peptides were not found in the deduced protein sequences of other *Laccaria* gene models nor in the *Populus trichocarpa* genome[11]. MISSP7 localization in *Laccaria*/*Populus* ectomycorrhizal root tips by indirect immunofluorescence is illustrated in Fig. 2. Control images in which the ectomycorhizae sections were not incubated with primary anti-MISSP7 antibody, but has been exposed to

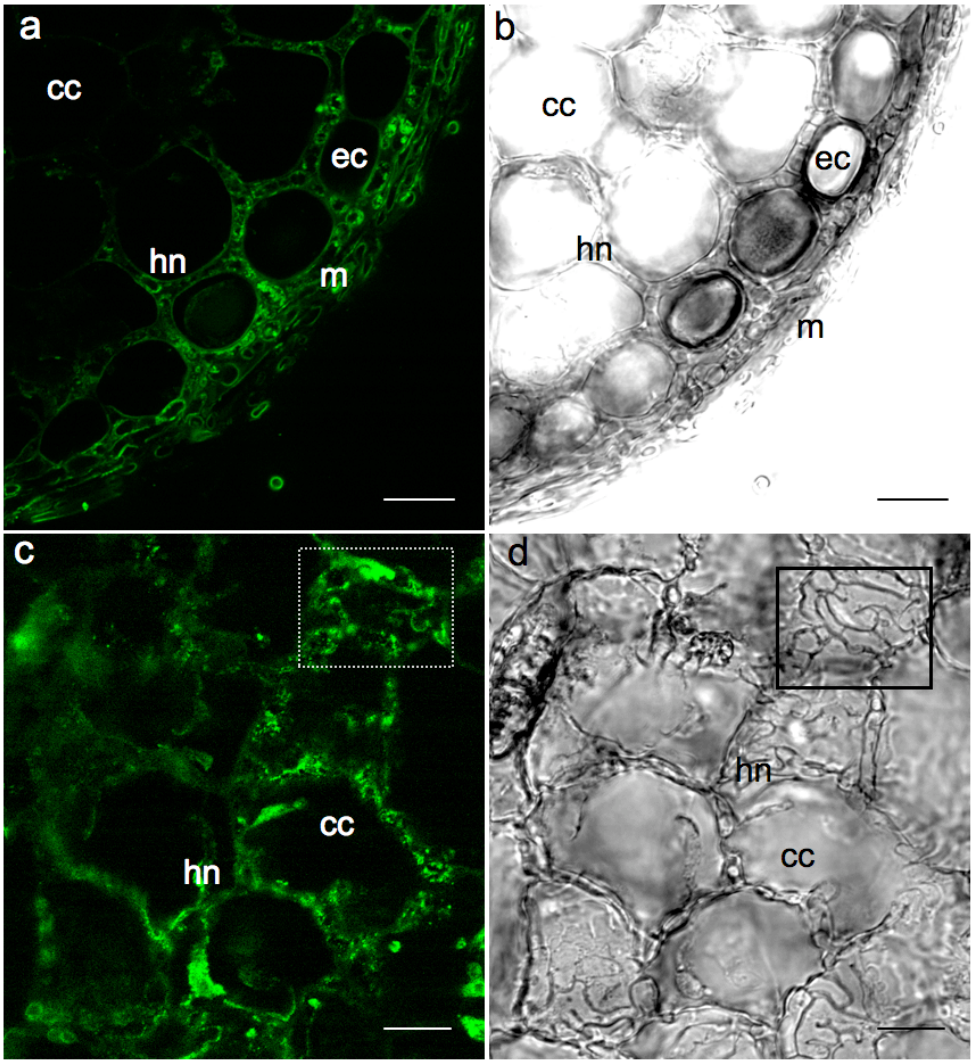


Figure 2. Indirect immunofluorescent localization of the small secreted protein MISSP7 in *Populus trichocarpa*/*L. bicolor* ectomycorrhizal root tips. Transverse (a, b) and longitudinal (c, d) sections of ectomycorrhizal tips. MISSP7 was detected with anti-MISSP7 IgG and secondary antibody conjugated with AlexaFluor 488 in the hyphae of the mantle (m) and the uniseriate Hartig net (hn) ensheathing the epidermal (ec) and cortical (cc) cells of the colonized roots. Dotted lines in panels (c) & (d) show the finger-like, labyrinthine hyphal system accumulating large amount of MISSP7. (b) and (d), phase contrast images. Bar = 10 µm.

the preimmunserum or the secondary label are shown in Supplementary Fig. S13. Where ectomycorrhizae were treated with anti-MISSP7 antibody followed by fluorescent-labeled secondary antibody, fluorescence is seen localized in the hyphae colonizing short roots (Fig. 2, Supplementary Fig. S12) and not detected in the free-living mycelium (Supplementary Fig. S13). Although MISSP7 was detected in the hyphal mantle layers ensheathing the root tips, the protein mainly accumulated in the finger-like, labyrinthine branch hyphal system (Hartig net) which provides a very large area of contact between cells of the two symbionts. It accumulated in the cytosol and cell wall of the fungal cells. The MISSP7 protein could interact with the plant components after secretion, but the mechanisms by which it may enter the plant cell are unknown. MISSP7 shares no sequence similarity or protein motif with other SSPs. Comparison of the MISPP sequences did not reveal a specific conserved motif, such as the RXLR motif of phytopathogenic *Phytophthora* or the malaria parasite, that could potentially contribute to the their function or to targeting to the host cell. Those SSP with an upregulated expression in fruiting body (Supplementary Table 6, Supplementary Fig. S11) may play a role in the differentiation of the sexual tissues and/or aggregation of sporophore tissues. Interestingly, they are a set of SSP genes showing significant changes in gene expression in both ectomycorrhizal root tips and fruiting body suggesting that both developmental processes recruit similar gene networks (e.g., those involved in hyphal aggregation).

Host trees are able to harness the formidable web of mycorrhizal hyphae, that permeates the soil and leaf litter, for their nutritional benefit. A process that is pivotal to the success of ectomycorrhizal interactions is the equitable exchange of nutrients between the symbiont and its host-plant[1,2,12]. A comparison with other basidiomycetes (Supplementary Table 13) revealed that the total number of predicted transporters has been expanded in *Laccaria* compared to *C. cinerea* and *P. chrysosporium*. The largest families of transporters in these fungi inhabiting soils and decaying litter encode members of the major facilitator (MFS) and ATP-binding cassette (ABC) superfamilies. Interestingly, *Laccaria* has multiple ammonia (AMT) transporters although it encodes a single nitrate permease. Ammonia is arguably the most important inorganic nitrogen source for ectomycorrhizal fungi[13]. One of the ammonia transporters (*LbAMT2.2*), for instance, is greatly

up-regulated in ectomycorrhizae (Supplementary Table 6). *Laccaria,* thus, shows an increased genetic potential in terms of nitrogen uptake when compared to other basidiomycetes. These capabilities are consistent with *Laccaria* being exposed to a range of nitrogen sources from organic matter decay[14]. *Laccaria* also has a greater array of metabolite efflux systems, such as ABC transporters, possibly reflecting a need to protect colonizing hyphae against toxic concentrations of metabolites in the rhizosphere and root apoplast. One of the seven putative *Laccaria* inorganic phosphate transporters is upregulated in ectomycorrhizal tips. Finally, *Laccaria* is capable of transporting a wide range of sugars derived from plant metabolism. These broad uptake capabilities likely contribute to rhizosphere and host root colonization, because sugars and amino acids are the main metabolites translocated between the symbiotic partners[15] and are also present in the rhizosphere as products of the organic matter degradation.

Although the *Laccaria* genome contains numerous genes coding for key hydrolytic enzymes, such as proteases and lipases, we observed an extreme reduction in the number of enzymes involved in the degradation of plant cell wall (PCW) oligo- and polysaccharides. Glycoside hydrolases (GH), glycosyltransferases (GT), polysaccharide lyases (PL), carbohydrate esterases (CE) and their ancillary carbohydrate-binding modules (CBM) were identified using the carbohydrate-active enzyme classification (http://www.cazy.org/). A comparison of the *Laccaria* candidate CAZymes with fungal phytopathogens confirms the adaptation of its enzyme repertoire to symbiosis and reveals the strategy used for the interaction with the host (Supplementary Tables 14 and 15). The reduction in PCW CAZymes affects almost all GH families culminating in the complete absence of several key families. For instance, there is only one candidate cellulase (GH5) appended to the sole fungal cellulose-binding module (CBM1) found in the genome and no cellulases from families GH6 and GH7 (Supplementary Table 15). Similar reductions or loss of hemicellulose and pectin degrading enzymes were also noted. These observations suggest that the inventory of *Laccaria* PCW degrading enzymes underwent massive gene loss as a result of its adaptation to a symbiotic lifestyle and that this species is now unable to use many PCW polysaccharides as a carbon source, including those found in soil and leaf litter. The remaining small set of secreted CAZymes with potential

action on plant polysaccharides (e.g. GH28-polygalacturonases) is probably required for cell wall remodeling during fungal tissue differentiation as their expression was up-regulated in both fruiting body and ectomycorrhizae (Supplementary Table 15, Supplementary Fig. S14). In contrast, transcripts coding for protein with expansin domain were only induced in ectomycorrhizae suggesting they may be used by *Laccaria* for penetrating into the root apoplastic space. To survive before its mycorrhizal association with its host, *Laccaria* appears to have developed a capacity to degrade non-plant (e.g. animal, bacterial) oligo- and polysaccharides which is suggested by retention of CAZymes from families GH79, PL8, PL14 and GH88 (Supplementary Table 15). Interestingly, there is no invertase gene in the *Laccaria* genome, implying that this fungus is unable to directly use sucrose from the plant. This is consistent with earlier observations[15] that *Laccaria* depends on its host plant to provide glucose in exchange for nitrogen. We also noticed an expansion of CAZymes involved in the fungal cell wall biosynthesis and rearrangement, almost entirely due to an increased number of putative chitin synthases and enzymes acting on b-glucans (Supplementary Table 15). Several of the corresponding genes are up- or down- regulated upon developmental processes requesting cell wall alterations such as formation of fruiting bodies or mycorrhizae (Supplementary Table 16, Supplementary Fig. S14).

Ectomycorrhizal fungi play a significant role in mobilizing N from well-decomposed organic matter[2,14]. The hyphal network permeating the soil might therefore be expected to express a wide diversity of proteolytic enzymes. The total number of secreted proteases (116 members) identified (Supplementary Fig. S15) is relatively large compared with other sequenced saprotrophic basidiomycetes, such as *C. cinerea* and *C. neoformans*. Secreted aspartyl-, metallo- and serine-proteases may play a role in degradation of decomposing litter[14] confirming that *Laccaria* has the ability to use nitrogen of animal-origin, as suggested previously[16]. They may also play a role in developmental processes as the expression of several secreted proteases is up- or downregulated in fruiting bodies and ectomycorrhizal root tips (Supplementary Table 17). Mycelial mats formed by *Laccaria* hyphae colonizing organic matter therefore possess the ability to degrade decomposing leaf litter.

Our analysis of the gene space reveals a multi-faceted mutualistic biotroph equipped to take advantage of transient occurrences of high-nutrient niches (living host roots and decaying soil organic matter) within a heterogeneous, low-nutrient environment. The availability of genomes from mutualistic, saprotrophic[4], and pathogenic[6] fungi, but also from the mycorrhizal tree *Populus trichocarpa*[11], now provides an unparalleled opportunity to develop a deeper understanding of the processes by which fungi colonize wood and soil litter, and also interact with living plants within their ecosystem in order to perform vital functions in the carbon and nitrogen cycles[2] that are fundamental to sustainable plant productivity.

Methods summary

Gene assembly and annotation.
The haploid genome of the strain S238N-H82 from *L. bicolor* (Maire) P.D. Orton was sequenced using a whole-genome shotgun strategy, assembled with JAZZ assembler, and annotated by combining *ab initio* models and alignments of proteins, ESTs and genomic DNA from *L. bicolor* and other basidiomycete species, which were integrated in JGI Annotation pipeline (SOM).

Secreted proteins.
Secreted proteins were predicted to carry a signal peptide both by the hidden Markov and the neural network algorithms implemented in SignalP 3.0. After eliminating predicted transmembrane proteins, we selected secreted cysteine-rich proteins with a size <300 AA, yielding 278 candidate small secreted proteins.

DNA-array analysis.
For transcript profiling, custom-designed NimbleGen oligoarrays were used. For each predicted gene, eight perfect oligoprobes were designed (SOM).

Supplementary information, full methods and any associated references are available in the online version of the paper at www.nature.com/nature

Methods

Genome sequencing.

The haploid genome of the strain S238N-H82 from *L. bicolor* (Maire) P.D. Orton was sequenced with the use of a whole-genome shotgun (WGS) strategy. All data were generated by paired-end sequencing of cloned inserts using Sanger technology on ABI3730xl sequencers. Supplementary Table 1 gives the number of reads obtained per library.

Genome assembly.

The data was assembled using release 1.0.1b of JAZZ, a WGS assembler developed at the JGI (Chapman, Putnam, Ho & Rokhsar, unpublished). Based on the number of alignments per read, the main genome scaffolds were at a depth of 9.88. The amount of sequence in the unplaced reads was 6.5 Mbp, which is sufficient to cover the main-genome gaps to a mean depth of 9.9. This is extremely close to the overall main genome scaffold depth, hinting that the unplaced reads might represent the gap sequence. This resulted in 21 scaffolds (i.e. 180 contigs), each longer than 784 kbp, covering a total of 30 Mbp or half of the main genome (*N50*). A total of 65 Mbp are captured in the scaffold assembly (Supplementary Table 2).

Genome annotation.

Several resources were used to build gene models automatically with FgenesH[17], homology-based FgenesH+[17] (www.softberry.com), Genewise[18], EST-based estExt (I. Grigoriev, unpublished), as well as EuGène[19] and TwinScan[20], and alignments of several cDNA resources (SOM). The JGI pipeline selected a best representative gene model for each locus based on EST support and similarity to known proteins from other organisms, and predicted 20,310 protein-coding gene models.

Secreted proteins.

For the prediction of secreted proteins, all *Laccaria* gene models predicted by the JGI machine annotation pipeline were searched for the presence of a signal peptide, using the TargetP and SignalP 3.0 algorithms[22]. A total of 2,931 proteins were predicted to carry a signal peptide both by the hidden Markov and the neural network algorithms. After eliminating predicted transmembrane proteins, we selected cysteine-rich secreted proteins with a size <300 AA, yielding 439 candidate small secreted proteins (SSPs). Gene models with similarity

with TE fragments were eliminated at which point 278 SSP genes had been identified in *Laccaria*.

Indirect immunofluorescent localization of MISSP7.
The peptides LRALGQASQGGDLHR and GPIPNAVFRRVPEPNF located in the N-terminal and C-terminal parts of the MISSP7 sequence (without the signal peptide) were synthesized and used as antigens for the generation of antibodies in rabbits according to the manufacturer's procedures (Eurogentec, Seraing, Belgium). The anti-MISSP7 IgG fraction was purified using MAbTrap kit (GE Healthcare) according to the manufacturer's recommendations. Subsequently, IgG-containing fraction was desalted using a HiTrap™ desalting column (GE Healthcare). The concentration of purified IgG preimmune serum was determined by Bradford assay using a Bio-Rad protein assay. Final concentration of anti-MISSP7 IgG was 0.16 mg/ml.

Immunolocalization was performed essentially as described by [23,24] with slight modifications. Radial and longitudinal sections of non-mycorrhizal root tips from 3-month-old *P. trichocarpa* (cv. 101-74), free-living mycelium of *Laccaria,* and ectomycorrhizal root tips from 3-month-old *P.trichocarpa* (cv. 101-74) were fixed for 2h in 4% formaldehyde in PME buffer (50 mM Pipes, 5 mM $MgSO_4$, and 10 mM EGTA), pH 6.9. The root segments were embedded in agarose 4% and cut into 25 mm longitudinal or transverse sections with a model 1000 Vibratome (Leica). Sections were retrieved with a brush and carefully transferred onto watch glasses and then were digested in 1% cellulase, 0.01% pectolyase, and 0.1% BSA in PME buffer for 10 min. After digestion, the segments were washed five times for 5 min each with PME buffer and then incubated in 1% BSA in PBS (135 mM NaCl, 25 mM KCl; and 10 mM Na_2HPO_4, pH 7.5) for one hour. The BSA was removed, and the segments were incubated overnight with purified anti-MISSP7 protein rabbit antibody diluted 1:1500 in PBS containing 0.5% (w/v) BSA at 4°C. The segments were then washed five times in PBS and incubated in the secondary antibody conjugate, a 1:80 dilution of goat anti-rabbit IgG-AlexaFluor 488 conjugate (Molecular Probes, Invitrogen™, Carlsbad, CA) in PBS for 2 h. After five more washes in PBS, sections were mounted in 80% glycerol (Merck), 20% PBS, 5% w/v propyl gallate (Fluka) and viewed by a Bio-Rad Radiance 2100 AGR3Q-BLD Rainbow microscope equipped with X20, X40 and X60, numerical

aperture 1.4. The excitation and emission wavelengths for the Alexa Fluor 488 dye were 500 to 550 nm, respectively. Optical sections were collected at 0.1 to 0.7 mm intervals with Kalman averaging. As a control, sections were incubated with IgG purified from pre-immune serum diluted to the same concentration as anti-MISSP7 IgG. Section shown in Fig. 1 was stained with propidium iodide (red) which colours cell walls and nuclei of plant and fungus, and Uvitex (green) which stains plant and fungal cell walls.

Gene expression.
For transcript profiling, custom-designed NimbleGen oligoarrays were used. Probe sets were designed on the basis of the JAZZ sequencing assembly.

Acknowledgements

The genome sequencing of *Laccaria bicolor* H82 was funded by the U.S. Department of Energy's Office of Science, Biological, and Environmental Research Program and by University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396. The EST sequencing and transcriptome analysis were funded by the U.S. Department of Energy, INRA 'AIP Séquençage' and Région Lorraine grants. Annotation was supported by grants from the INRA postdoctoral fellowship programme (to J.W. and M.P.O.L.S.), Région Lorraine, the U.S. Department of Energy's Office of Science, Biological, and Environmental Research Program, the U.S. National Science Foundation (EF-0412016), the European Network of Excellence EVOLTREE, the European Commission (STREP FungWall, LSHB-CT-2004-511952) and the Swedish Research Council. We would like to thank S Rombauts and L Sterck for helpful discussion on the automated annotation and multigene family analysis, K Vandepoele for help in promoter analysis, G Werner and his colleagues for support of JGI annotation tools, S Pitluck and K Zhou for help with GenBank submission, B Hilselberger for help with transcript profile analysis and J Gérard for her assistance in immunolocalization analysis. F.M. thanks Prof. N. Talbot for critical reading of an early draft of the manuscript.

References and notes

1. Smith, S. E. & Read, D. J. Mycorrhizal Symbiosis (2nd edition, Academic Press, London) (1996).

2. Read, D. J. & Perez-Moreno, J. Mycorrhizas and nutrient cycling in ecosystems - a journey towards relevance? New Phytol. **157**, 475-492 (2003).

3. Galagan, J. E., Henn, M. R., Ma, L. J., Cuomo, C. A., Birren, B. Genomics of the fungal kingdom: insights into eukaryotic biology. Genome Res. **15**, 1620-1631 (2005).

4. Martinez, D. *et al.* Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. Nature Biotech. **22**, 695-700 (2004).

5. Loftus, B. J. *et al.* The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. Science **307**, 1321-1324 (2005).

6. Kämper, J. *et al.* Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. Nature **444**, 97-101 (2006).

7. *Coprinus cinereus D*atabase: http://www.broad.mit.edu/annotation/genome/coprinus_cinereus/Home.html

8. Kulkarni, R. D., Kelkar, H. S., Dean, R. A. An eight-cysteine-containing CFEM domain unique to a group of fungal membrane proteins. Trends Bioch. Sci. **28**, 118-118 (2003).

9. Catanzariti, A. M., Dodds, P. N., Lawrence, G. J., Ayliffe, M. A., Ellis, J. G. Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. Plant Cell **18**, 243-256 (2006).

10. Kamoun, S. A. Catalogue of the effector secretome of plant pathogenic oomycetes. Annu. Rev. Phytopathol. **44**, 41-60 (2006).

11. Tuskan, G. A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science **313**, 1596-1604 (2006).

12. Martin, F., Kohler, A., Duplessis, S. Living in harmony in the wood underground: ectomycorrhizal genomics. Curr. Opin. Plant Biol. **10**, 204-210 (2007).

13. Chalot, M., Blaudez, D., Brun, A. Ammonia: a candidate for nitrogen transfer at the mycorrhizal interface. Trends Plant Sci. **11**, 263-266 (2006).

14. Lindahl, B. D., Ihrmark, K., Boberg, J., Trumbore, S. E., Högberg, P., Stenlid, J., Finlay, R. D. Spatial separation of litter decomposition and mycorrhizal nitrogen uptake in a boreal forest. New Phytol. **173**, 611-620 (2007).

15. Nehls, U., Grunze, N., Willmann, M., Reich, M., Küster, H. Sugar for my honey: Carbohydrate partitioning in ectomycorrhizal symbiosis. Phytochem. **68**, 82-91 (2007).

16. Klironomos, J. N. & Hart, M. M. Animal nitrogen swap for plant carbon. Nature **410**, 651 (2001).

17. Salamov, A, Solovyev, V. *Ab initio* gene finding in *Drosophila* genomic DNA. Genome Res **10**: 516-522 (2000).

18. Birney, E, Clamp, M, Durbin, R. GeneWise and genomewise. Genome Res **14**: 988-995 (2004).

19. Schiex, T, Moisan, A, Rouzé, P. EuGène: an eukaryotic gene finder that combines several sources of evidence. In: *Computational Biology*, Gascuel O, Sagot MF, eds, LNCS 2066, pp. 111-125 (2001).

20. Tenney, A. E. *et al.* Gene prediction and verification in a compact genome with numerous small introns. Genome Res **14:** 2330-2335 (2004).

21. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. J Mol Biol **215**: 403-10 (1990).

22. Emanuelsson, O., Brunak, S., von Heijne, G., Nielsen, H. Locating proteins in the cell using TargetP, SignalP, and related tools. Nature Protocols **2**: 953-971 (2007).

23. Blancaflor, E.B., Zhao, L., Harrison, M.J. Microtubule organization in root cells of *Medicago truncatula* during development of an arbuscular mycorrhizal symbiosis with *Glomus versiforme*. Protoplasma **217**: 154–165 (2001).

24. Harrison, M.J., Dewbre, G.R., and Liu, J.Y. A phosphate transporter from *Medicago truncatula* involved in the acquisition of phosphate released by arbuscular mycorrhizal fungi. Plant Cell **14**: 2413–2429 (2002).

The WGS project has been deposited at GenBank/EMBL/DDBJ under project accession ABFE00000000. The version described in this paper including assembly and annotation is the first version ABFE01000000. Annotations and further information on the project are also available from JGI Genome Portal (http://www.jgi.doe.gov/laccaria) and INRA *Laccaria* web site (http://mycor.nancy.inra.fr/IMGC/LaccariaGenome/). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to F.M. (fmartin@nancy.inra.fr).

Manuscript

Expansion of protein families in the symbiotic fungus *Laccaria bicolor*

Balaji Rajashekar[1], Annegret Kohler[2], Jason E. Stajich[3], Yao-Cheng Lin[4], Pierre Rouzé[4], Francis Martin[2], Anders Tunlid[1] and Dag Ahrén[1*]

[1]*Department of Microbial Ecology, Ecology Building, Lund University, SE-223 62, Lund, Sweden.* [2]*UMR1136, INRA-Nancy Université, Interactions Arbres/Microorganismes, INRA-Nancy, 54280 Champenoux, France.* [3]*Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720-3102, USA.* [4]*Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, B-9052 Ghent, Belgium.*

II

* Corresponding author: Dag Ahrén, Department of Microbial Ecology, Ecology Building, Lund University, SE-223 62, Lund, Sweden. Tel.: +46-46-222 3757, Fax: +46-46-222 4158.
Email: Dag.Ahren@mbioekol.lu.se

Key words: basidiomycetes, comparative genomics, *Coprinopsis cinerea*, *Cryptococcus neoformans*, ectomycorrhizae, gene gain, gene loss, *Laccaria bicolor*, *Phanerochaete chrysosporium*, protein family evolution, protein kinase, small GTPases, symbiosis, *Ustilago maydis*

Abstract

Background: Gene duplications and loss are major mechanisms generating evolutionary novelties and pruning specialized functions. Here, we have studied the role of duplication and deletion events for the evolution of ectomycorrhizae, a symbiotic association between fungal hyphae and plant roots. Gene duplicates and gene families in the genomes of the symbiont *Laccaria bicolor*, the saprophytes *Coprinopsis cinerea* and *Phanerochaete chrysosporium*, the human pathogen *Cryptococcus neoformans* and the plant pathogen *Ustilago maydis* were analyzed.

Results: *L. bicolor* contained a higher number of gene duplicates when compared to the other basidiomycetes. The density of young gene duplicates was also highest in the *Laccaria* genome. The differences in gene duplicates had a pronounced effect on the number and the size of multigene families. In total the 7352 protein families were identified in all five basidiomycete genomes. The percentage of proteins found in families was largest in *L. bicolor*. Furthermore, this fungus contained the largest number of lineage-specific (1077 families) and expanded (1064) protein families. Conversely, the numbers of extinct and contracted protein families were lower in *L. bicolor* than in the other basidiomycetes. A large fraction (523 out of 1824) of the young gene duplicates of *L. bicolor* were found within 55 gene families, as a criterion containing more than 25 members. A majority of these families did not contain any known protein domain. However, results from microarray experiments showed that many of them contain proteins that are differentially expressed in mycorrhizal root tips as compared to fruiting bodies and mycelia. Among the protein domains identified were those predicted to have a role in signal transduction mechanisms. Phylogenetic analyses of two such families including protein kinases and small GTPases showed that *L. bicolor* contained clusters of paralogs that have arisen though duplication events in the *Laccaria* lineage.

Conclusions: Our analysis suggests that the evolution of symbiosis in *L. bicolor* has been associated with the duplication of genes and the expansion of large, multigene families. The functions of these new genes are largely unknown. Many of them are differentially expressed during symbiosis.

Background

Mycorrhiza, the symbiotic interactions between fungi and plant roots, are almost universal in terrestrial plants. In these ecologically important interactions, the fungal partner obtains photosynthetic sugars from the host plant while, in return, the plant receives mineral nutrients from the fungus [1]. Fossils from the Rhynie chert, 400 million years ago (Mya), suggest that the mycorrhizal associations, similar to those of the arbuscular mycorrhizae, facilitated the colonization of land by plants [2]. During the subsequent history of plants, other kinds of mycorrhizae evolved. The major type is ectomycorrhizae (EM), which is formed by more than 5000 species of fungi, primarily homobasidiomycetes and some ascomycetes and zygomycetes [2]. About thirty plant families form EM and many of them are the dominating trees in boreal and temperate forests. The EM associations are highly complex involving the development of specialized cells for nutrient and carbon exchange at the fungus-plant interface, and morphological adaptations of the fungus for nutrient assimilation and transport and in the soil. EM fungi are facultative symbionts, as they can grow both as biotrophs and saprophytes [1].

The first occurrence of fossil of EM associations dates 50 Mya [3], but several observations suggest that this symbiosis evolved earlier. First, Pinaceae and many of the angiosperm families, including the Dipterocarpaceae, whose current members establish EM symbiosis were extant well before 50 Mya, along with the major fungal lineages with modern ectomycorrhizal representatives [4,5]. Second, biogeographic analyses of mycorrhizal associations in the neotropical genus of Dipterocarpaceae indicates that EM associations evolved in Gondwana 135 Mya [6].

Molecular phylogenetic analyses have shown that EM interactions have evolved repeatedly from saprophytic precursors [4,7]. For example, within the homobasidiomycetes, EM fungi occur in six independent clades [4]. The genomic mechanisms that could account for the emergence of the EM life style in fungi are not known. In principle, there are four compatible mechanisms that could account for these patterns. First, symbiosis may be associated with the presence of novel genes. In eukaryotes, new genes are mainly acquired by duplications of

genes or larger chromosome regions [8]. Second, adaptations to the symbiotic life style might result from gene loss and deletions. Such processes have been observed in many symbiotic bacteria [9]. Third, symbiotic adaptations may occur as a result of mutations in existing genes. Fourth, fungal symbiosis could be associated with quantitative differences in gene expression.

The first complete genome of the EM fungi, the basidiomycete *Laccaria bicolor* has recently been published (Martin *et al*., manuscript). The 65 Million base pair genome assembly contains ~ 20,000 predicted protein-encoding genes, and is the largest sequenced fungal genome published so far [10-13]. The large size could partly be explained by a higher content of transposable elements (TE) than in other fungal genomes. The *Laccaria* genome also contained a larger percentage of proteins in multigene families compared to other basidiomycetes. Analysis of the *L. bicolor* genome revealed several features that demonstrate the dual biotrophic and saprotrophic lifestyle of EM fungi. Among them is the expression of an array of small secreted proteins, displaying similarity to so-called effector proteins of biotrophic plant pathogens. The *Laccaria* genome also contain numerous genes including extracellular proteases that may play a role in the degradation of litter (Martin *et al*., manuscript)

In this study, we have used comparative genomics, coupled with evolutionary analysis, to get insights into the gain and loss of genes that have occurred during the evolution of mutualism in *L. bicolor*. *L. bicolor* is a member of the Tricholomataceae in the order Agaricales of the homobasidiomycetes. Comparison was made with four other sequenced basidiomycetes, including the saprophytic species *Coprinopsis cinerea* (Psathyrellaceae, Agaricales) and *Phanerochaete chrysosporium* (Corticaceae, Corticiales), the human pathogen *Cryptococcus neoformans* (Tremellaceae, Tremellales) and the plant pathogen *Ustilago maydis* (Ustilaginaceae, Ustilaginales). The five basidiomycete genomes compared are covering about 550 million years of evolution (Figure 1A). We demonstrate that the evolution of symbiosis in *L. bicolor* has been associated with a significant expansion of specific multigene families.

The function for a majority of these families is not yet known, but many of them contain members that are differentially expressed during symbiosis. The evolution of two of the expanded families including protein kinases and Ras GTPases were examined in more detail.

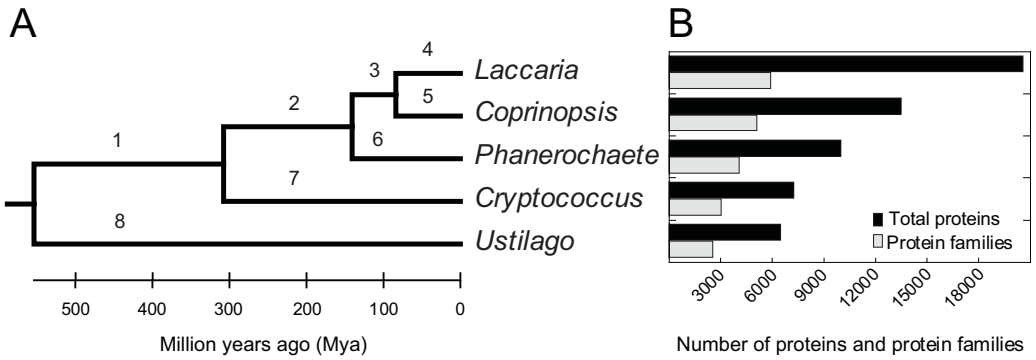


Figure 1. Phylogenetic relationship, number of protein and protein families in the five basidiomycetes used in this study. (A) A linearized Neighbor-Joining tree with 1000 bootstrap replicates of 18S ribosomal DNA sequences from *Laccaria bicolor*, *Coprinopsis cinerea*, *Phanerochaete chrysosporium*, *Cryptococcus neoformans* and *Ustilago maydis* was constructed. *Aspergillus niger* was used to root the tree and the divergence time between *A. niger* and *U. maydis* [14] was used to date other lineages. The branch lengths are in million years indicated by the scale and the bootstrap support for each node was above 98 (values not shown). The numbers on the branches are used for labeling various parts of the tree from the CAFE analysis (c.f. Table 2). (B) Total number of proteins and protein families in the five basidiomycetes.

Results

Number of gene duplicates

The genome of *L. bicolor* contained the highest number of gene duplicates among the compared basidiomycetes. The total number of gene-pairs displaying >90% sequence identity was 1482 in *L. bicolor*, 155 in *C. cinerea*, 1277 in *P chrysosporium*, 227 in *C. neoformans* and 7 in *U. maydis*. The number of gene-pairs with 60 to 90% identity was 1984 in *L. bicolor*, 138 in *C. cinerea*, 565 in *P chrysosporium*, zero in *C. neoformans*, and 9 in *U. maydis*. Further analysis of the duplicated gene-pairs showed that *L. bicolor* contains a significantly higher number of recently duplicated genes as compared to the other basidiomycetes (Figure 2).

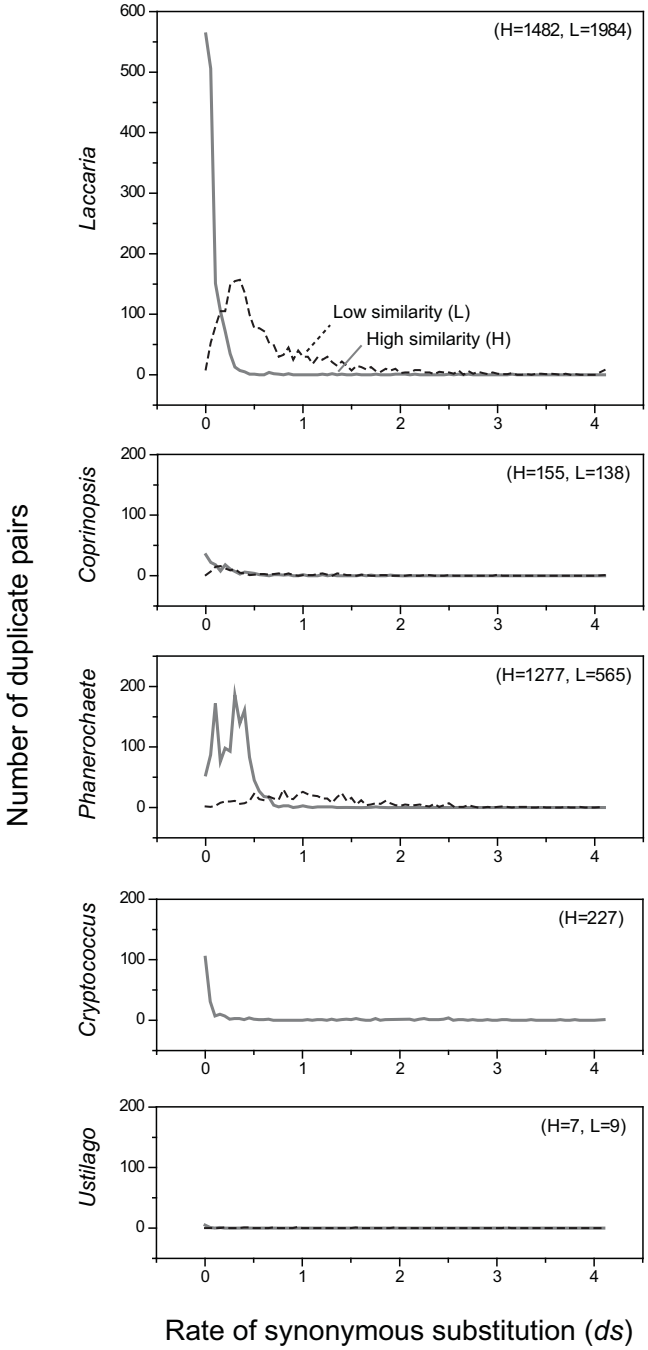


Figure 2. Age distribution of duplicate gene pairs measured by rate of synonymous substitutions per silent site (*ds*). Gene duplications located anywhere in the genomes of *Laccaria bicolor*, *Coprinopsis cinerea*, *Phanerochaete chrysosporium*, *Cryptococcus neoformans* and *Ustilago maydis* were identified. Based on the protein sequence similarity, the gene duplicates in each genome were divided into high (H, above 90% identity) and low (L, 60 to 90% identity) sequence similarity.

The total number of proteins being members of duplicated gene-pairs in *L. bicolor* was 2783, of which 1824 were considered as young duplicates (*ds*<0.2). The corresponding numbers in *C. cinerea* was 387/212 (number of duplicated proteins/young duplicates), *P chrysosporium* 958/324, *C. neoformans* 420/256 and *U. maydis* 32/16.

Protein families

To examine the role of gene duplications for generating novel genes in *L. bicolor*, the protein sequences predicted from the genome sequences of the five compared basidiomycete were clustered into protein families. The protein sequences clustered into 7352 families. As previously reported (Martin *et al*., manuscript), the number of protein families and the size of the families increased with the genome size and were highest in *L. bicolor*, followed by *C. cinerea*, *P. chrysosporium*, *C. neoformans*, and *U. maydis* (Figure 1B) (Table 1).

Table 1. The number of protein families in the genomes of *Laccaria bicolor*, *Coprinopsis cinerea*, *Phanerochaete chrysosporium*, *Cryptococcus neoformans* and *Ustilago maydis*[a]

	Laccaria	***Coprinopsis***	***Phanerochaete***	***Cryptococcus***	***Ustilago***
Families present	5947	5148	4126	3056	2583
Families not present	1405	2204	3226	4296	4769
Proteins in families	17,134	10,614	8118	5001	3820
Proteins not in families	3480	2930	1930	2301	2702
Total number of proteins	20,614	13,544	10,048	7302	6522
Proteins per family (mean)	2.88	2.06	1.97	1.64	1.48

[a]Predicted protein sequences from all genomes were clustered into families using the TRIBE-MCL algorithm [15]. In total 7352 protein families (containing at least two sequences) were identified.

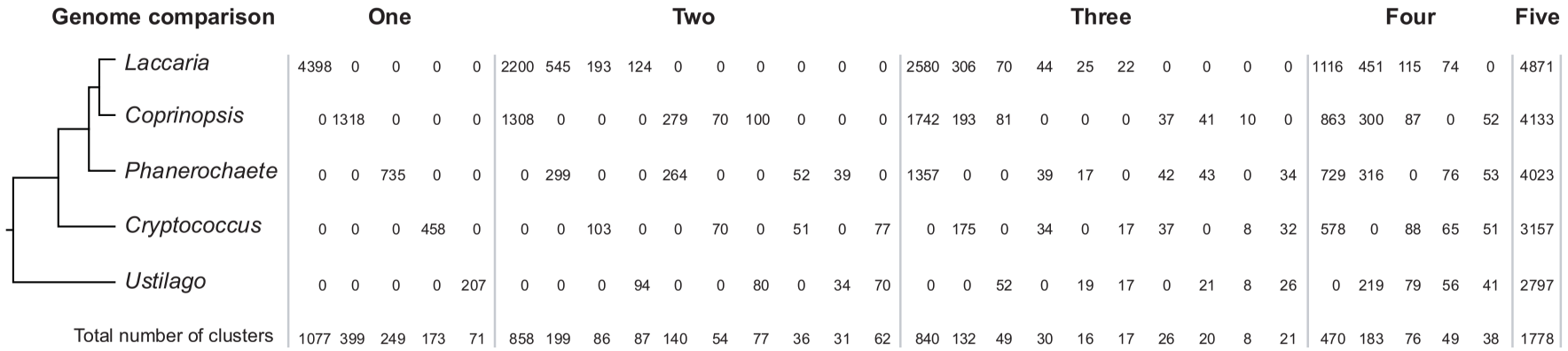


Figure 3. Taxonomic distribution of protein families in basidiomycetes. Protein sequences predicted from the genome sequences of *Laccaria bicolor*, *Coprinopsis cinerea*, *Phanerochaete chrysosporium*, *Cryptococcus neoformans* and *Ustilago maydis* were clustered into families using the TRIBE-MCL algorithm [15]. In total, 7352 protein families (containing at least two sequences) were identified. The matrix shows the number of proteins that were found in families present in only one genome (i.e. lineage-specific), or shared among two, three, four or five (all) genomes. The bottom row shows the total numbers of protein families identified in the different comparisons.

An analysis of the taxonomic distribution of the 7352 protein families showed that 24.2% were present in all five basidiomycetes, 49.0% were shared between any two, three or four species, and the remaining 26.8% were specific for one species (Figure 3). The *L. bicolor* genome contained a considerably higher number of lineage-specific families (1077 families) compared to the other species (Figure 3). The average size of the *L. bicolor* lineage-specific families (4.08 proteins / family) was larger than the average size of the *L. bicolor* families containing members from all basidiomycetes (2.74 proteins / family). Furthermore, the average size of the lineage-specific families in *L. bicolor* was larger than the size of the lineage-specific families identified in the other species (Figure S1, Supplementary material).

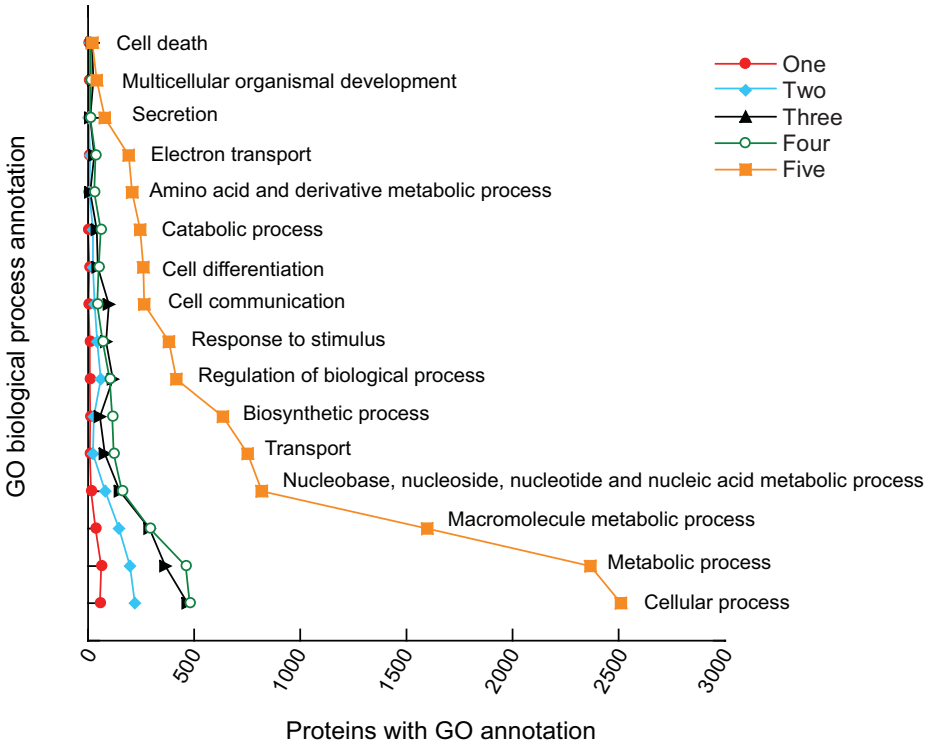


Figure 4. Functional annotations of proteins in families of *Laccaria bicolor*. Protein sequences predicted from the genome sequences of *Laccaria bicolor* were annotated into various biological processes based on homology searches against the SwissProt database. Assignments to biological processes were done based on the Gene Ontology (GO) classifications. The number of proteins with GO annotations in families represented in one, two, three, four or five genomes was counted (c.f. Figure 3). Shown are 16 biological processes having at least 20 annotated proteins. In total, 4577 proteins in *L. bicolor* were assigned to biological processes. The numbers of proteins having GO annotations for families represented in one, two, three, four and five genomes were 74, 254, 537, 594 and 3118, respectively.

The number of *L. bicolor* proteins that could be assigned to a GO annotation was highest in universal protein families shared between all basidiomycetes in comparison to those present in one (i.e. lineage-specific), two, three, or four genomes (Figure 4). The lineage-specific families (1077) contained 4398 proteins. A majority of them (70.4%) did not show any significant homology to sequences present in the fungal genome database of the Broad Institute or the NCBI nr protein database.

Evolution of protein family size in basidiomycetes

Protein families that were significantly expanded or contracted in the five basidiomycete genomes were identified using the CAFE tool [16]. The analysis showed that the largest number of protein families that showed an expansion have occurred along the branch leading to *L. bicolor*. The number of protein families that had expanded in size along the *L. bicolor* branch was estimated to 1064, and along the branches leading to *C. cinerea*, *P. chrysosporium*, *C. neoformans*, and *U. maydis* 459, 371, 307 and 96, respectively (Table 2).

Table 2. The number of gene families that showed expansion, no change, contractions, or extinctions along the branches of the phylogenetic tree of basidiomycetes[a]

No.	Branch	Divergence time (Mya)	Expansions	No change	Contractions	Extinctions	Average Expansion
1	*Lac/Copr/Phan/Cryp*	246	109	5248	26	0	0.036
2	*Lac/Copr/Phan*	167	426	4873	22	62	0.178
3	*Lac/Copr*	57	393	4855	47	88	0.130
4	*Lac*	84	1064	3844	112	363	0.695
5	*Copr*	84	459	4111	329	484	0.056
6	*Phan*	140	371	3291	277	1444	-0.169
7	*Cryp*	308	307	2272	304	2500	-0.519
8	*Ust*	554	96	2043	373	2871	-0.655

[a] A total of 5383 protein families present in at least two genomes were analyzed using the CAFE program [16]. Of these families, 4870 were represented in *L. bicolor*, 4749 in *C. cinerea*, 3877 in *P. chrysosporium*, 2833 in *C. neoformans* and 2512 in *U. maydis*. The branch numbers and lengths, calculated as divergence times (Mya), are shown in Figure 1A. The last column shows the average gene family expansion among all families along each branch, where a contraction is counted as a negative expansion.

Conversely, the numbers of contracted and extinct protein families were lower in the *L. bicolor* branch as compared to the branches of the other basidiomycetes. The CAFE analysis revealed 112 contracted and 363 extinct protein families in the *L. bicolor* lineage. In comparison, the branch of the biotrophic plant pathogen *U. maydis* contained 373 contracted and 2871 extinct gene families (Table 2). A statistical analysis showed that 11 protein families were significantly (P <0.001) contracted and 2 families were extinct in the *L. bicolor* lineage. The two extinct families were significantly expanded in *C. cinerea* lineage and had representatives in *P. chrysosporium* but were absent in *C. neoformans* and *U. maydis* (Table S1, Supplementary material). Members of these two families showed sequence similarities to a protein kinase domain and a fungal cellulose binding domain, respectively. The Pfam domains identified among the contracted protein families included cytochrome P450, ankyrin repeat, GMC oxidoreductase, glycosyl hydrolase family 7, and fungal cellulose binding domain (Table S1, Supplementary material).

Large protein families in* L. bicolor *that are expanded or lineage-specific

A large fraction of the recent duplicated genes in *L. bicolor* belonged to a limited number of large protein families that showed significant expansion or were lineage-specific. Thus, 523 out of the 1824 young duplicates were found in 55 large families containing more than 25 proteins (Table 3). In total, these families contained 3448 proteins. Similarity searches showed that 75.6% proteins in the large families showed significant similarity to sequences in the fungal genome database (Broad Institute). Searches in the Pfam database showed that 19 out of the 55 families had members displaying significant sequence homologies to protein domains. Several of the domains have roles in protein-protein interactions (NACHT, WD 40 and ankyrin repeats) and signal transduction mechanisms (protein kinases and GTPases) (Table 3).

The expression patterns of genes in the 55 large families were examined using data from 11 microarray experiments encompassing three different tissues - mycelia, mycorrhizal root tips and fruiting body. Based on a principal component analysis (PCA), 18 out of 55 protein families showed differential expression in mycorrhizal root tips as compared to fruiting bodies and mycelia (Figure 5) (Table 3).

Table 3. Large protein families in *Laccaria bicolor* that showed significant expansion, or were lineage-specific[a]

Protein family	Number of proteins[b]					Gene expression[c]	Duplications[d]			Pfam accession[d]	Pfam description
	Lac	*Copr*	*Phan*	*Cryp*	*Ust*	*Lac*	Tandem	Segmental	Recent		
Expanded in *Lac*											
1	**216**	97	91	75	74	163*	2	7	22	PF00400	WD domain, G-beta repeat
2	**150**	113	109	86	74	144*	3		2	PF00069, PF07714	Protein kinase domain, Protein tyrosine kinase
22	**102**	13	2	1	0	71	4		4		
30	**97**	1	0	0	0	50*	16		32	PF00023	Ankyrin repeat
9	**96**	49	0	0	0	43			17		
25	**91**	11	6	1	0	25	4		42		
32	**91**	1	0	0	0	58	2		16		
23	**86**	26	3	0	0	71*	2	4	8	PF05729	NACHT domain
42	**73**	3	0	0	0	41	2		4		
6	**59**	30	27	24	25	41	4		19	PF00071, PF08477, PF01926	Ras family, Miro-like protein, GTPase of unknown function
49	**58**	4	3	2	1	24			9	PF00467	KOW motif
56	**57**	1	3	0	0	41*	2		12	PF01926, PF08477	GTPase of unknown function, Miro-like protein
60	**56**	1	0	2	0	19			18		
65	**54**	0	2	0	0	34*		2	11		
62	**54**	0	3	0	0	42			4		

Protein family	Number of proteins[b]					Gene expression[c]	Duplications[d]			Pfam accession[d]	Pfam description
	Lac	*Copr*	*Phan*	*Cryp*	*Ust*	*Lac*	Tandem	Segmental	Recent		
44	**53**	19	1	0	0	41*	5		6		
36	**52**	28	2	2	0	17			8		
35	**48**	28	10	0	0	44	6		4	PF01926, PF04548, PF00350	GTPase of unknown function, AIG1 family, Dynamin family
69	**47**	6	1	0	0	18	2		8		
38	**44**	28	8	0	0	18			8		
75	**41**	5	2	1	1	29*			4	PF00867, PF00752	XPG I-region, XPG N-terminal domain
98	**41**	1	0	0	0	26			4		
72	**41**	4	4	0	2	10			20		
50	**40**	11	2	0	13	29			5	PF03184	DDE superfamily endonuclease
103	**39**	1	1	0	0	26		2	8		
71	**38**	8	5	0	0	26*			2		
91	**38**	2	5	0	0	15			11	PF05699	hAT family dimerisation domain
81	**34**	7	6	0	0	28					
119	**34**	0	0	1	0	7			23		
107	**33**	4	2	0	0	21*			4	PF07714, PF00069	Protein tyrosine kinase, Protein kinase domain
78	**32**	10	7	0	0	28*	6				
67	**32**	13	9	0	0	20*		4	11	PF00503, PF00025	G-protein alpha subunit, ADP-ribosylation factor family

Protein family	Number of proteins[b]					Gene expression[c]	Duplications[d]			Pfam accession[d]	Pfam description
	Lac	*Copr*	*Phan*	*Cryp*	*Ust*	*Lac*	Tandem	Segmental	Recent		
76	**32**	9	3	4	2	21			7	PF00271, PF00270, PF04851	Helicase conserved C-terminal domain, DEAD/DEAH box helicase, Type III restriction enzyme, res subunit
100	**31**	1	9	0	0	25*	4		6		
59	**31**	12	16	0	0	26*	2		4	PF01753	MYND finger
114	**31**	5	0	0	0	10			6		
93	**29**	9	6	0	0	20*			2		
89	**29**	12	4	0	0	23			4	PF05699	hAT family dimerisation domain
129	**29**	3	0	0	0	5		2	16		
66	**28**	18	7	1	1	26	8				
132	**27**	4	0	0	0	19			6		
127	**27**	1	1	1	2	10	2		4	PF03987, PF03986	Autophagocytosis associated protein, C-terminal domain, Autophagocytosis associated protein, N-terminal domain
147	**27**	1	1	0	0	25	2		2		
122	**27**	2	2	1	2	12			10		
149	**27**	0	2	0	0	6			15	PF00098	Zinc knuckle
118	**26**	1	7	1	0	18			2		

Protein family	Number of proteins[b]					Gene expression[c]	Duplications[d]			Pfam accession[d]	Pfam description
	Lac	*Copr*	*Phan*	*Cryp*	*Ust*	*Lac*	Tandem	Segmental	Recent		
Lineage-specific in *Lac*											
5	**206**	0	0	0	0	110*	4	10	31	PF00931, PF05729	NB-ARC domain, NACHT domain
17	**128**	0	0	0	0	116*	26	2	8		
64	**56**	0	0	0	0	35			14		
88	**46**	0	0	0	0	32		2	4		
120	**35**	0	0	0	0	21*	2				
130	**32**	0	0	0	0	26			3		
143	**30**	0	0	0	0	20			2		
148	**29**	0	0	0	0	9	2				
165	**27**	0	0	0	0	12					

[a]Listed are 55 protein families containing more than 25 proteins in *L. bicolor* that were either significantly (P <0.001) expanded along the *L. bicolor* branch (branch 4 in Figure 1), or present only in the *L. bicolor* genome (lineage-specific, c.f. Figure 3). In total, *L. bicolor* contains 71 protein families having more than 25 members.

[b]Number of proteins in respective protein family - *Laccaria bicolor (Lac)*, *Coprinopsis cinerea (Copr)*, *Phanerochaete chrysosporium (Phan)*, *Cryptococcus neoformans (Cryp)* and *Ustilago maydis (Ust)*.

[c]Data from microarray experiments of *L. bicolor* (*Lac*). The numbers in the column show the total number of gene models in each protein family that could be recognized with at least one specific 60-mer probe. (*) indicates 18 protein families showing differential expression levels in mycorrhizal root tips as compared to fruiting bodies and mycelia. The statistical significance at P <0.001 was analyzed by ANOVA on principal component 1 scores using data from 11 microarray experiments (c.f. Figure 5).

[d]Number of genes in the *L. bicolor* protein families that were present as tandem, segmental, and recent (ds<0.2) duplicates.

[e]The putative protein domains search was performed for *L. bicolor* proteins against the Pfam-LS database [17] with E-value threshold of 0.05. Domains present in more than 25% of the members in the protein family are shown.

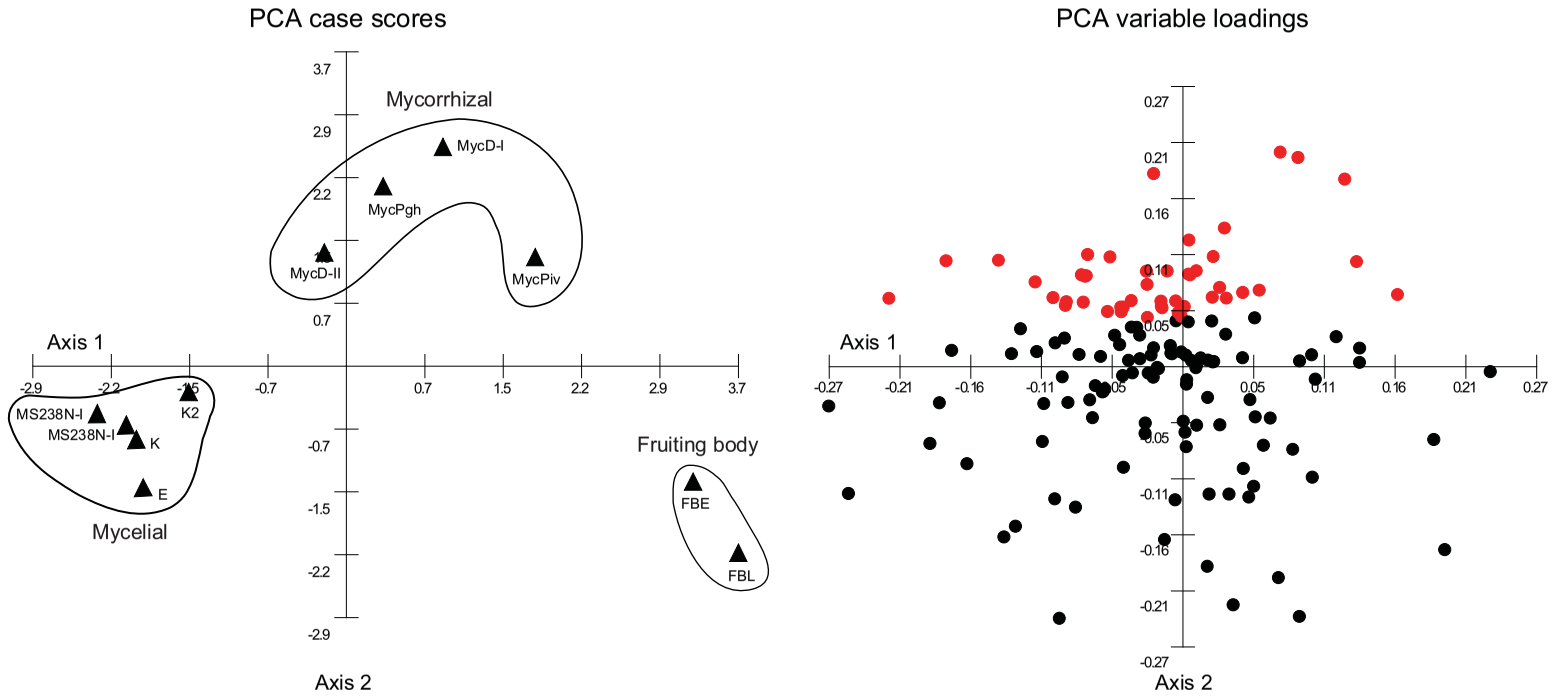


Figure 5. Principal component analysis (PCA) on the expression levels of *L. bicolor* genes in Family-2. Members in this family contained sequences with significant similarity to the protein kinase domain. Expression levels (log_2 values) were retrieved from 11 microarray experiments including four mycorrhizal (MycD-I, MycD-II, MyPgh, MycPiv), two fruiting body (FBE, FBL) and five mycelial (MS238N-I, MS238N-II, K, K2, E) samples. Family-2 contained 150 members (c.f. Table 3), of which 144 could be analyzed using the microarray after removing signals from cross-hybridizing probes. The PCA scatter plot shows the case scores and loadings scores of principal components 1 and 2. The components 1 and 2 accounted for 61% of the variation. The symbol (●) indicates 46 genes located above 0.05 (loading score) on Axis 2, which were considered as significantly expressed in mycorrhizal root tips.

Evolution of protein kinases

The evolution of two gene families that were significantly expanded in *L. bicolor* was examined in more detail. The first family was Family-2 which contained sequences with homology to the protein kinase domain PF00069 (Table 3). This conserved catalytic domain defines a single superfamily of protein kinases that carries out protein phosphorylation in almost all eukaryotes [18].

In the analyzed basidiomycetes, the PF00069 domain was found in 41 protein families with 1120 proteins. To verify the clustering done by the TRIBE-MCL algorithm and to examine the relationship between Family-2 and other PF00069-containing families, a weighted graph of the families based on their sequence similarities were generated (Figure 6A). Clearly, the large families were found in cohesive clusters in the obtained network. Family-2 is the largest family with 512 basidiomycete sequences (140 in *L. bicolor*) and was located central in the network connecting to all of the other families. Among the families visualized in this network was also Family-24, which was extinct in *L. bicolor* (Table S1, supplementary material). The network properties of all the basidiomycete protein families showed a power-law degree distribution (data not shown). The average connectivity of the protein kinase superfamily was 238 and the network diameter was nine showing that despite the clear division into smaller families, the protein kinase superfamily is highly conserved. The maximum connectivity (562) was found among the proteins in Family-2.

Analysis of protein kinases in several model organisms including *S. cerevisiae* have shown that these proteins can be clustered into a number of groups and subgroups with conserved functions [18]. To classify the protein kinases of Family-2 into such groups, a phylogenetic tree was reconstructed, which also included sequences of yeast protein kinases (Figure 6B). In total, 77 subgroups of protein kinases were identified (Table 4). Forty-six of them contained sequences of yeast protein kinases ("Orthologs in *S. cerevisiae*"), 31 subgroups did not contain any yeast homolog ("No orthologs in *S. cerevisiae*"), and a few sequences were not clustered into any subgroups ("Not clustered").

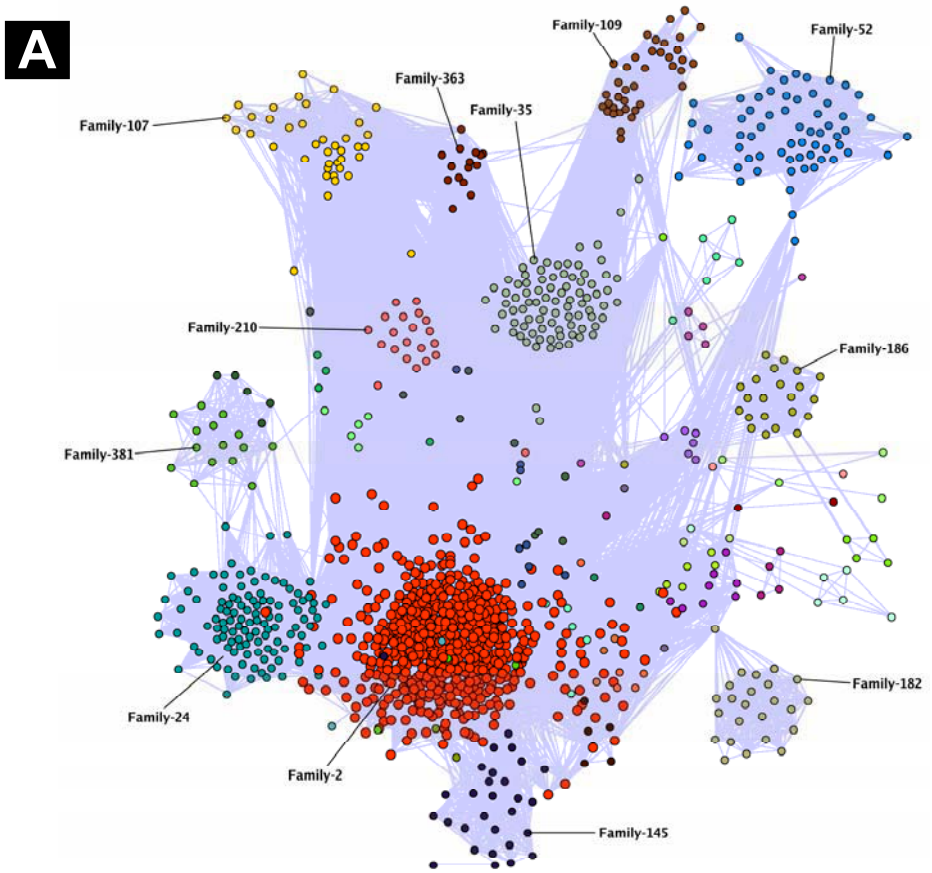
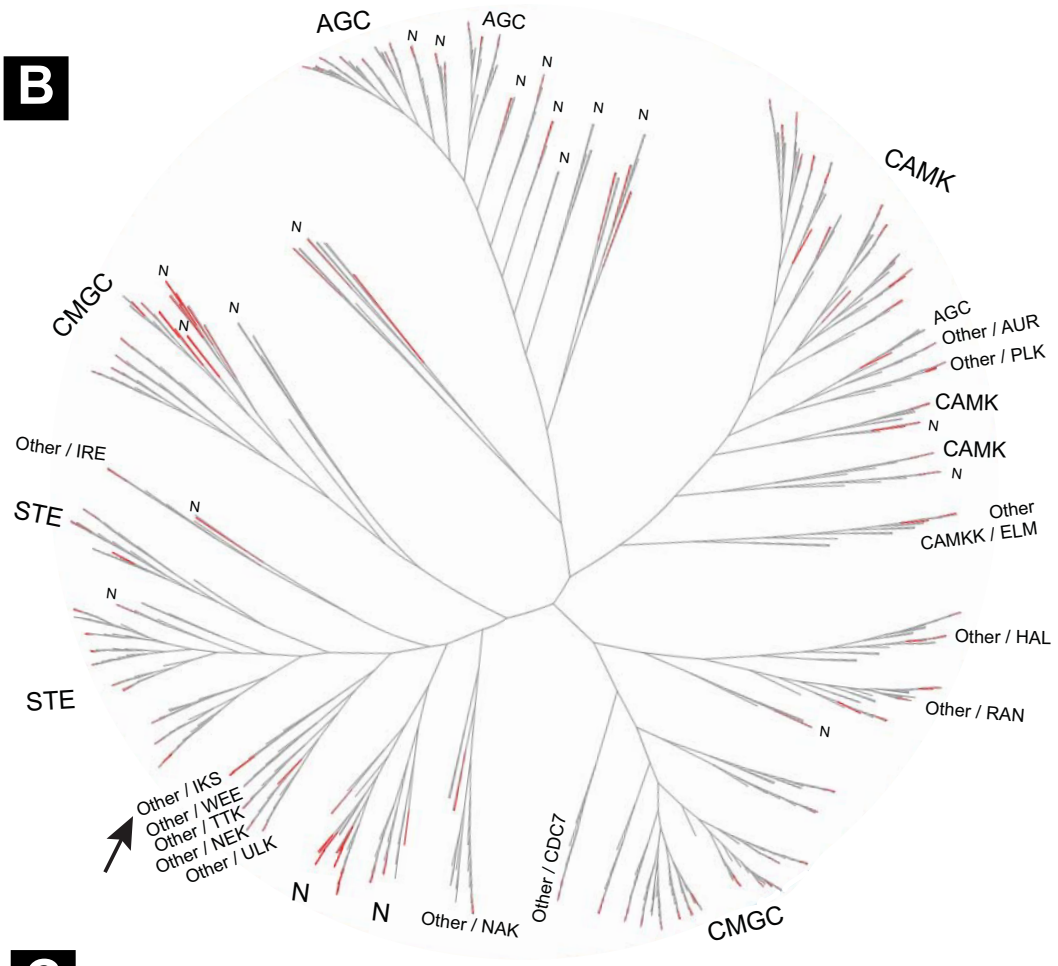


Figure 6. Evolution of the protein kinase superfamily. (A) A weighted graph representing proteins belonging to the protein kinase superfamily in the genomes of *Laccaria bicolor*, *Coprinopsis cinerea*, *Phanerochaete chrysosporium*, *Cryptococcus neoformans* and *Ustilago maydis*. The graph is based on the analysis of all protein families in these genomes containing sequences with homology the Pfam domain PF00069 (41 families with 1120 proteins). Following a sequence similarity search (BlastP), the graph was generated using the Biolayout graph layout algorithm [19]. The protein families were mapped onto the graph (coloring of the nodes) and twelve of the largest protein families (≥15 members) were labeled with family identies. Family-2 is the largest family and it is also significantly expanded in *L. bicolor* (Table 3). (B) Phylogeny of protein kinases in Family-2. The displayed hyperbolic tree (log_2-transformed) [20] was derived from a Neighbor-Joining (NJ) tree, which was constructed using the PF00069 domains from 512 sequences of Family-2 and 121 yeast protein kinases. Terminal branches of the 140 *L. bicolor* sequences are shown in red and those of other species in gray. The protein kinases were classified into a number of sub-groups corresponding to clades identified in the NJ tree (c.f. Table 4). "N" denotes clusters that do not contain any yeast protein kinase. ⟶



(C) Expansion of *L. bicolor* protein kinases within the "Other/IKS1 sub-group". A portion of the NJ tree from Panel B is enlarged (indicated by arrow). The bootstrap support was estimated from 100 replicates. The subgroup contains three *L. bicolor* paralogs (red branches). The symbol "▲" indicates two of the paralogs that were recently duplicated (*ds* <0.2) and that were expressed above background levels.

Table 4. Classification of protein kinases in Family-2[a]

Group	Sub-groups	Number of proteins					
		Lac	Mycorrhizal expression[b]	*Copr*	*Phan*	*Cryp*	*Ust*
Orthologs in *S. cerevisiae*							
AGC	6	12	4	11	11	9	9
CAMK	13	16	8	14	14	13	9
CMGC	9	22	6	20	22	20	17
Other	13	25	8	17	20	18	13
STE	5	15	8	16	12	13	13
No orthologs in *S. cerevisiae*							
Lac	6	16	4	0	0	0	0
Copr	2	0	0	4	0	0	0
Phan	2	0	0	0	7	0	0
Lac/*Copr*	6	6	0	6	0	0	0
Lac/*Phan*	2	2	0	0	2	0	0
Lac/*Copr*/*Phan*	3	5	2	4	3	0	0
Lac/*Copr*/*Phan*/*Cryp*	1	1	1	1	1	1	0
Lac/*Copr*/*Phan*/*Cryp*/*Ust*	5	6	2	6	6	5	5
Other combinations	4	3	1	5	3	0	3
Not clustered		11	2	5	6	5	3
Total	**77**	**140**	**46**	**109**	**107**	**84**	**72**

[a]In total, Family-2 contains 512 proteins containing the protein kinase domain PF00069. These proteins were classified into a number of sub-groups based on a phylogeny (Neighbor-Joining tree) of the PF00069 domain. A sub-group corresponds to a clade of sequences having a bootstrap support value ≥ 50. The phylogenetic distribution of these subgroups is shown in Figure 6B. "Orthologs in *S. cerevisiae*" includes sub-groups (clades) containing sequences of yeast kinases. AGC includes cyclic-nucleotide and calcium-phospholipid-dependent kinases, ribosomal S6-phosphorylating kinases, G protein-coupled kinases, and all close relatives of these groups; CAMK comprises of calmodulin-regulated kinases; CMGC includes cyclin-dependent kinases, mitogen-activated protein (MAP) kinases; Other consists of kinases not classified into any of the above groups; STE includes many kinases functioning in the MAP kinase cascades (For a description of these subgroups see http://kinase.com/). "No orthologs in *S. cerevisiae*" comprises subgroups that do not contain any yeast kinase homologs. The "Not clustered" protein kinases include those that are located outside any of the clades identified in the NJ tree. *Laccaria bicolor (Lac)*, *Coprinopsis cinerea (Copr)*, *Phanerochaete chrysosporium (Phan)*, *Cryptococcus neoformans (Cryp)* and *Ustilago maydis (Ust)*.

[b]The number of genes considered to be significantly regulated in the mycorrhizal root tips based on a PCA analysis (c.f. Figure 5).

Expansion of protein kinases in *L. bicolor* has mainly occurred among the "Not clustered" proteins and in subgroups lacking yeast orthologs (Table 4). In the latter category, 19 paralogs were identified in five subgroups. We define paralogs as gene duplicates evolved after the speciation event which corresponds to inparalogs as proposed by Sonnhammer and Koonin [21]. Only paralogs with the lineage-specific branch having a bootstrap value of 50% or higher was considered in the analysis. Expansion of protein kinases in *L. bicolor* was also observed in two of the subgroups having orthologs in *S. cerevisiae*, "Other/IKS1" (3 paralogs) (Figure 6C) and "Other-PLK" (2 paralogs).

Analysis of the microarray data showed that 136 out of 140 protein kinases in *L. bicolor* were expressed above the background level (Table 4). Forty-six of these genes were differentially expressed in the mycorrhizal root tips. A majority (34) of them belonged to the yeast superfamilies of protein kinases whereas the remaining 12 were found in the "No orthologs in *S. cerevisiae*" and "Not clustered" group.

Evolution of Ras GTPases

The second significantly expanded protein family in *L. bicolor* that was examined in detail was Family-6 containing proteins with significant homology to the Ras family of small GTPases (Pfam PF00071) (Table 3). The Ras small GTPases are divided into five subfamilies, the Ras, Rho, Rab, Ran and Arf [22, 23]. The Ras, Rho, Rhab and Ran subfamilies contain the Ras domain (Pfam PF00071) whereas the Arf subfamily contains the Arf domain (Pfam PF00025) and was therefore not included in the analysis. In *L. bicolor*, the Ras domain is found in 61 proteins and 55 of them are present in Family-6. Based on a phylogenetic analysis, the small GTPases of Family-6 were classified into 29 subgroups (Figure 7A). Fourteen of these subgroups contained orthologs to the Ras, Rho, Rab and Ran GTPases in *S. cerevisiae* (Table 5).

In total, we identified 26 *Laccaria* paralogs in 6 subgroups of the GTPases. One cluster of four paralogs was found in the RAS1/RAS2 subgroup (Figure 7B). The four *Laccaria* paralogs in the RAS1/RAS2 subgroup form two-pairs of young duplicates (*ds* <0.2). One pair was expressed above background level. Members in the other pair had

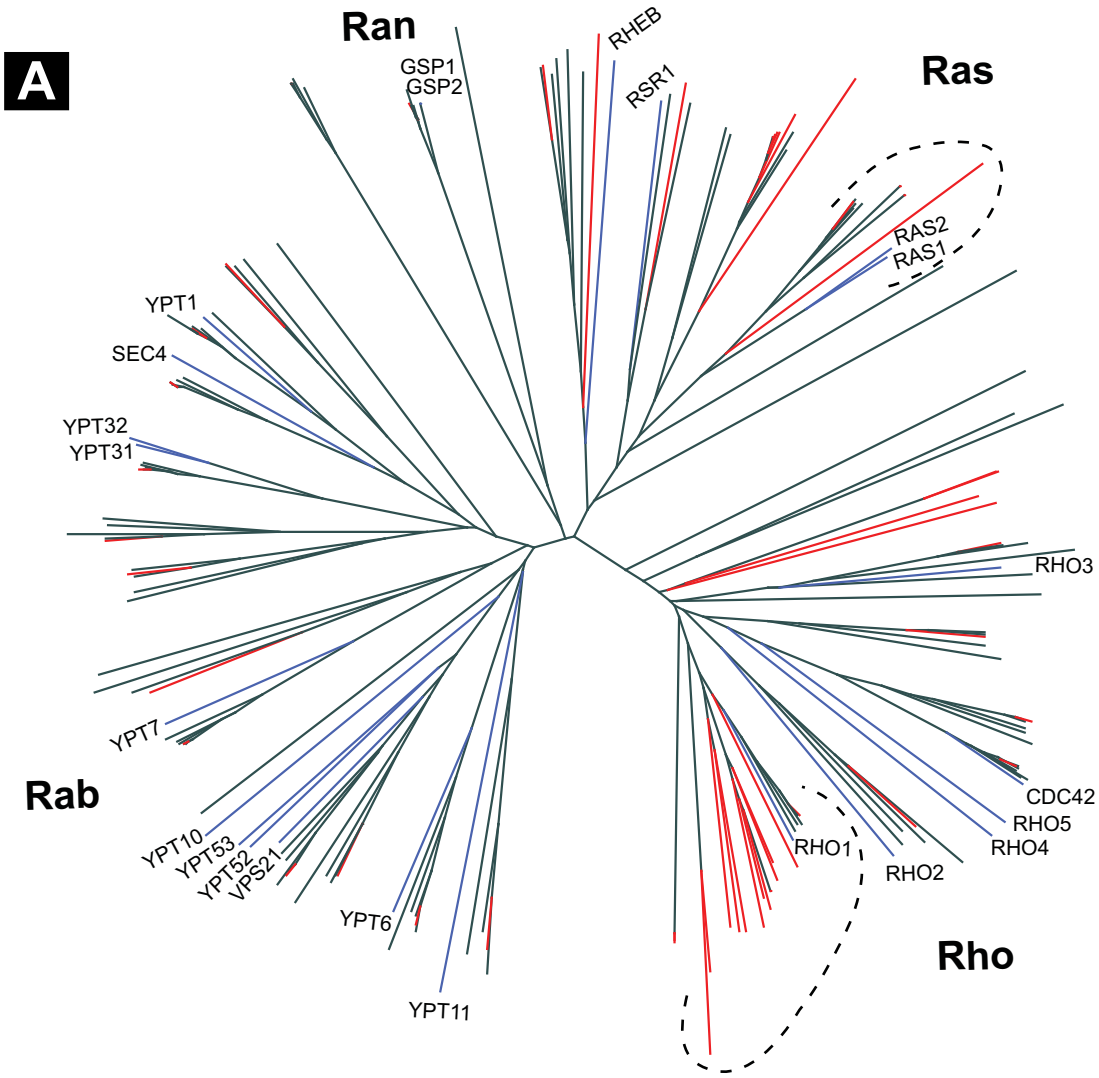


Figure 7. Phylogeny of Ras GTPases from five basidiomycetes. (A) An unrooted, radial Neighbor-Joining (NJ) tree showing the phylogeny of the Ras domain PF00071 from 160 proteins in Family-6, and 23 *S. cerevisiae* proteins. Family-6 contains members from all of the analyzed basidiomycetes, and is significantly expanded in *L. bicolor* (c.f. Table 3). The terminal branches of the 55 *L. bicolor* sequences are shown in red lines, 23 *S. cerevisiae* sequences in blue and remaining sequences in black lines. (B) Expansion of the Ras1/Ras2 subgroup of small GTPases in *L. bicolor*. Shown is part of the NJ tree displayed in Panel A. The subgroup contains six *L. bicolor* sequences (red branches). Four of them were recently duplicated (*ds* <0.2) paralogs labelled with the symbols "▲" (expressed above background level), and "▲" (hybridization signals below background level). *L. bicolor* proteins labelled in boldface are differentially expressed according to PCA analysis. (C) A large cluster of 10 paralogs in *L. bicolor* adjacent to the Rho1 subgroup of ras GTPases. Shown is part of the NJ tree displayed in Panel A. The same symbols are used as in Panel B.

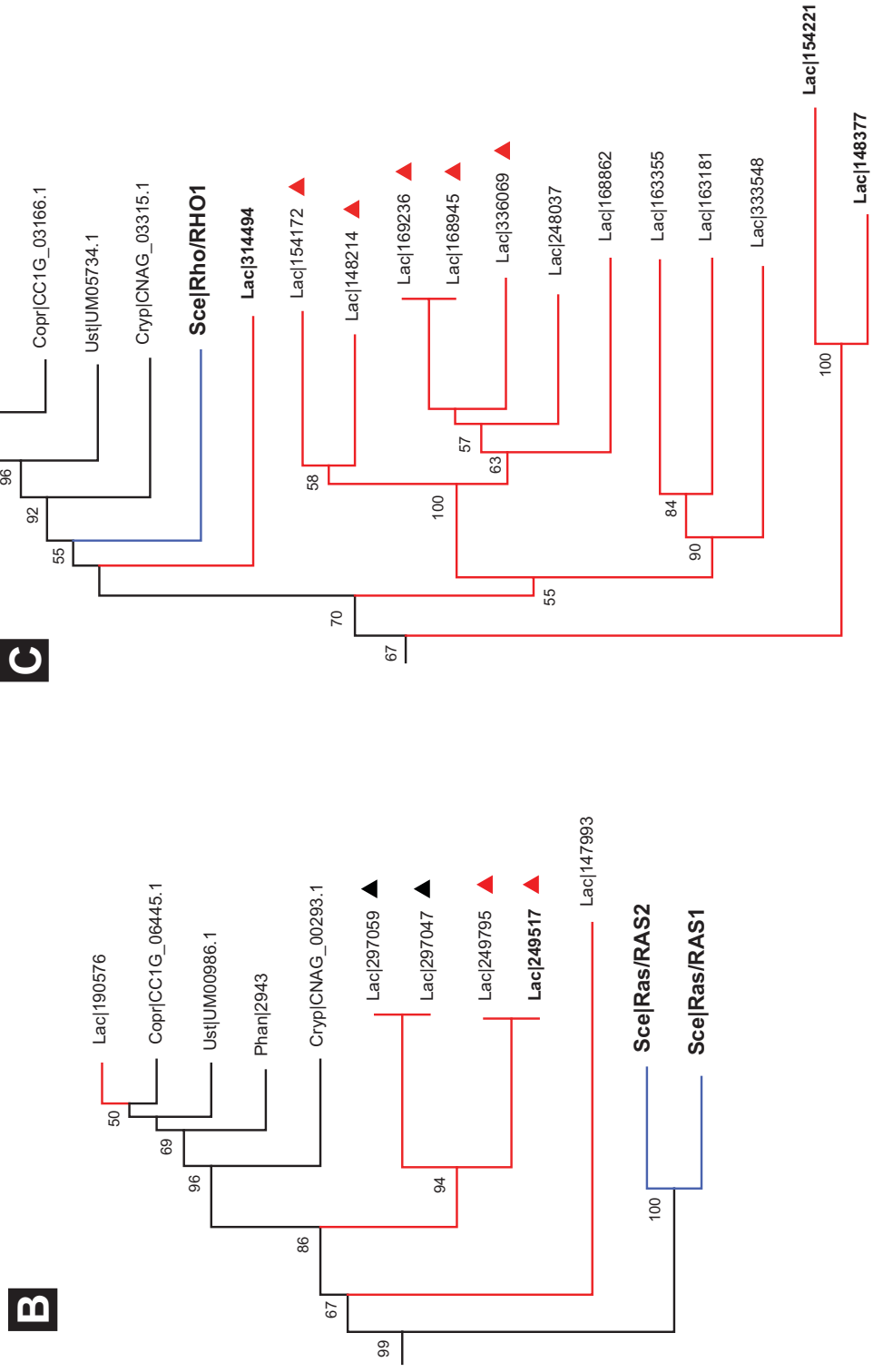

C
Phan|131276
Lac|175325
Copr|CC1G_03166.1
Ust|UM05734.1
Cryp|CNAG_03315.1
Sce|Rho/RHO1
Lac|314494
Lac|154172
Lac|148214
Lac|169236
Lac|168945
Lac|336069
Lac|248037
Lac|168862
Lac|163355
Lac|163181
Lac|333548
Lac|154221
Lac|148377
85
100
96
92
55
58
100
57
63
84
90
55
70
67
100
B
Lac|190576
Copr|CC1G_06445.1
Ust|UM00986.1
Phan|2943
Cryp|CNAG_00293.1
Lac|297059
Lac|297047
Lac|249795
Lac|249517
Lac|147993
Sce|Ras/RAS2
Sce|Ras/RAS1
50
69
96
86
94
67
99
100

Table 5. Classification of small GTP-binding proteins in Family-6[a]

Group	Sub-groups	Number of proteins				
		Lac	*Copr*	*Phan*	*Cryp*	*Ust*
Orthologs in *S. cerevisiae*						
Ras						
RAS1/RAS2	1	6	1	1	1	1
RSR1	1	1	0	1	0	1
RHEB	1	2	2	1	1	1
Rho						
RHO1	1	1	1	1	1	1
RHO2	1	1	1	1	1	1
RHO3	1	1	1	1	1	1
CDC42	1	1	1	1	2	1
Rab						
YPT1	1	1	1	2	1	1
YPT6	1	1	1	1	1	1
YPT7	1	1	2	1	1	1
YPT31/YPT32	1	1	1	1	1	1
SEC4	1	1	1	1	1	1
VPS21	1	1	1	1	1	1
Ran						
GSP1/GSP2	1	1	1	1	1	1
No orthologs in *S. cerevisiae*						
Lac	4	16	0	0	0	0
Copr	1	0	2	0	0	0
Lac/Copr/Phan/Ust	3	3	3	3	0	3
Lac/Copr/Phan/Cryp/Ust	7	13	7	7	8	7
Not clustered		3	2	2	2	1
Total	**29**	**55**	**29**	**27**	**24**	**25**

[a]In total, Family-6 contains 160 proteins containing the Ras domain PF00071. These proteins were classified into a number of sub-groups based on a phylogeny (Neighbor-Joining tree) of the PF00071 domain (Figure 7). A sub-group corresponds to a clade of sequences having a bootstrap support value ≥ 50. The phylogenetic distribution of these sub-groups is shown in Figure 7A. "Orthologs in *S. cerevisiae*" includes sub-groups (clades) containing sequences of small GTP-binding proteins from yeast. Ras superfamily is broadly classified into five subfamilies (sub-groups), Ras (rat sarcoma), Rho (Ras homologs), Rab (rat brain), Arf (ADP ribosylation factors) and Ran (Ras-related nuclear) [22]. The Arf family (Pfam PF00025) is not represented in Family-2 but are related to Ras family (Pfam PF00071) by a common clan of G-protein superfamily (clan id : CL0017, a clan contains two or more Pfam families that have arisen from a single evolutionary origin) [17]. "No orthologs in *S. cerevisiae*" comprises sub-groups that do not contain any yeast small GTP-binding proteins.

hybridization signals below the background level and are presumably pseudogenes. They were the only genes among the 55 genes in Family-6 with hybridization signals below the background level. Even though the Family-6 was not significant in the ANOVA analysis for axis 1, the PCA plot (Principal component 1 and 2) separates the mycorrhizal experiments from the fruiting body and mycelia (data not shown). One of the proteins (Protein id 249517) was differentially expressed in mycorrhizal root tips based on the PCA analysis and is shown in Figure 7B.

The largest group of paralogs in Family-6 consisted of a cluster of 10 sequences adjacent to the Rho1 subgroup (Figure 7C). The cluster could not unambiguously be assigned to any of the *S. cerevisiae* proteins due to low bootstrap values. Five out of the 10 paralogs were young duplicates (ds<0.2).

Discussion

In this study, we have analyzed gene gain and loss at the genomic scale by studying the expansion and contraction of gene families in the whole genomes of five basidiomycetes. Large differences were observed in the family sizes. This variation can either be due to random processes like genetic drift and mutations, or be the result of adaptive processes [24]. To account for the differences due to random processes, the data were analyzed in a statistical framework that provides an estimate on the variation in gene family sizes when gains and losses occur randomly [16]. As a result, it is possible to identify the branches in a phylogenetic tree where larger-than-expected expansions or contractions occur that potentially indicate the action of natural selection.

←

The "Not clustered" small GTP binding proteins include those that are located outside any of the clades identified in the NJ tree. Six of the small GTP-binding proteins in yeast (YPT53, YPT10, YPT11, YPT52, RHO4 and RHO5) were found outside the clusters. *Laccaria bicolor (Lac)*, *Coprinopsis cinerea (Copr)*, *Phanerochaete chrysosporium (Phan)*, *Cryptococcus neoformans (Cryp)* and *Ustilago maydis (Ust)*.

Our analysis showed that *L. bicolor* genome has gained a large number of genes that are not present in the genomes of saprophytic and parasitic basidiomycetes. Many of these novel genes are found in large multigene families that were significantly expanded in the *Laccaria* branch or in families that are unique to this fungus. Moreover, the *L. bicolor* genome has retained a majority of the gene families supposed to be present in the common ancestor of basidiomycetes, and the genome contains a majority (5947 out of 7352) of the extant basidiomycete gene families. These findings suggest that *L. bicolor* has acquired numerous genes needed for the symbiotic interactions with the host plants but also kept many of the genes needed for saprophytic growth. Phylogenetic analyses have indicated that there have been several switches from symbiotic to free-living, saprophytic, life styles in basidiomycetes [4]. The fact that the *L. bicolor* genome has not lost significant portions of the proteins present in saprophytic species suggests that such reversals can occur without the need for gaining large sets of novel genes.

Changes in genome repertoire occurring through gene acquisition and deletion have been shown to be the major events underlying the emergence and evolution of bacterial symbionts [25]. Two contrasting evolutionary trends have been identified. The massive reductions in genome size that are found in obligate and ancient symbionts, and the expansion of genomes that are typically for recent and facultative symbionts [26]. Thus, the expansion of gene content observed in the EM fungus *L. bicolor* is similar to that observed in facultative symbiotic bacteria like *Bradyrhizobium japonicum* [27], *Sinorhizobium meliloti* [28] and *Mesorhizobium loti* [29].

The expansion of gene families in *L. bicolor* is due to a large number of small single duplication events rather than whole genome or large segmental duplications (Martin *et al.*, manuscript). A majority of the duplicated genes are dispersed in the *Laccaria* genome and only a small number of them are located in tandem or in segments. Compared to *L. bicolor*, the other filamentous basidiomycetes contained a lower number of gene duplicates. However, the total numbers of tandem duplicates were similar (244 in *L. bicolor*, 214 in *C. cinerea*, 233 in *P. chrysosporium*, Martin *et al.*, manuscript). Most probably, the large numbers of transposable elements detected in *L. bicolor* (Martin *et al.*,

manuscript) have had a major impact on the duplication events by inducing rearrangements in the genome and thereby generating novel gene duplications, and disruption of tandem duplicates.

A considerable part of the identified gene duplicates in *L. bicolor* appear to be the result of recent duplication events (*ds*<0.2) (Fig. 2). Almost all (98%, 1788 out of 1824) of these young duplicates belonged to the lineage-specific or expanded families, particularly to large-sized gene families (Table 3). A majority of these protein families did not contain any known protein domains and the functions of them are not known. However, the fact that many of the genes in these families were differentially expressed during interaction with the host plant suggests that the expansion of the gene families have evolved in response to selection for symbiotic growth.

One of the largest expanded gene families in *L. bicolor* was Family-2 that contains 150 members, and with a majority of them containing a protein kinase domain. Protein phosphorylation controls many cellular processes including stress responses, cell cycle, development and differentiation [18]. Our phylogenetic analysis showed that 26 new duplicates (paralogs) have arisen in Family-2 along the *L. bicolor* branch since its divergence from *C. cinerea*. The paralogs were found in 9 clusters and only two of them were found in subgroups having orthologs in *S. cerevisiae*, namely IKS1 and PLK (Polo-Like Kinase). IKS1 encodes a putative serine/threonine kinase with unknown function. IKS1 mutants are hypersensitive to copper sulphate and resistant to sorbate [30]. PLK, encoded by CDC5, has multiple functions in mitosis and cytokinesis [30]. A majority (136) of the genes in Family-2 was expressed above background levels and 46 of them were differentially expressed in mycorrhizal root tips.

The second family that was studied was Family-6 containing 55 proteins with the domain of Ras small GTPases. This is a group of highly similar proteins which functions as versatile molecular switches in diverse processes including signal transduction, cell polarity, cytoskeleton regulation, and vesicle trafficking [31]. It has previously been suggested that the Ras protein (Lbras) in *L. bicolor* may be involved in signal transduction pathways responding to diverse external stimuli such as for example hormones, growth factors and other elicitors [32].

The Lbras is most similar to a protein (sequence id 147993, in Figure 7B).

Family-6 has rapidly expanded in *L. bicolor* with 26 paralogs that cluster in six different groups. One of the expanded groups contains Ras1/Ras2. RAS1 of *S. cerevisiae* is involved in G-protein signaling in the adenylate cyclase-activating pathway and plays a role in cell proliferation. RAS2 is a GTP-binding protein that regulates the nitrogen starvation response, sporulation, and filamentous growth [30]. The differential expression in mycorrhizal root tips of one of the recently duplicated Ras1/Ras2 paralogs in *L. bicolor* suggests a duplication event followed by an adaptation to a novel function either through subfunctionalization or neofunctionalization. However, the largest expansion occurred in a group of 10 recent paralogs forming a sister clade to the Rho1 subgroup (Figure 7C). The Rho1 is directly regulating the reorganization of the actin cytoskeleton as well as 1,3 β-glucan synthase which is a major component of the cell wall and further suggests the importance of a Rho1-like protein in the formation of mycorrhizal tissue in response to environmental signals.

A small number (87) of basidiomycete protein families were restricted to the symbiont *L. bicolor* and the biotroph *U. maydis*. A majority of these families are of small in size (1.4 *L. bicolor* proteins / family). Furthermore, only a small fraction of the proteins shared by *L. bicolor* and *U. maydis* display significant sequence similarity to proteins in other fungal genomes, including ascomycetes. Notably, four of the 87 families contain putative small secreted fungal proteins as defined by Martin *et al* (unpublished). However, the functions of these families are largely unknown. In addition, clustered secreted proteins in *U. maydis* [11] were examined and compared to the protein families in our study. In total, 32 protein families contained clustered secreted *U. maydis* proteins and out of these, five families contained exclusively *L. bicolor* and *U. maydis* proteins. Family-516 was the largest and was significantly expanded in the *L. bicolor* branch (one protein from *U. maydis* and 11 proteins from *L. bicolor*).

A limited number of the protein families were contracted or extinct in the *L. bicolor* genome. Among the significantly contracted families was Family-4 with members containing the cytochrome P450

domain. Proteins of this family are thought to be of importance for the degradation of various complex carbon sources in *P. chrysosporium* [10]. The *L. bicolor* P450 Family-4 contained 39 proteins, significantly higher number than in *C. neoformans* and *U. maydis* but less than in *C. cinerea* and *P. chrysosporium* reflecting different life styles of the species compared. Among the extinct families in the *L. bicolor* branch, 38 families had proteins from all other four basidiomycete species, suggesting that they may belong to core functions that have been lost in *L. bicolor*. One example of an extinct family in our study is the pectin lyase domains which was previously reported to be lacking from the genome of *L. bicolor* (Martin *et al*., manuscript) but is found as multigene families in *C. cinerea* and *P. chrysosporium* and is thought to be of importance for the pathogenicity in necrotrophic fungi [33].

One obvious limitation in our analysis of gene duplications and protein families is the limited number of genomes currently available for the basidiomycota crown group (i.e. taxon sampling problem). However, *C. cinerea* is closely related to *L. bicolor* and both belong to the order of Agaricales (family Psathyrellaceae and Hydnangiaceae, respectively). The Psathyrellaceae and Hydnangiaceae families are sister groups in the species-rich Agaricales genus [34] and thus *C. cinerea* provided us with a close non-symbiotic relative to *L. bicolor* which is crucial for the analysis. *P. chrysosporium* (Corticiales, Corticiaceae), *C. neoformans* (Tremellales, Tremellaceae) and *U. maydis* (Ustilaginales, Ustilaginaceae) are more distantly related to *L. bicolor* and the species cover approximately 550 million years of divergence [14]. However, the substitution rate in the basidiomycete crown group was shown to be significantly lower as compared to the ascomycetes [35]. The divergence in number of substitutions in this analysis is therefore similar to the protein family analysis of the hemiascomycete yeasts [36]. Since our analysis is not depending on the divergence time but rather on the relative distances between the species, the controversy on the divergence time [35] does not affect our analysis. Another limitation of the comparative genomics approach is that the differences identified in the *L. bicolor* lineage could be due to adaptations other than to a symbiotic life style. However, by correlating independent analyses such as differential expression- phylogenetic- and functional analyses with the comparative genomics results, candidate genes involved in EM symbiosis were identified.

This is the first study examining the genomic changes that could account for the evolution of the EM in basidiomycetes. Several studies have suggested that mycorrhizal symbiosis have evolved repeatedly in basidiomycetes [4,7,37]. Recently JGI funded a program to sequence the genome of *Paxillus involutus* which belongs to the Boletales, in which EM trait has evolved independently from that in Agaricales. Notably, the genome of *P. involutus* appears to be significantly smaller than that of *L. bicolor* [38]. Comparative analyses of the genomes of these two fungi will provide insights into common as well as unique mechanisms underlying the evolution of mutualism in the basidiomycetes.

Materials and Methods

Genomes used

Genome sequence data for the five basidiomycetes were extracted from the databases at the Joint Genome Institute [39] and the Broad Institute [40]. The 65 Megabase pair (Mbp) genome of *Laccaria bicolor* S238N-H82 encodes 20,614 predicted proteins (release version 1.0, March 2005; JGI), the genome of *Coprinopsis cinerea* (*Coprinus cinereus*) Okayama7#130 contains ~37 Mbp and 13,544 proteins (release 1, July 2003; Broad), the *Phanerochaete chrysosporium* RP78 genome contains 35 Mbp and 10,048 proteins (release version 2.0, Feb. 2005; JGI), the *Cryptococcus neoformans* H99 genome contains 19.5 Mbp and 7302 proteins (Assembly 1, May 2003 & Gene set 3.0, Feb. 2006; Broad) and the *Ustilago maydis* 521 genome contains 19.7 Mbp and 6522 proteins (Release 2, March 2004; Broad).

Identification of gene duplicates

Duplicated gene pairs located anywhere in the basidiomycete genomes were identified by bidirectional best-hit analysis using BlastClust program [41]. Thresholds of 60% and 90% sequence identity over 90% alignment lengths were used to make two protein datasets. The "high similarity" set contained protein pairs having sequence identity above 90%, and the "low-similarity" set comprised pairs with an identity between 60 to 90%.

To estimate the age of divergence of the duplicates, the rates of synonymous substitutions per silent site (*ds*) were determined for each

gene pairs. The proteins pairs with high similarity and low similarity duplicates for each genome were aligned using ClustalW. Subsequently, the Tranalign (EMBOSS, version 4.0.0) [42] was used to align coding nucleic acid sequences based on protein alignments. The *ds* for the aligned nucleotide pairs were estimated using the SNAP program [43]. Genes in gene-pairs with *ds* value below 0.2 were considered as recent duplicates.

Protein family identification

An all-against-all bidirectional sequence similarity search was performed on the entire set of basidiomycete proteins (58,030 proteins) using the BlastP program [44] with a threshold E-value of 1e-5. Based on the score values, the proteins were cluster into related protein families (containing at least two sequences per family) using the TRIBE-MCL algorithm with default parameters [15].

Annotation and homology searches

Annotations into Gene Ontology (GO) categories [45] were obtained by searching the sequences of *L. bicolor* against the SwissProt database (Release 52.5) using the BlastP algorithm with an E-value threshold of 1e-10. GO annotations were inferred by retrieving information from the SwissProt entry corresponding to the best BlastP hit. All classified genes were mapped to their parents term using GO slim groupings [46]. Putative protein domains in the basidiomycete protein sequences were identified by searching the Pfam-LS database (version 22.0) [17] using an Hidden Markov Model (HMM) algorithm as implemented in the HMMPfam program (HMMER software, version 2.3.2, http://hmmer.janelia.org/), with an E-value threshold of 0.05. Searches for putative homologs to members within the large protein families of *L. bicolor* were performed by similarity search against the sequences of 33 fungal genomes (Broad Institute) using BlastP (E-value threshold 1e-5). Sequences of *L. bicolor* were also searched against the non-redundant (nr) protein sequence database at NCBI [47] (E-value threshold 1e-5).

Phylogenetic analysis

A phylogeny of the five analyzed basidiomycetes was constructed based on 18S rDNA sequences. Alignment was performed by ClustalW 1.81 [48] and MEGA 3.1 [49] was used for constructing a linearized Neighbor-Joining (NJ) tree with 1000 bootstrap replicates. *Aspergillus*

niger was used as an outgroup to root the tree and the estimated divergence between *A. niger* and *U. maydis* [14] was used as a reference point to estimate the divergence time between other lineages.

Evolutionary changes in the size of the protein families among the five basidiomycetes were analyzed using the CAFE tool [16]. It uses a stochastic birth and death process to model the evolution of gene family sizes over a phylogeny. Given a phylogeny and protein family sizes in extant species, the CAFE tool can infer the most likely gene family size at internal nodes and identify families that have accelerated rates of gain or losses, quantified by a *P*-value. The 18S rDNA tree and the sizes of 5383 TRIBE protein families, which included those present in at least two of the basidiomycete genomes, were used as inputs in the CAFE analysis. The rate of birth and death of proteins was estimated from the dataset to 0.0018 gains and losses per million years which is similar to previous reports in five yeast species (0.002) [50].

A phylogenetic tree of fungal protein kinases was constructed by extracting the protein kinase domain (Pfam PF00069) from 512 members of basidiomycete Family-2 (c.f. Table 3) using the Seqret software (EMBOSS). In addition, the PF00069 domain was extracted from 121 *Saccharomyces cerevisiae* protein kinases present in the KinBase database (http://kinase.com). All PF00069 domains (512+121) were aligned against a new profile HMM using the HMMAlign program (HMMER). The new profile HMM was constructed using the HMMbuild program (HMMER) from a full alignment of 25,281 protein kinase domains obtained from the Pfam database. The Neighbor-Joining (NJ) tree with 100 bootstrap replicates was constructed using Quicktree [51]. A tree viewer (ATV) [52] and the Hypertree program [20] were used to visualize and classify the protein kinase groups (Figure 6B). Clusters having a bootstrap support of 50% or higher in the Neighbor-Joining tree were classified into different groups and sub-groups. Using a similar procedure, a phylogenetic tree of ras GTPases was reconstructed by extracting the Pfam domain PF00071 from 160 proteins in Family-6 (Figure 7A, c.f. Table 3). These domains, together with 29 *S. cerevisiae* PF00071 domains obtained from Pfam database, were aligned against new profile HMM constructed from an alignment of 5239 Ras domains present in the Pfam database. The conserved G box motifs elements (G1,

G2, G3, G4 and G5) were identified in vast majority of the protein sequences in Family-6 [53].

Characterization of the basidiomycete protein kinase superfamily

A graph theory approach was adopted to characterize the protein families that belong to the protein kinase superfamily in the five basidiomycetes. In total, 1120 protein sequences were retrieved from 41 families containg sequences with significant homology to the Pfam domain PF00069 (E-value threshold <0.05). An all-against-all similarity search of the 1120 proteins using BlastP with an E-value threshold of 1e-5 was performed. The sequence similarities (E-values) were used to generate a weighted graph using the Biolayout graph algorithm [19]. The protein families identified by the TRIBE-MCL algorithm were mapped onto the graph and the topology of the protein families were inspected visually. Network properties, such as average connectivity, average node degree and maximum connectivity, of the protein kinase superfamily were calculated using the Biolayout program.

Gene expression analysis

The whole-genome *L. bicolor* microarray contains 20,226 gene model representatives and for each eight independent, non-identical, 60-mer probes. To remove probes that could hybridize to several gene models (i.e. cross-hybridize), the sequences of all probes were searched against the *L. bicolor* gene sequences using BlastN. Probes having more than 85% similarity to genes other than the genes used as a template for designing the probes were removed. After this filtering, 16,977 gene models remained in the dataset that could be identified with at least one 60-mer probe.

The array have been used for analyzing gene expression levels in three tissues (mycelium, mycorrhizal root tips and fruiting bodies) from 11 experiments (MS238N-I, MS238N-II, K, K2, E, MycD-I, MycD-II, MycPgh, MycPiv, FBE and FBL). Detailed descriptions on these microarray experiments can be obtained from Martin *et al*., (manuscript). Briefly, the free living mycelium was grown onto agar plates before harvesting the peripheral hyphal tips. Three weeks mycelium was isolated for MS238N-I and MS238N-II while two weeks mycelium without thiamine was isolated for K, K2 and E. The mycorrhizal root tips of *L. bicolor*/Douglas fir were extracted from nine

month ectomycorrhiza for MycD-I and MycD-II, *L. bicolor*/*Populus trichocarpa* were extracted from three month ectomycorrhiza for MycPgh, and *L. bicolor*/*Polpulus tremula* were extracted from one month ectomycorrhiza for MycPiv. Fruiting bodies of *L. bicolor* were collected from Douglas fir seedings grown in greenhouse for FBE and FBL.

Expression patterns for genes of the large *L. bicolor* protein families were subjected to principal component analysis (PCA) using the multivariate statistical analysis package MVSP [54]. PCA was performed on normalized $\log_2$ mean expression values for all genes within a protein family. The first principal components (PC1) were separated into three clusters - mycelia, mycorrhizal root tips and fruiting body tissues. Single factor ANOVA was performed on the three clusters to identify protein families showing significant (P <0.001) differential expression in the mycorrhizal root tips as compared to fruiting bodies and mycelia.

Acknowledgements

This study has been supported by grants from the Swedish Research Council.

References

1. Smith SE, Read DJ: *Mycorrhizal Symbiosis*. San Diego, CA: Academic Press; 1997.

2. Malloch DW, Pirozynski KA, Raven PH: **Ecological and evolutionary significance of mycorrhizal symbioses in vascular plants (A Review).** *Proceedings of the National Academy of Sciences of the United States of America* 1980, **77:**2113-2118.

3. Lepage BA, Currah RS, Stockey RA, Rothwell GW: **Fossil ectomycorrhizae from the middle Eocene.** *American Journal of Botany* 1997, **84:**410-412.

4. Hibbett DS, Gilbert LB, Donoghue MJ: **Evolutionary instability of ectomycorrhizal symbioses in basidiomycetes.** *Nature* 2000, **407:**506-508.

5. Wang XQ, Tank DC, Sang T: **Phylogeny and divergence times in Pinaceae: Evidence from three genomes.** *Molecular Biology and Evolution* 2000, **17:**773-781.

6. Moyersoen B: ***Pakaraimaea dipterocarpacea* is ectomycorrhizal, indicating an ancient Gondwanaland origin for the ectomycorrhizal habit in Dipterocarpaceae.** *New Phytologist* 2006, **172:**753-762.

7. Bruns TD, Szaro TM, Gardes M, Cullings KW, Pan JJ, Taylor DL *et al*: **A sequence database for the identification of ectomycorrhizal basidiomycetes by phylogenetic analysis.** *Molecular Ecology* 1998, **7:**257-272.

8. Ohno S: *Evolution by Gene Duplication*. Berlin: Springer Verlag; 1970.

9. Moran NA: **Symbiosis as an adaptive process and source of phenotypic complexity.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104 Suppl 1:**8627-8633.

10. Martinez D, Larrondo LF, Putnam N, Gelpke MD, Huang K, Chapman J *et al*: **Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78.** *Nature Biotechnology* 2004, **22:**695-700.

11. Kämper J, Kahmann R, Bolker M, Ma LJ, Brefort T, Saville BJ *et al*: **Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*.** *Nature* 2006, **444:**97-101.

12. Galagan JE, Henn MR, Ma LJ, Cuomo CA, Birren B: **Genomics of the fungal kingdom: Insights into eukaryotic biology.** *Genome Research* 2005, **15:**1620-1631.

13. Loftus BJ, Fung E, Roncaglia P, Rowley D, Amedeo P, Bruno D *et al*: **The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*.** *Science* 2005, **307:**1321-1324.

14. Berbee ML, Taylor JW: **Fungal molecular evolution: gene trees and geologic time.** In *The Mycota: A Comprehensive Treatise on Fungi as Experimental Systems for Basic and*

Applied Research. Edited by McLaughlin DJ, McLaughlin EG, Lemke PA. Springer; 2001:229-245.

15. Enright AJ, Van DS, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families.** *Nucleic Acids Research* 2002, **30:**1575-1584.

16. De Bie T., Cristianini N, Demuth JP, Hahn MW: **CAFE: a computational tool for the study of gene family evolution.** *Bioinformatics* 2006, **22:**1269-1271.

17. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T *et al*: **Pfam: clans, web tools and services.** *Nucleic Acids Research* 2006, **34:**D247-D251.

18. Manning G, Plowman GD, Hunter T, Sudarsanam S: **Evolution of protein kinase signaling from yeast to man.** *Trends in Biochemical Sciences* 2002, **27:**514-520.

19. Enright AJ, Ouzounis CA: **BioLayout - an automatic graph layout algorithm for similarity visualization.** *Bioinformatics* 2001, **17:**853-854.

20. Bingham J, Sudarsanam S: **Visualizing large hierarchical clusters in hyperbolic space.** *Bioinformatics* 2000, **16:**660-661.

21. Sonnhammer EL, Koonin EV: **Orthology, paralogy and proposed classification for paralog subtypes.** *Trends in Genetics* 2002, **18:**619-620.

22. Takai Y, Sasaki T, Matozaki T: **Small GTP-binding proteins.** *Physiological Reviews* 2001, **81:**153-208.

23. Wennerberg K, Rossman KL, Der CJ: **The Ras superfamily at a glance.** *Journal of Cell Science* 2005, **118:**843-846.

24. Lynch M: **The frailty of adaptive hypotheses for the origins of organismal complexity.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104 Suppl 1:**8597-8604.

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

26. Dale C, Moran NA: **Molecular interactions between bacterial symbionts and their hosts.** *Cell* 2006, **126:**453-465.

27. Kaneko T, Nakamura Y, Sato S, Minamisawa K, Uchiumi T, Sasamoto S *et al*: **Complete genomic sequence of nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum* USDA110.** *DNA Research* 2002, **9:**189-197.

28. Galibert F, Finan TM, Long SR, Puhler A, Abola P, Ampe F *et al*: **The composite genome of the legume symbiont *Sinorhizobium meliloti.*** *Science* 2001, **293:**668-672.

29. Cases I, de L, V, Ouzounis CA: **Transcription regulation and environmental adaptation in bacteria.** *Trends in Microbiology* 2003, **11:**248-253.

30. **The *Saccharomyces* Genome Database** [http://www.yeastgenome.org/]

31. Ridley AJ: **Rho family proteins: coordinating cell responses.** *Trends in Cell Biology* 2001, **11:**471-477.

32. Sundaram S, Kim SJ, Suzuki H, Mcquattie CJ, Hiremah ST, Podila GK: **Isolation and characterization of a symbiosis-regulated ras from the ectomycorrhizal fungus *Laccaria bicolor*.** *Molecular Plant Microbe Interaction* 2001, **14:**618-628.

33. Idnurm A, Howlett BJ: **Pathogenicity genes of phytopathogenic fungi.** *Molecular Plant Pathology* 2001, **2:**241-255.

34. Matheny PB, Curtis JM, Hofstetter V, Aime MC, Moncalvo JM, Ge ZW *et al*: **Major clades of Agaricales: a multilocus phylogenetic overview.** *Mycologia* 2006, **98:**982-995.

35. Taylor JW, Berbee ML: **Dating divergences in the Fungal Tree of Life: review and new analyses.** *Mycologia* 2006, **98:**838-849.

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

37. James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, Cox CJ *et al*: **Reconstructing the early evolution of Fungi using a six-gene phylogeny.** *Nature* 2006, **443:**818-822.

38. Le Quere A., Johansson T, Tunlid A: **Size and complexity of the nuclear genome of the ectomycorrhizal fungus *Paxillus involutus*** *Fungal Genetics and Biology* 2002, **36:**234-241.

39. **The Joint Genome Institute** [www.jgi.doe.gov]

40. **The Broad Institute** [www.broad.mit.edu/annotation/fgi}

41. **BlastClust** [ftp://ftp.ncbi.nih.gov/blast]

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

43. **The HIV Databases** [www.hiv.lanl.gov]

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

45. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM *et al*: **Gene Ontology: tool for the unification of biology.** *Nature Genetics* 2000, **25:**25-29.

46. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J *et al*: **The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology.** *Nucleic Acids Research* 2004, **32:**D262-D266.

47. **The Non Redundant Protein Database** [ftp://ftp.ncbi.nih.gov/blast/db]

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

49. Kumar S, Tamura K, Jakobsen IB, Nei M: **MEGA2: molecular evolutionary genetics analysis software.** *Bioinformatics* 2001, **17:**1244-1245.

50. Hahn MW, De BT, Stajich JE, Nguyen C, Cristianini N: **Estimating the tempo and mode of gene family evolution from comparative genomic data.** *Genome Research* 2005, **15:**1153-1160.

51. Howe K, Bateman A, Durbin R: **QuickTree: building huge Neighbour-Joining trees of protein sequences.** *Bioinformatics* 2002, **18:**1546-1547.

52. Zmasek CM, Eddy SR: **ATV: display and manipulation of annotated phylogenetic trees.** *Bioinformatics* 2001, **17:**383-384.

53. Bourne HR, Sanders DA, McCormick F: **The GTPase superfamily: conserved structure and molecular mechanism.** *Nature* 1991, **349:**117-127.

54. Kovach WL. *MVSP: A Multivariate Statistical Package for Windows.* Petraeth, UK: Kovach Computing Services; 1998.

Supplementary material

Table S1. Protein families in *Laccaria bicolor* that showed contractions or extinctions[a]

Protein family	Number of proteins[b]					Pfam accession[c]	Pfam description
	Lac	*Copr*	*Phan*	*Cryp*	*Ust*		
Contractions in *Lac*							
0	52	337	280	50	2	PF00665, PF00078, PF00385	Integrase core domain, Reverse transcriptase, chromo domain
4	39	78	85	1	7	PF00067	Cytochrome P450
8	5	75	65	12	0	PF07727, PF00665	Reverse transcriptase Integrase core domain
26	6	30	16	12	42	PF00665, PF07727	Integrase core domain, Reverse transcriptase
28	9	77	6	4	3	PF00023, PF05729	Ankyrin repeat, NACHT domain
29	17	37	34	2	9	PF05199, PF00732, PF01266	GMC oxidoreductase, GMC oxidoreductase, FAD dependent oxidoreductase
31	13	53	31	0	0		
46	3	67	2	0	0		
92	1	37	2	3	1		
169	1	12	10	1	2	PF00107[d]	Zinc-binding dehydrogenase
273	1	9	8	0	0	PF00840, PF00734[d]	Glycosyl hydrolase family 7, Fungal cellulose binding domain
Extinctions in *Lac*							
24	0	88	26	0	0	PF00069[d]	Protein kinase domain
240	0	16	4	0	0	PF00734	Fungal cellulose binding domain

[a]Listed are the 13 protein families showing significant (P <0.001) contractions and extinctions in the *L. bicolor* lineage (branch 4 in Figure 1A).

[b]Number of proteins in respective protein family - *Laccaria bicolor (Lac)*, *Coprinopsis cinerea (Copr)*, *Phanerochaete chrysosporium (Phan)*, *Cryptococcus neoformans (Cryp)* and *Ustilago maydis (Ust)*.

[c]The putative protein domains search was performed for *C. cinerea* proteins against the Pfam-LS database [Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T *et al*: *Nucleic Acids Research* 2006, **34:**D247-D251.] with E-value threshold of 0.05. Domains present in more than 25% of the members in the protein family are shown.

[d]Putative protein domains search (procedure similar to above) was performed against *P. chrysosporium*.

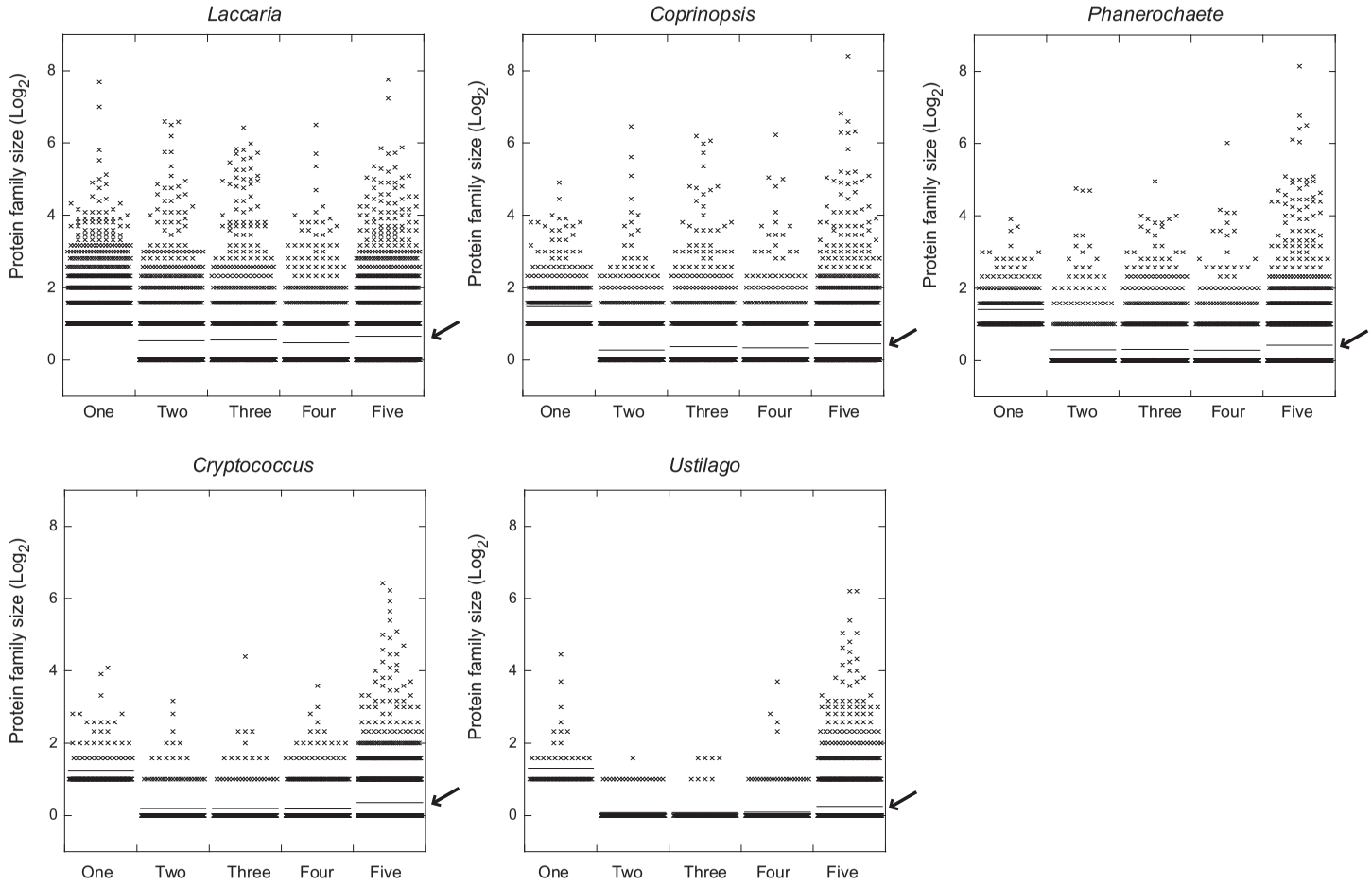


Figure S1. Expansion of protein family sizes in the genomes of *Laccaria bicolor*, *Coprinopsis cinerea*, *Phanerochaete chrysosporium*, *Cryptococcus neoformans* and *Ustilago maydis*. Shown on y-axis are the number of proteins (transformed $\log_2$ values) in protein families that are present in one genome (i.e. lineage-specific), or shared between any two, three, four or all five of the analyzed genomes (c.f. Figure 3). Zero on the y-axis indicates single or two proteins within a protein family. The horizontal thin line (indicated by arrow) in each category represents average protein family size.

Molecular Ecology (2006) **15**, 535–550

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.

III

doi: 10.1111/j.1365-294X.2005.02796.x

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

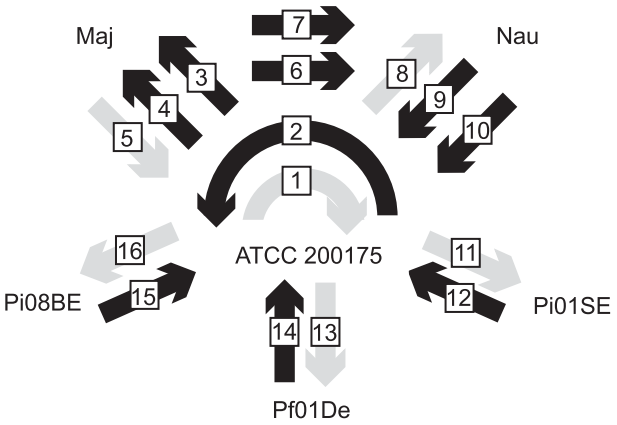


Fig. 1 Experimental design for the CGH analysis of five strains of *Paxillus involutus* (ATCC, Pi01SE, Pi08BE, Maj and Nau) and one strain of *Paxillus filamentosus* (Pf01De). Each arrow indicates paired dual-label microarray hybridizations of reference (ATCC) and test DNA, except for hybridizations 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

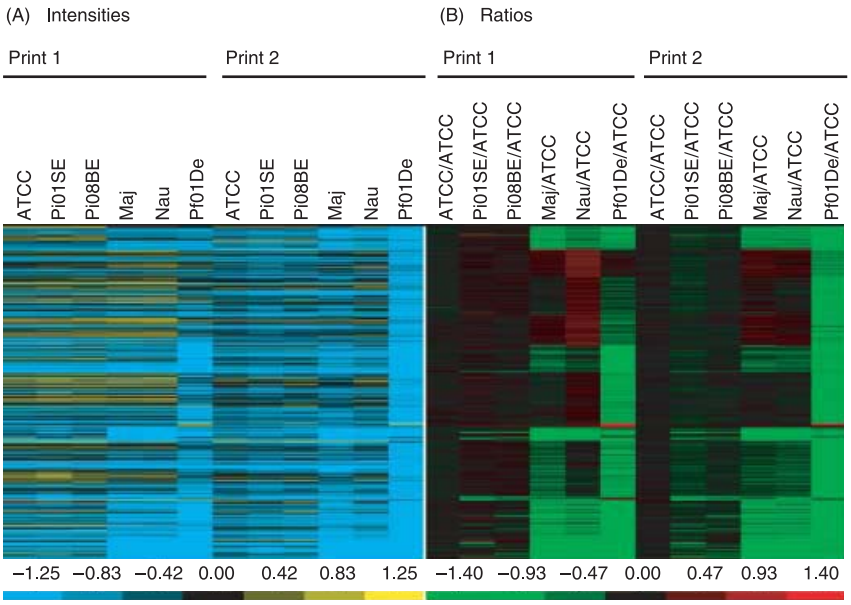


Fig. 2 CGH analyses of five strains of *Paxillus involutus* (ATCC, Pi01SE, Pi08BE, Maj and Nau) and one strain of *Paxillus filamentosus* (Pf01De). The array was printed using PCR-amplified cDNA derived from a collection of EST clones obtained from the ATCC strain (reference) and contained 1076 unique 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).

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

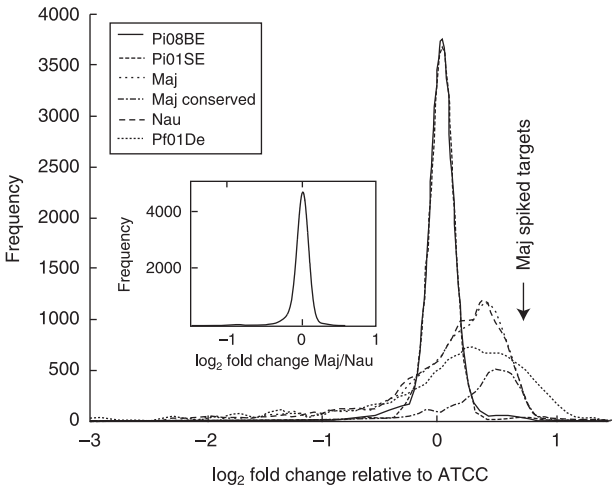


Fig. 3 Distribution of $\log_2$ fold changes in hybridization signals for various *Paxillus* strains relative to the reference strain ATCC. 'Maj conserved' show the distribution of fold changes for a conserved subset of genes in the strain Maj. Those were identified by a TBLASTX search including all 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

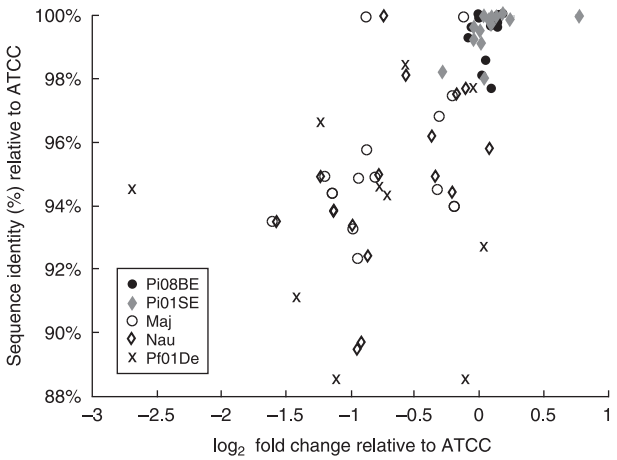


Fig. 4 Relationship between sequence identity and fold changes in hybridization signals for 17 loci in various *Paxillus* strains relative to the ATCC reference strain. The sequence identity in exon regions for each gene (cf. Table 2) is plotted against its corresponding $\log_2$ fold change in hybridization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

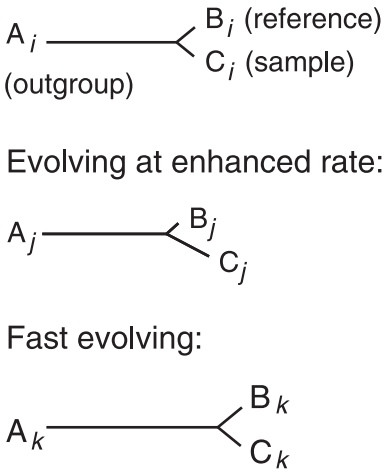


Fig. 5 The EPLP screen for identifying locally divergent genes. (A) represents an organism with a fully sequenced genome located outside the clade of strains analysed by CGH. (B) here is the reference strain, which is also the strain used for the construction of the microarray. (C) is a closely related 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.

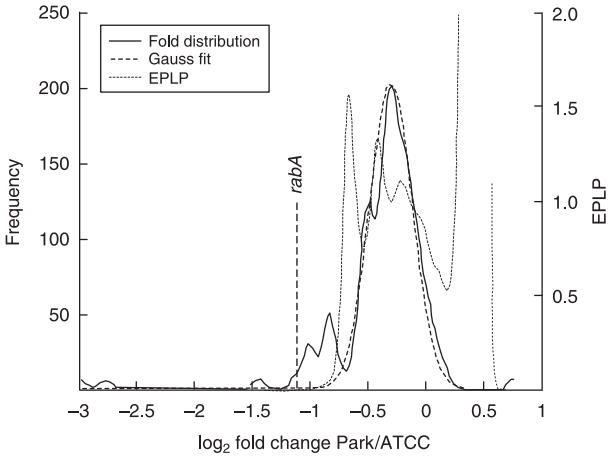


Fig. 6 Application of the EPLP algorithm. The method is illustrated here for the *rabA* gene (cf. Table 2), which has a TBLASTX score of 105 against *Coprinus cinereus*. The solid line shows the distribution of $\log_2$ fold change in hybridization signals between Park and ATCC for 102 genes with a similar 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

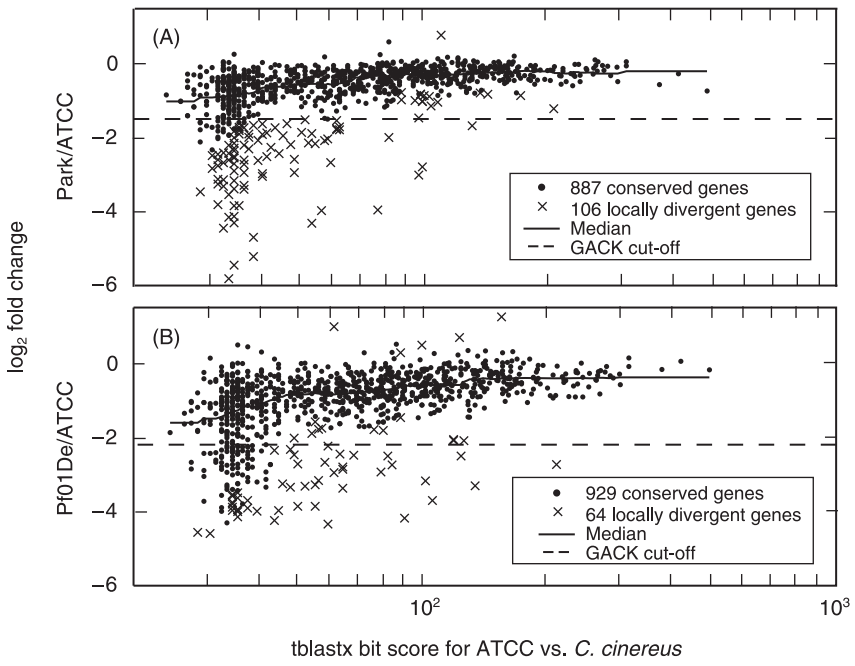


Fig. 7 Locally divergent genes identified in *Paxillus involutus* using the EPLP procedure. The scatter plots show the $\log_2$ fold changes in hybridization signals between (A) the reference strain ATCC and Park, and (B) the ATCC and the *Paxillus filamentosus* strain Pf01De, vs. the sequence similarity (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 underrepresentation 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

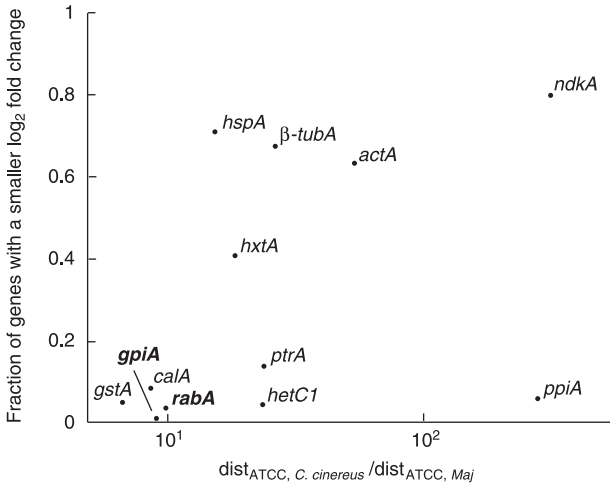


Fig. 8 Validation of the EPLP procedure. Plotted is the relationship between an EPLP estimate of divergence vs. a sequence-based measures of accelerated evolution. Both methods analyse how fast genes are evolving between Maj (from the Park clade) and ATCC (Forest clade) strains of *Paxillus involutus* 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.

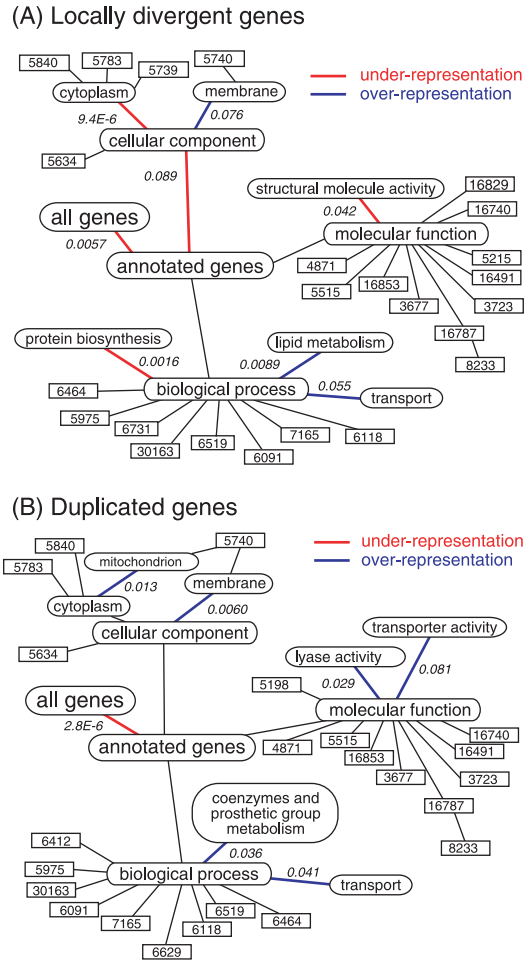


Fig. 9 Functional annotations of (A) locally divergent and (B) duplicated genes. Based on sequence similarities, the *Paxillus* genes were annotated into GO categories organized as molecular function, biological process and cellular component (Ashburner *et al*. 2000). The relationships between GO categories 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

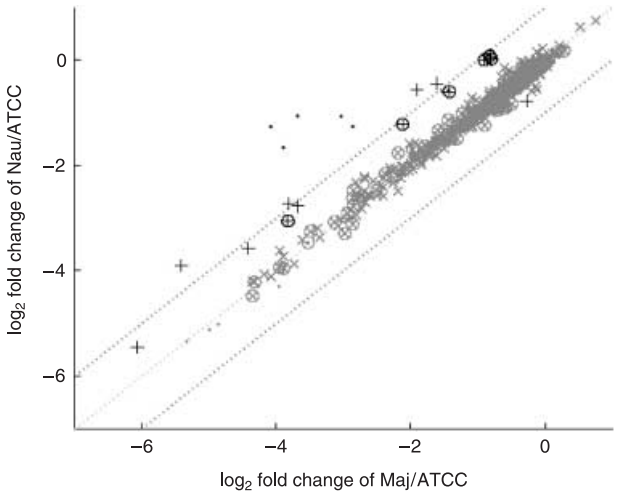


Fig. 10 Identification of genes being duplicated in the compatible strain Maj and the incompatible strain Nau (i.e. not infecting birch or poplar). The scatter plot shows the $\log_2$ fold change in hybridizations signals of Maj relative to ATCC (reference strain) vs. the $\log_2$ fold change of Nau 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.

New Phytologist

Research

Evolution of nucleotide sequences and expression patterns of hydrophobin genes in the ectomycorrhizal fungus *Paxillus involutus*

Balaji Rajashekar*, Peter Samson*, Tomas Johansson and Anders Tunlid

Department of Microbial Ecology, Lund University, Ecology Building, SE–223 62, Lund, Sweden

Author for correspondence:
Anders Tunlid
Tel: +46 46 2223757
Fax: +46 46 2224158
Email: Anders.Tunlid@mbioekol.lu.se

Received: *23 October 2006*
Accepted: *22 December 2006*

Summary

• Hydrophobins are small, secreted proteins that play important roles in the development of pathogenic and symbiotic fungi. Evolutionary mechanisms generating sequence and expression divergence among members in hydrophobin gene families are largely unknown.

• Seven hydrophobin (*hyd*) genes and one *hyd* pseudogene were isolated from strains of the ectomycorrhizal fungus *Paxillus involutus*. Sequences were analysed using phylogenetic methods. Expression profiles were inferred from microarray experiments.

• The *hyd* genes included both young (recently diverged) and old duplicates. Some young *hyd* genes exhibited an initial phase of enhanced sequence evolution owing to relaxed or positive selection. There was no significant association between sequence divergence and variation in expression levels. However, three *hyd* genes displayed a shift in the expression levels or an altered tissue specificity following duplication.

• The *Paxillus hyd* genes evolve according to the so-called birth-and-death model in which some duplicates are maintained for a long time, whereas others are inactivated through mutations. The role of subfunctionalization and/or neofunctionalization for preserving the *hyd* duplicates in the genome is discussed.

Key words: ectomycorrhiza, expression patterns, gene duplication, hydrophobin, neofunctionalization, *Paxillus involutus*, subfunctionalization.

New Phytologist (2007) **174**: 399–411

doi: 10.1111/j.1469-8137.2007.02022.x

Introduction

Duplications of genes or larger chromosome regions in combination with mutations that cause functional divergence of the duplicates is considered to be the most important mechanisms generating evolutionary novelties, including new gene functions and expression patterns (Ohno, 1970). Such duplications have most likely played a substantial role both in the rapid change in organismal complexity apparent in deep evolutionary split, and in the adaptation and diversification of more closely related strains or species (Prince & Pickett, 2002; Long *et al.*, 2003). Analyses of genome sequences suggest that gene duplications arise by very high rates. Some of these new genes are preserved, but a majority are silenced within a few millions of years (Lynch & Conery, 2000). Accordingly, there is a relatively narrow time window for evolutionary explorations until gene inactivation becomes the most likely outcome. The fate of the recent duplicated genes and the evolutionary forces that drive their fixation and divergence are, however, not yet clear (Long *et al.*, 2003).

Recently, we have used DNA microarrays to screen for duplicated and rapidly evolving genes that could be associated with symbiotic adaptations in the ectomycorrhizal (ECM) fungus *Paxillus involutus* (Basidiomycetes; Boletales). Strains of *P. involutus* with various abilities to form ECM were analysed by comparative genomic hybridizations using a cDNA

*These authors have contributed equally to this work.

IV

Table 1 Fungal strains used in this study

Species/strain	Abbreviation	Site and mycorrhizal host	Origin	References
Paxillus involutus				
ATCC 200175	AT	Isolated close to birch trees. Forms ECM with birch, pine, spruce and poplar (in the laboratory)	Scotland	Chalot *et al*. (1996)
Pi01Se	Se	Isolated from a pine forest	Sweden	S. Erland (unpublished)
Pi08Be	Be	Forms ECM with pine, spruce and poplar (laboratory)	Belgium	Blaudez *et al*. (1998)
Maj	Mj	Isolated close to poplar trees. Forms ECM with poplar and birch (laboratory)	France	Gafur *et al*. (2004)
Nau	Nu	Isolated close to oak trees. Does not form ECM with poplar or birch but with oak (laboratory)	France	Gafur *et al*. (2004)
Paxillus filamentosus				
Pf01De	Pf	Isolated close to alder trees	Germany	Jarosch & Bresinsky (1999)

ECM, ectomycorrhiza.

microarray containing 1076 putative unique genes. Approximately 17% of the gene representatives available on the array were detected as rapidly and presumably nonneutrally evolving within *Paxillus* (Le Quéré *et al.*, 2006). Among them were several genes encoding hydrophobins. Hydrophobins are small, secreted proteins that can self-assemble and form visible aggregations of protein rodlets on fungal surfaces (Kershaw & Talbot, 1998; Wösten, 2001). Hydrophobins are known to play a role in a range of different processes related to growth and development in fungi. For example, hydrophobins are involved in the formation of aerial structures such as spores and fruiting bodies, they can mediate adhesion of pathogenic fungi to plant host surfaces and they can function as toxins (Kershaw & Talbot, 1998; Wösten, 2001). Hydrophobins have also been shown to be developmentally regulated by ECM-forming fungi during the infection of the host plant (Tagu *et al.*, 1996; Mankel *et al.*, 2002; Duplessis *et al.*, 2005; Le Quéré *et al.*, 2005). Hydrophobin genes are commonly found in multigene families. The members in these families display a large divergence in nucleotide sequences and expression patterns (Wessels *et al.*, 1991; Segers *et al.*, 1999; Duplessis *et al.*, 2001).

In this study we have examined the evolutionary mechanisms that could be responsible for generating sequence and expression divergence among members of the hydrophobin gene family in *P. involutus* ATCC 200175. Seven hydrophobin (*hyd*) genes were characterized. Orthologs were isolated from several *Paxillus* strains and one closely related species (Le Quéré *et al.*, 2004) and compared with available sequences from other fungi using phylogenetic analysis and tests for selection. Gene expression patterns were inferred using data from several cDNA microarray experiments.

Materials and Methods

Fungal cultures and DNA extractions

Five strains of *P. involutus* (Batsch: Fr.) and one closely related species *Paxillus filamentosus* (Table 1) were grown on cellophane-covered agar plates and liquid medium, and DNA was prepared as previously described (Le Quéré *et al.*, 2002). The obtained DNA extracts were finally treated with RNase A (Promega, SDS Biosciences, Falkenberg, Sweden).

Isolation of *hydA* to *hydG*

Putative *hyd* genes were identified in a *P. involutus* ATCC 200175 expressed sequence tag (EST) database. The database contains 19 188 ESTs (assembled into a uniset of 3700 fungal gene representatives) originating from 11 different cDNA libraries and have partly been published (Johansson *et al.*, 2004). The uniset can be estimated to correspond to approx. 50% of the total number of genes in *P. involutus*, assuming a gene content of 7700 (Le Quéré *et al.*, 2002). The cDNA libraries represent material harvested from different tissues and under different growth conditions including saprophytically growing mycelium and ECM root tips associated with birch (*Betula pendula*; Johansson *et al.*, 2004), mycelium, cords and ECM root tips growing in soil (D. Wright *et al.* unpublished) and mycelium growing in liquid cultures with different nutrient composition (M. Caillau *et al.* unpublished). Homology searches of the EST sequences were performed against public databases using the BLASTX program (Altschul *et al.*, 1990). Within this database seven contigs displaying a hydrophobin signature motif (Kershaw & Talbot, 1998) were identified.

For the complete sequencing of cDNAs corresponding to these hydrophobin genes (designated *hydA* to *hydG* (Table 2), for gene nomenclature refer to the Supplementary Material, Table S3), plasmid clones were retrieved from the collection of EST clones (Johansson *et al.*, 2004). Plasmids were prepared and the inserts were amplified and sequenced using vector-specific or template-specific primers (see the Supplementary Material, Table S1) as previously described (Wright *et al.*, 2005). Overlapping sequences were aligned and trimmed using the SEQUENCHER software (v. 3.1.1b4) (Gene Codes Corp., Ann Arbor, MI, USA).

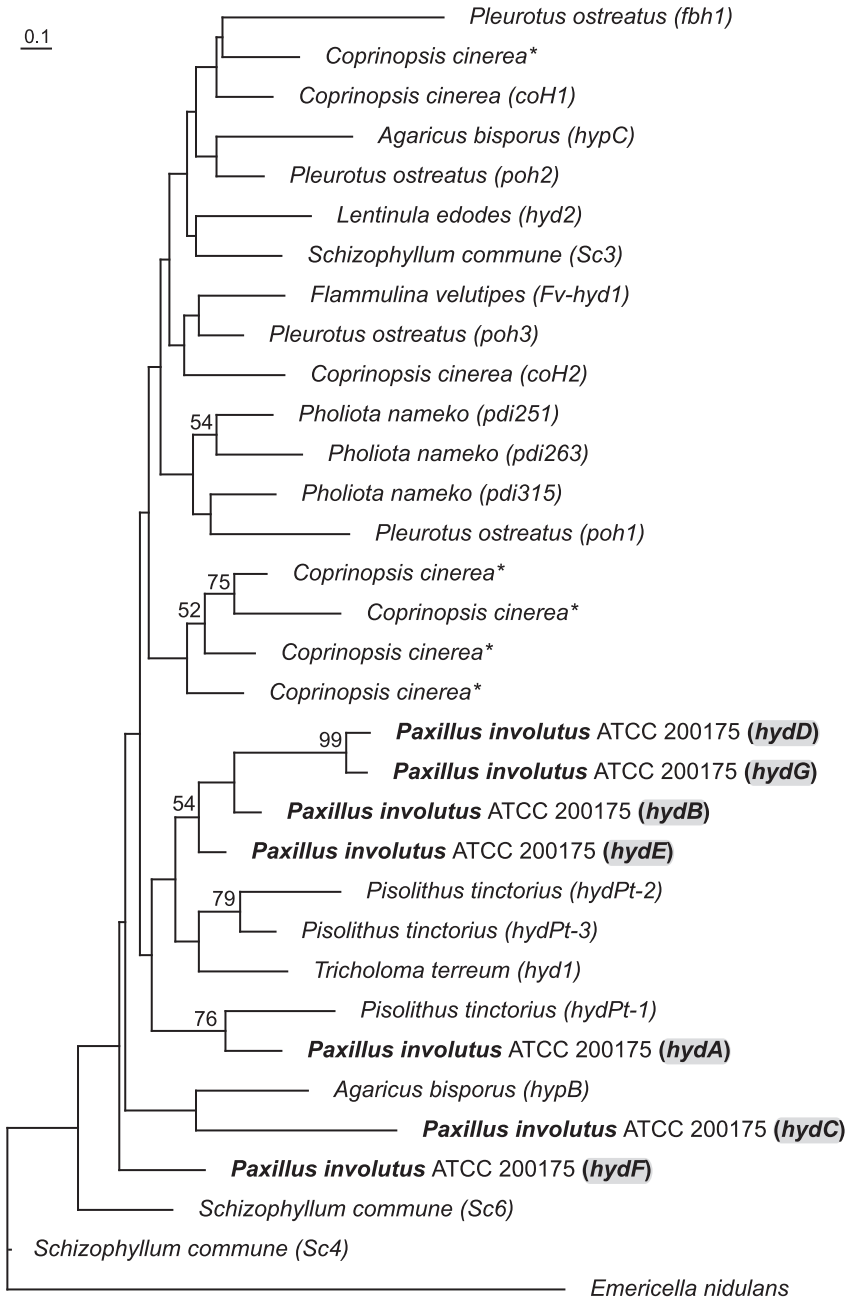


Fig. 1 Phylogenetic relationships between hydrophobins from *Paxillus involutus* ATCC 200175 and those from other basidiomycetes. The maximum likelihood tree shows values at the nodes indicating the bootstrap support in percent for 100 replicates (only values > 50 are shown). Accession numbers for the nucleotide sequences are from top to bottom: AJ319663, Scaffold 6* (292824-293084, 293144-293185), Y10627, X90818, AJ225061, AF217808, M32329, AB126686, Y16881, Y10628, AB079128, AB079129, AB079130, AJ225060, Scaffold 25* (103205-103417, 103474-103515), Scaffold 24* (134686-104898, 104988-135029), Scaffold 18* (76134-75859, 75794-75750), Scaffold 7* (95562-95287, 95225-95184), U29606, AF097516, AY048578, U29605, Y15940, AJ007504, M32330 and M61113 (*Emericella nidulans*) as an outgroup for rooting the tree. Asterisks (*) indicate sequences obtained from Broad Institute (http://www.broad.mit.edu/annotation/). The shaded boxes indicate the seven hydrophobin genes (*hydA* to *hydG*) from *P. involutus* ATCC 200175 and their accession numbers are given in Table 2 and supplementary material Table S3.

Phylogenetic analyses

Putative homologues for *hydA* to *hydG* were identified by querying the GenBank database (www.ncbi.nlm.nih.gov) and genome sequences of *Coprinopsis cinerea* (*Coprinus cinereus*, Sequencing Project, Broad Institute of MIT and Harvard, www.broad.mit.edu) using BLASTN search algorithm. Nucleotide sequences (in total 82) with an *E*-value ≤ 10e-5 were translated into polypeptide sequences which were then aligned using the MUSCLE (version 3.6) (Edgar, 2004) and BIOEDIT version 7.0.4.1 (www.mbio.ncsu.edu/BioEdit/bioedit.html) softwares. Sequences that were shorter than expected, or without identified start and stop codons, or not containing codons for all the eight cysteine residues according to the hydrophobin signature motif (Kershaw & Talbot, 1998), or sequences distantly related to *hydA* to *hydG* were excluded from further analysis. A final set of 26 protein sequences remained after filtering. These sequences together with *hydA* to *hydG* and a hydrophobin from *Emericella nidulans* (outgroup) (cf. Figure 1) were realigned using MUSCLE. The resulting protein alignment was used as a template to align the corresponding nucleotide sequences. In

the nucleotide alignment, codons upstream to the first cysteine codon (i.e. upstream to the hydrophobin core motif; Wösten, 2001), along with gaps and a few ambiguous sites were removed before the phylogenetic analysis. The final alignment containing 201 bp were used to construct a phylogenetic tree by the Maximum Likelihood method in PAUP* software (version 4.0b8) (Swofford, 1998). The software MODELTEST (Posada & Crandall, 1998) was used to evaluate the appropriate models and parameters.

The evolutionary relationships between the seven *P. involutus* ATCC 200175 *hyd* genes were analysed further using the SplitsTree method (Huson, 1998; see the Supplementary Material, Fig. S2).

Analysis of hydrophobin gene fragments in *Paxillus* strains

Based on the hydrophobin cDNA sequences obtained from the *P. involutus* ATCC 200175 (*hydA* to *hydG*), new primers were designed (see the Supplementary Material, Table S1) to amplify the corresponding genomic regions of *hydA* to *hydG* in various strains of *P. involutus* and *P. filamentosus*. The amplicons were ligated into the pGEM-T Easy Vector System I (Promega) and transformed into *Escherichia coli* DH5α. For each amplicon/primer pair, several clones were randomly picked and in total 98 genomic fragments were sequenced as already described. The sequences of the putative *hydA* – *hydG* genes were aligned using MUSCLE and BIOEDIT software. The identity between the aligned sequences were estimated using MATGAT software (v 2.02) (Campanella *et al.*, 2003). Sequences showing pairwise identities of > 99% to existing sequences and four truncated sequences identified as pseudogenes were removed. The phylogenetic relationships of the remaining 48 gene fragments were analysed by constructing a neighbour-joining (NJ) tree (Galtier *et al.*, 1996). The phylogeny was constructed using a 50%-majority rule consensus of 1000 neighbour-joining bootstrap replicates, adjusted with the Kimura-2 model. Accession numbers of the amplified gene fragments are given in the Supplementary Material (Table S3).

The rates of nonsynonymous (d_n) and synonymous (d_s) nucleotide substitutions per site for each of the *hydA* to *hydG* orthologous groups were estimated separately using the CRANN software (Creevey & McInerney, 2003). The hydrophobin gene (*Sc4*) of *Schizophyllum commune* was used as outgroup.

Analysis of expression data

The expression patterns for six (*hydA* to *hydF*) of the seven *P. involutus* ATCC 200175 *hyd* genes (*hydG* was excluded because no reporter was available on the array) were examined by collecting processed relative expression levels from dual-label cDNA-microarray experiments (each including dye swaps and at least three biological replicates) from a number of studies (cf. Table 3) which included 25 different treatments covering a wide range of developmental and physiological conditions (Table 3 and Fig. 5). The microarray design, the sample preparation, the hybridization procedures and the statistical analyses of the array data have previously been described in detail by Johansson *et al.* (2004) and Le Quéré *et al.* (2005), and involved normalization and statistical analysis (mixed-model analysis of variance (ANOVA)) as described by Wolfinger *et al.* (2001). From those analyses we retrieved the estimated relative expression level ($\log_2$), for each *hyd* gene and in each treatment, for the construction of transcriptional profiles (Fig. 5).

The expression divergences for all *hyd* gene pairs were estimated by calculating the Euclidean distances of the $\log_2$-transformed expression levels (residuals) using data from 13 of the treatments listed in Table 3 (designated A to H, APO, ASU, BSA, CHI and GLN). These distances were related to the corresponding pair-wise nucleotide sequence divergence measured either as d_s or as amino acid (aa) sequence distances (d); d_s was calculated as described earlier. *hydC* was excluded from the analysis because of significant higher substitution rates at the second and third codon positions in comparisons to the other *P. involutus* ATCC 200175 *hyd* genes. The PROTDIST program from the PHYLIP-3.64 package was used to estimate d from the alignment using the Dayhoff PAM matrix along with default parameters.

The estimated $\log_2$ relative expression levels of *hydA* to *hydF* were subjected to principal component analysis (PCA) using the Multivariate Statistical Package MVSP (Kovach, 1998).

Results

Characterization of *hydA* to *hydG*

Seven putative hydrophobin genes designated *hydA* to *hydG*, were identified within a collection of EST clones originating from the *P. involutus* ATCC 200175. The corresponding cDNA sequence information were translated into polypeptide sequences and showed predicted sizes ranging between 107 and 141 aa residues (Table 2). All polypeptides were predicted to contain an *N*-terminal signal peptide as well as eight conserved cysteine residues, which are characteristic features of fungal hydrophobins (see the Supplementary Material, Fig. S1).

Three distinct gene structures were observed in the *hyd* genes. The *hydA* gene contained three introns, *hydB* to *hydF* genes contained two introns whereas the *hydG* gene consisted of a single uninterrupted exon (Table 2). The nucleotide sequences of *hydA* to *hydG* varied extensively. The *hydD* and *hydG* (88%) showed the highest pairwise identity followed by the pair *hydB* and *hydE* (76%). The nucleotide sequence of *hydC* appeared most divergent and displayed low identities (53–56%) against all the other *P. involutus* ATCC 200175

Table 2 Characterization of hydrophobin genes and their gene products in *Paxillus involutus* ATCC 200175

Gene[a]	Length (bp)	Exons[b] (bp)	Introns (bp)	Number of introns	GC[c] (%)	Protein (aa)	ESTs (library)[d]	Mean expression level (log_2)[e]	Strains[f]
hydA	929	426	169	3	57	141	23 (1)	0.78	6 (AT,Be,Se,Mj,Nu,Pf)
hydB	637	336	113	2	63	111	114 (9)	0.72	4 (AT,Be,Se, –,Nu, –)
hydC	779	324	116	2	53	107	21 (5)	0.54	6 (AT,Be,Se,Mj,Nu,Pf)
hydD	607	327	112	2	61	108	5 (1)	–1.15	5 (AT,Be,Se,Mj,Nu, –)
hydE	614	324	107	2	59	107	16 (3)	2.58	3 (AT, –, –,Mj,Nu, –)
hydF	602	351	107	2	60	116	1 (1)	–0.59	6 (AT,Be,Se,Mj,Nu,Pf)
hydG	596	327	0	0	58	108	2 (2)	NA	1 (AT, –, –, –, –, –)

EST, expressed sequences tag.
[a]Accession numbers are given in the Supplementary Material, Table S3.
[b]The boundaries of exons and introns were identified by comparing genomic and cDNA sequences.
[c]The GC content in exons (GC) were calculated using the DAMBE-4.13 program (Xia & Xie, 2001).
[d]Total number of ESTs clustered for each gene. The numbers in the brackets refer to the distribution of ESTs among 11 cDNA libraries.
[e]Means of relative expression levels (log_2-transformed) calculated from 25 microarray experiments (Table 3). NA, No data available because the array did not contain reporters for *hydG*.
[f]Total number and names of *Paxillus* strains (cf. Table 1) in which orthologs of *hydA* to *hydG* were amplified (cf. Figure 2). Accession numbers for all sequences are given in the Supplementary Material, Table S3.

hyd genes (see the Supplementary Material, Table S2). In addition, *hydC* had a considerable lower GC content in its exon (53%) compared with that of other *P. involutus* ATCC 200175 *hyd* genes (57–63%) (Table 2).

Phylogeny of *hydA* to *hydG*

A phylogentic analysis showed that the *hydA* to *hydG* genes from *P. involutus* ATCC 200175 were found among genes of the Class I group of hydrophobins found in basidiomycetes (Wessels, 1997) (Fig. 1). The *hydD*, *hydG*, *hydB* and *hydE* clustered into one clade (bootstrap support value 54) with *hydD* and *hydG* (bootstrap support 99) as the most recently duplicated genes. The *hydA*, *hydC* and *hydF* genes were dispersed among other basidiomycete sequences. The *hydA* was found in a clade (support value 76) containing the *hydPt-1* gene from *Pisolithus tinctorius* (Tagu *et al.*, 1996). Two other hydrophobin genes (*hydPt-2* and *hydPt-3*) have been isolated from *P. tinctorius* (Tagu *et al.*, 1996) which were found outside this clade. The orthologs *hydA* and *hydPt-1* encode for considerably longer polypeptides (141 and 140 aa) than those encoded by *hydB* to *hydG* and *hydPt-2* and *hydPt-3* (in the range 107–117 aa) (Table 2) (Tagu *et al.*, 1996).

Owing to the rather poor resolution in the maximum likelihood tree, the evolutionary relationships between *hydA* to *hydG* of the *P. involutus* ATCC 200175 were also analysed using the split decomposition method, as implemented in the SplitsTree (Huson, 1998). The SplitsTree graph supported the finding that *hydD* and *hydG* are closely related. In addition, the analysis also indicates that *hydE* is more closely related to *hydB* than to *hydD* and *hydG* (see the Supplementary Material Fig. S2).

Evolution of hydrophobin genes within the *Paxillus* clade

Attempts were made to amplify, by polymerase chain reaction (PCR), genomic fragments corresponding to each of the seven *hyd* genes in five strains of *P. involutus* and in one closely related species *P. filamentosus* (Table 1). Sequence analysis showed that *hydA*, *hydC* and *hydF* orthologs were successfully amplified from all the six *Paxillus* strains/species (Fig. 2) and were found in well-resolved clades supported with high bootstrap values (82, 100 and 91, respectively). By contrast, *hydB*, *hydD*, *hydE* and *hydG* orthologs were less universal and were amplified in one to five strains/species. The relationships of the clades of *hydB* and *hydE*, and *hydG* and *hydD*, respectively, supports the finding (cf. above) that these genes represent two recently duplicated gene pairs. In addition, with the primers used, hydrophobin gene fragments were amplified that formed two additional clades, designated *hydX* and *hydY*. The *hydX* clade was closely related to the *hydD* and *hydG* clades, and the *hydY* clade was closely related to the clade of *hydF*.

Analysis of hydrophobin pseudogenes

Among the gene fragments amplified with primers designed against *hydE*, there were four sequences that appeared to be truncated, containing only part of the conserved hydrophobin signature motif. Two sequences (Mj-*hydP1* and Mj-*hydP2*) were amplified from the strain Maj and another two (Nu-*hydP1* and Nu-*hydP2*) from the closely related strain Nau (Le Quéré *et al.*, 2004). These four sequences were 100% identical and the sequence of Mj-*hydP1* was analysed in detail. From comparisons with the other *hyd* genes it was revealed that the above sequences represent a pseudogene of *hydE*

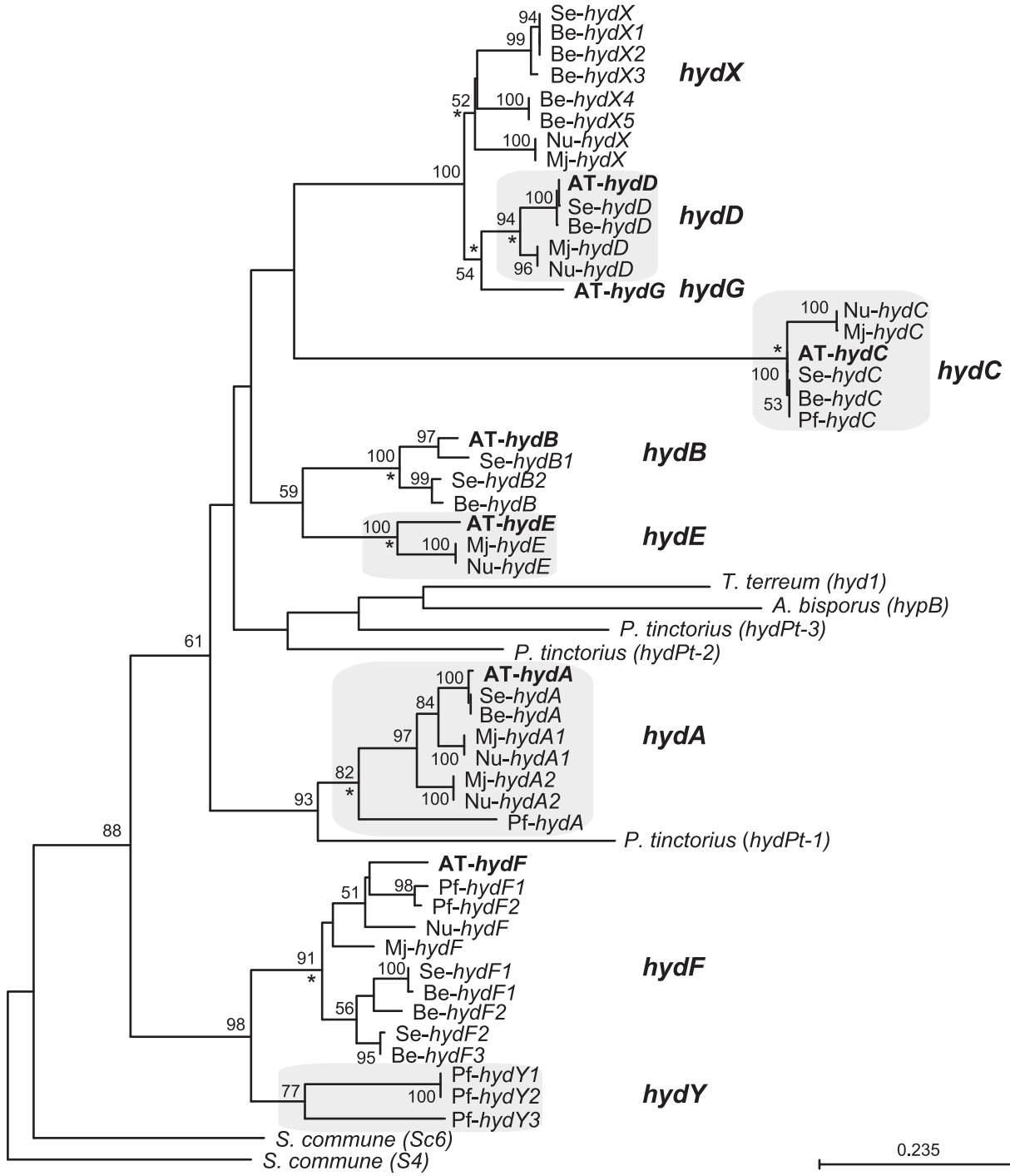


Fig. 2 Phylogenetic tree showing relationships of hydrophobin genomic fragments that were polymerase chain reaction (PCR)-amplified from different strains of *Paxillus*. The gene fragments from *Paxillus* strains were amplified using gene-specific primers based on the *hydA* to *hydG* genes (see the Supplementary Material, Table S1). The first alphabets in the gene name represent the name of strain (e.g. AT) or species (e.g. Pf), separated by '-', followed by gene names abbreviated as three lower-case italics and an upper-case italic letter (e.g. *hydA*). The italic number after the locus name refers to a clone number. Accession numbers of all the genes are given in Table S3. The tree was constructed using the neighbour joining algorithm on 48 hydrophobin sequences from the *Paxillus* strains along with seven hydrophobin genes from other closely related basidiomycetes *Tricholoma terreum*, *Agaricus bisporus*, *Pisolithus tinctorius* and *Schizophyllum commune* (cf. Fig. 1). The tree was rooted using a sequence from *S. commune* (*S4*). Values at nodes (only > 50 are displayed) represent the bootstrap support value in per cent of 1000 replicates. An asterisk (*) at the node indicates well-resolved clades that contain genes considered to have evolved directly from the same ancestral locus (i.e. they represent well-defined orthologs). *hydX* and *hydY* represent two clades of sequences with not yet identified homologs in the *P. involutus* ATCC 200175. Alternate *hyd* gene groups have been shaded.

Fig. 3 Reconstruction of events leading to the formation of a hydrophobin pseudogene in *Paxillus involutus*. The pseudogene (Mj-*hydP1*) is compared against *hyd* genes AT-*hydE* (*P. involutus* ATCC 200175 strain), Mj-*hydE* and Mj-*hydX* (Maj strain). The pseudogene (Mj-*hydP1*) was amplified from the strain Maj using primers designed for *hydE*. (a) Organization of the fully sequenced AT-*hydE* gene The gene has three exons (shaded boxes) and two introns (open boxes). The translational start codon (M), the codons for the eight cysteine residues(C) of the hydrophobin signature motif and the stop codon (*) are indicated. In the pseudogene (Mj-*hydP1*), the region corresponding to Exon 2 and Exon 3 is well conserved whereas a number of degenerative mutations have occurred in the region corresponding to Exon 1 and Intron 1 (underlined). The nucleotide (nt) sequence identity between the region (alignment had 216 sites including gaps and the stop codon) containing Exons 2 and 3 of the pseudogene and AT-*hydE* is 60%, and between the pseudogene and the Mj-*hydE* ortholog Maj is 71%. The Exons 2 and 3 of the pseudogene also display high sequence identity (62%) to another amplified hydrophobin gene fragment of Maj (Mj-*hydX*) that is found among a uncharacterized cluster of hydrophobin genes named *hydX* (cf. Fig. 3). (b) Enlargement of the region corresponding to Exon 1 and Intron 1. Enclosed in a box are the translated sequences (amino acid (aa) in upper case letters) corresponding to the region of Exon 1 in AT-*hydE*. Region corresponding to the primer binding site are excluded (+). Note the presence of a stop codon (*) in the pseudogene. Furthermore, the pseudogene lacks part of the nucleotides encoding Exon 1 and Intron 1 (nt in lower case letters). (c) Enlargement of the boxed region displayed in (b). The nt alignment of Exon 1 has been divided into four segments (1–4). Segment 1 is the primer binding site (+). Segment 2 shows a region where two nt have been inserted (position 4 and 19) in the pseudogene compared with the Mj-*hydE*, Mj-*hydX* and AT-*hydE* genes. Segment 3 of the pseudogene contains a variable region (enclosed in box) where insertions and deletions of nt are observed. Segment 4 shows the part of Exon 1 that has been deleted in the pseudogene. (d) Translation of the nt sequences in Exon 1 after removing the inserted nt (at positions 4 and 19 indicated in panel (c)) in the pseudogene. In the segment 2 of the pseudogene, seven of the first 10 aa residues are identical. The remaining nt sequences of Exon 1 that are absent in the pseudogene correspond to a region containing three of the eight cysteine residues of the hydrophobin motif.

(Fig. 3; Supplementary Material, Table S3). The Exons 2 and 3 which contain five out of eight conserved cysteine residues in the hydrophobin motif were intact in the pseudogene and this region displayed high sequence identity to *hydE* from the Maj strain. The Exon 1 and Intron 1 of this pseudogene displayed several degenerative mutations leading to the formation of a nonfunctional gene: one mutation introducing a stop codon, insertions of two single nucleotides in the first exon and a deletion of a larger fragment encompassing part of the first exon and the entire first intron.

Rate of nucleotide substitutions

The rates and patterns of nucleotide substitutions within the *Paxillus hyd* gene family were analysed by calculating the d_n and d_s values for the orthologs of *hydA–hydF,* and *hydX.* In total 143 pair-wise comparisons were performed of which 14 showed a brief period of d_n/d_s ratio > 1, indicating relaxed or positive selection (Fig. 4). Four of these comparisons included *hydB*, six *hydD* and the remaining four comparisons involved genes of the *hydX* group.

Divergence in expression profiles

The expression levels of the *P. involutus* ATCC 200175 *hyd* genes varied extensively depending on the growth conditions and the tissues being analysed (Fig. 5). Overall, the pattern of expression profiles for *hydA*, *hydB*, *hydC* and *hydE* were similar, whereas *hydD* and *hydF* were expressed at lower levels and showed different patterns of regulation.

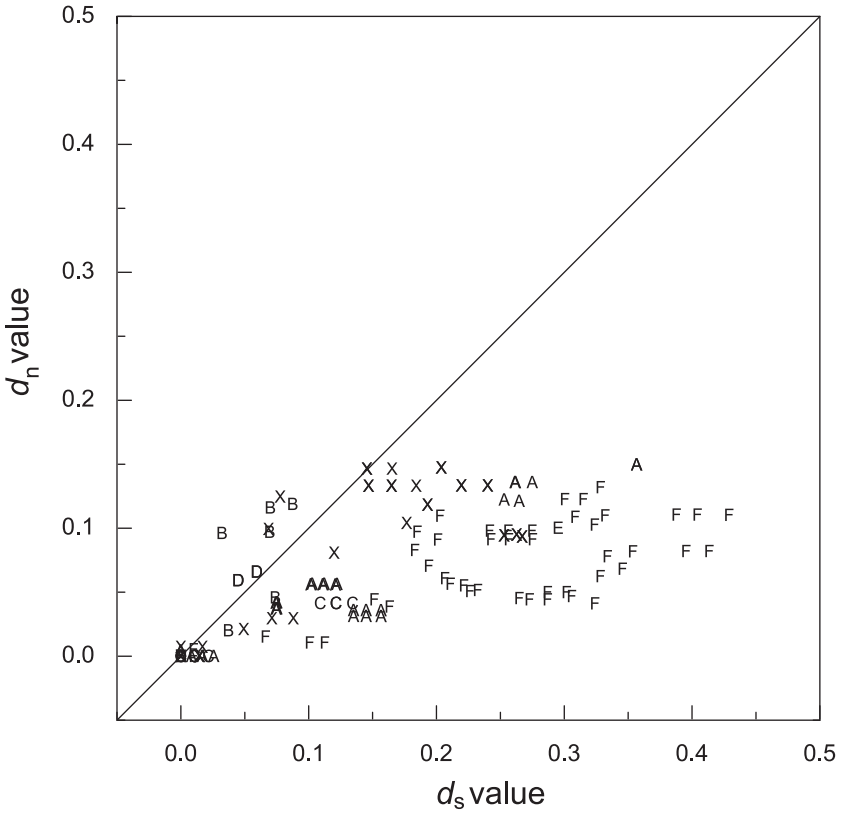


Fig. 4 Comparison of rates of nonsynonymous (d_n) and synonymous (d_s) nucleotide substitutions per coding site in the hydrophobin genes. Each alphabetic character represents a pairwise comparison made between the true orthologous hydrophobin genes belonging to *hydA* to *hydX* from *Paxillus involutus* (strains ATCC 200175, Pi08Be, Pi01Se, Maj and Nau) and *P. filamentosus* (Pf01De) (Fig. 2). The letter A corresponds to pairwise comparisons of hydrophobin genes within the true orthologous group indicated by an asterisks (*) in the *hydA* node in Fig. 2. The remaining letters, B, C, D, E, F and X, correspond to gene comparison of orthologous group *hydB*, *hydC*, *hydD*, *hydE*, *hydF* and *hydX*, respectively (Fig. 2). The diagonal line shows the neutral expectation where d_n is equal to d_s, if $d_n/d_s > 1$ then the pairwise comparisons occur above the diagonal line.

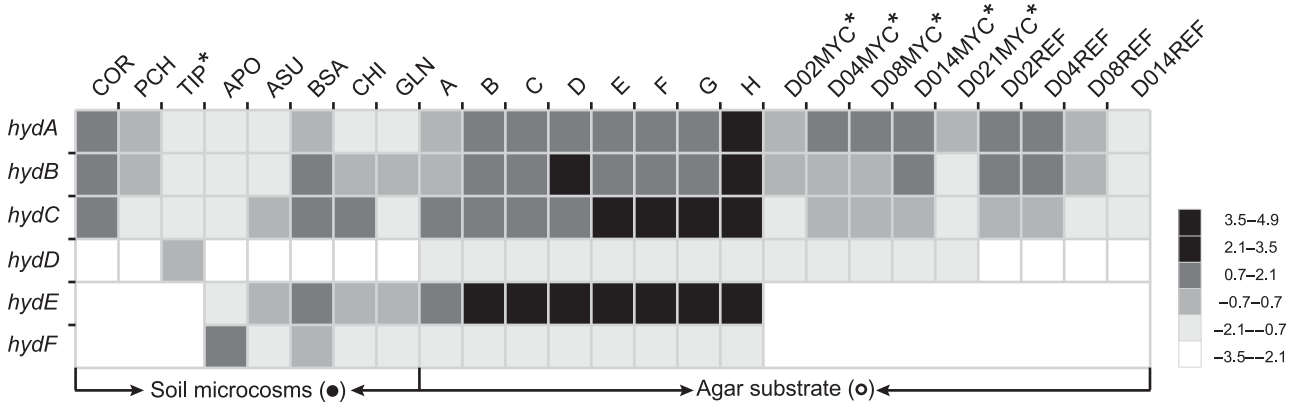


Fig. 5 Transcriptional profiles for six *hyd* genes (*hydA* to *hydF*) from *Paxillus involutus* ATCC 200175 based on relative expression levels retrieved from a number of different dual-label cDNA microarray studies as listed in Table 3. Columns represent 25 different treatments (Table 3) and are covering a wide range of developmental and physiological conditions. Treatments indicated by an asterisk (*) are ectomycorrhizal (ECM) tissues from associations with birch (*Betula pendula*). The mycelia grown in soil microcosms (●) and on agar substrates (○) are indicated with respective symbols. The scale shows relative expression levels on a $\log_2$ scale (c.f. Materials and Methods).

To examine whether the divergence in expression levels were related to the divergence in sequence between gene duplicates, the distances in expression level were related to d_s. This measure can be used as a proxy for the divergence time between gene duplicates (Gu *et al.*, 2002). Although the expression divergence increased with sequence divergence, the association with d_s was not statistically significant ($r = 0.29$, $P = 0.45$, $n = 9$). In addition, we tested the correlation between expression divergence and protein sequence divergence (d) of the hydrophobin gene pairs (Wagner, 2000). This association was weak and not statistically significant ($r = -0.33$, $P = 0.22$, $n = 15$).

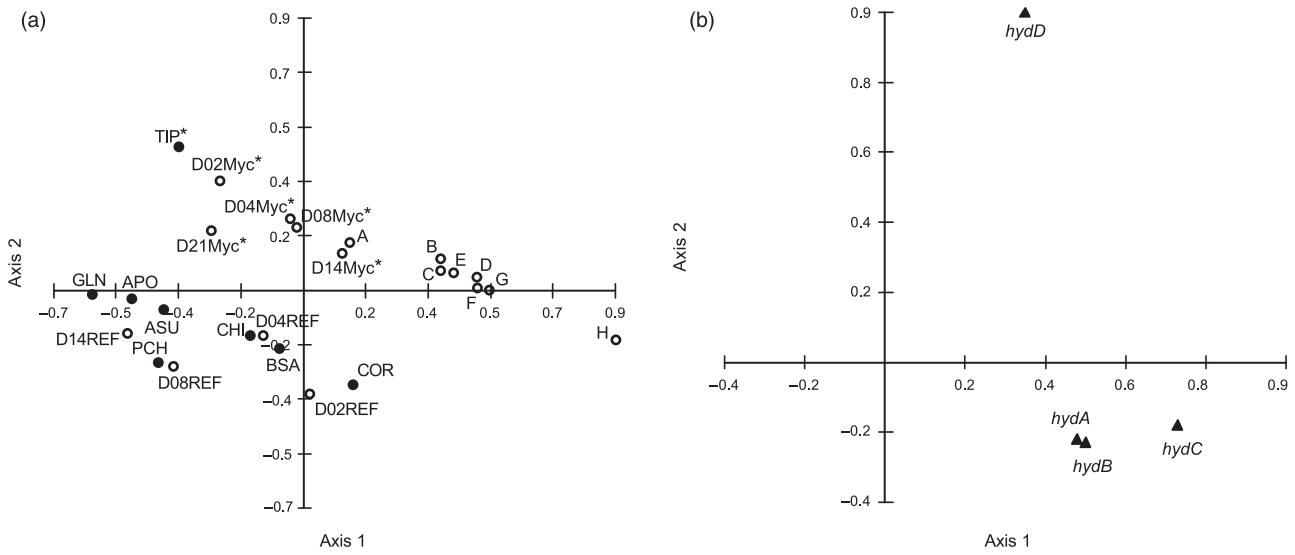


Fig. 6 Principal component analysis (PCA) of the patterns of expression levels of the hydrophobins *hydA* to *hydD* from 25 different microarray experiments (Table 3, Fig. 5). (a) Sample plots of 25 different microarray experiments. The PCA axis 1 and axis 2 explained 68% and 18%, respectively, of the total variation in the data. Treatments indicated by an asterisk (*) are ectomycorrhizal (ECM) tissues from associations with birch (*Betula pendula*). The mycelia grown in soil microcosms (●) and on agar substrates (○) are indicated with respective symbols. (b) Loading values of *hydA* to *hydD*. PCA axis 1 explains 68% and axis 2 18% of the variation.

To characterize the patterns of variation in expression levels more clearly, the data was examined by principal component analysis (PCA). We performed PCA on four *hyd* genes (*hydA* to *hydD*) for which expression data was available for a total of 25 microarray experiments. The first two principal components (PC1 and PC2) accounted together for 86% of the variation (Fig. 6a). The PC1 separated samples growing in soils from those growing on defined agar medium. The PC2 separated the samples of ECM root tips from those of mycelia and cords. A projection of the hydrophobins on the sample plane showed that the expression levels for *hydA*, *hydB* and *hydC* were positively correlated and explained mainly the variation along the PC1 (Fig. 6b). Conversely, *hydD* was more closely projected to the ECM samples along the PC2. Thus *hydD* appear to be specifically regulated in ECM root tips.

Discussion

Seven *hyd* genes were identified in a collection of EST clones from *P. involutus* ATCC 200175. Owing to the fact that the complete genome sequence of *P. involutus* is not available, the isolated genes might not include all the members of the *hyd* gene family in this fungus. The phylogenetic analyses indicated that four of the *hyd* genes characterized – *hydB*, *hydD*, *hydE* and *hydG* – have diverged rather recently, presumably after the separation of the *Paxillus* clade (suborder Paxillinieae) within the Boletales (Binder & Bresinsky, 2002) (Figs 1 and 2 and the Supplementary Material, Fig. S2). Furthermore, *hydD*/*hydG* and *hydB*/*hydE* represent two recently duplicated gene pairs. These apparent young duplicates were not found in all the *Paxillus* strains analysed. By contrast, orthologs for three other hydrophobins, *hydA*, *hydC* and *hydF*, were identified from all the *Paxillus* strains examined (Fig. 2). However, the evolutionary history of these *hyd* genes differed. *hydA* was found in a clade containing *hydPt-1* from *P. tinctorius*. The two genes translate into proteins sharing a unique primary structure that is different from those encoded by other hydrophobin genes in *P. involutus* and *P. tinctorius*. Considering the fact that *P. involutus* and *P. tinctorius* belong to two evolutionary distant lineages (suborders) within the Boletales, Paxillineae and Sclerodermatineae, respectively (Binder & Bresinsky, 2002), we conclude *hydA* is an ancient copy, which has been maintained for a long time in the genome of *Paxillus*. The *hydC* is found in a branch containing no other *hyd* genes or orthologs from other species (Fig. 2). The *hydC* gene showed the lowest GC content in the exons and displayed the lowest pair-wise nucleotide identity towards the other *P. involutus* *hyd* genes. By contrast, *hydF* is found in a clade with another closely related but not yet characterized *hyd* gene (*hydY*) (Fig. 2).

In addition to *hydA* to *hydG*, we identified a *hyd* pseudogene (*hydP*) in the Maj and Nau strains. Phylogenetic analyses of internal transcribed spacer (ITS) sequences have shown that these strains are closely related and comprise a well-resolved clade within *P. involutus* (Le Quéré et al., 2004). Comparison of the *hyd* pseudogenes and their cognate ORFs in *P. involutus* showed that the pseudogene had a truncated hydrophobin motif retaining five out of the eight conserved cysteine residues (Fig. 3). Studies of the crystal structure of

Table 3 Microarray experiments of the *Paxillus involutus* ATCC 200175

Growth conditions and tissues		Array[a]	References
Treatment[b]	Description		
P. involutus in association with birch (*Betula pendula*) in soil microcosms (●):		Print 85	Wright *et al*. (2005)
COR	Cords (rhizomorphs)		
PCH	Extramatrical mycelium growing in nutrient patch		
TIP*	ECM root tips		
Mycelium colonizing nutrient patches in soil microcosms. The following nutrients were added to the patches (●):		Print 154	Wright *et al*. (unpublished)
APO	Ammonium phosphate		
ASU	Ammonium sulphate		
BSA	Bovine serum albumin		
CHI	Chitin		
GLN	Glutamine		
Mycelium grown on agar with the following nutrient amendments (○):		Print 154	Caillau *et al*. (unpublished)
A	Ammonia		
B	Ammonia + patch of phosphate		
C	Ammonia + patch of phosphate + birch seedlings		
D	Complete medium		
E	Complete medium + birch seedlings		
F	Modified complete medium + birch seedlings		
G	Phosphate + patch of ammonia		
H	Phosphate + patch of ammonia + birch seedling		
P. involutus in association with birch (Betula pendula) grown on agar substrate (○):		Print 85	Le Quéré et al. (2005)
D02MYC*	ECM root tips developed after 2, 4, 8, 14 and 21 d, respectively,		
D04MYC*			
D08MYC*			
D14MYC*			
D21MYC*			
D02REF	Mycelium grown axenically for 2, 4, 8, 14 and 21 d, respectively,		
D04REF			
D08REF			
D14REF			

ECM, ectomycorrhiza.
[a]Two different batches of cDNA microarrays (Print 85 and 154) were used in the experiments. Both arrays were printed with reporters obtained from a nonredundant set of expressed sequences tag (EST) clones originating from the *P. involutus* ATCC 200175. Each reporter was replicated in at least quadruplicates on the array. Print 85 contained reporters for *hydA* to *hydD* and Print 154 contains reporters for *hydA* to *hydF*. A full description of the Print 85 array design is available from the EMBL-EBI ArrayExpress database (http://www.ebi.ac.uk/arrayexpress/; Accession number A-MEXP-92) whereas the Print 154 array design is in preparation for submission.
[b]Treatments indicated by an asterisk (*) are ECM tissues from associations with birch (*Betula pendula*).

hydrophobins and analysis of mutants have shown that all eight cysteine residues are needed for producing a functional hydrophobin (Hakanpaa *et al*., 2004; Kershaw *et al*., 2005). There were several partly degraded pseudogenes recognized within gene families of *Saccharomyces cerevisiae* (Lafontaine *et al*., 2004). However, such pseudogenes are not expected according to the classical model of gene degradation, which assumes that random mutations are accumulated in the absence of a selection pressure. This raises the question whether pseudogenes can remain in the genome because they are involved in homology-dependent gene silencing mechanisms (Cogoni, 2001; Lafontaine *et al*., 2004).

The above data suggests that the *Paxillus hyd* gene family evolves according to the birth-and-death model (Nei & Rooney, 2005). This model predicts that new genes are created by gene duplication, and whereas some duplicates remain functional in the genome for long period of time, others are inactivated or deleted. Because of the lack of genomic data, the mechanisms by which the *Paxillus hyd* genes become duplicated are largely unknown. Nevertheless, the lack of introns and the phylogenetic position of the *hydG* gene suggest that this gene has arisen by retrotransposition of *hydD* transcript (Long *et al*., 2003). This is probably a recent event since *hydG* was only identified in the *P. involutus* ATCC 200175 strain. There is also a possibility that *hydD* and *hydG* correspond to alleles of same gene. Since the 5′- and 3′-untranslated regions of *hydG* are much longer than *hydD* (see the Supplementary Material, Table S3), we still favour the

supposition that the *hydG* has arisen by retrotransposition of *hydD*.

The classical model for the origin of functional novelties following gene duplication postulates that gene duplication creates a redundant locus that is free to accumulate otherwise deleterious mutations as long as the original copy maintains the ancestral function (Ohno, 1970). The most likely outcome of this period of relaxed selection is that the redundant gene degenerates to become a pseudogene (nonfunctionalization). A less frequent outcome is that the redundant copy evolves a new function by a process known as neofunctionalization. In addition to neofunctionalization, it has also been proposed that preservation of gene duplicates can be brought by the process of subfunctionalization. In this model, the original function of the single-copy ancestral gene is partitioned between the two daughters (Prince & Pickett, 2002). The patterns of selection predicted by these models have been revealed by analysing gene duplicates in eukaryotes with sequenced genomes (Lynch & Conery, 2000). Similar to the results of these studies, we observed a period of relaxed selection in several of the young *hyd* duplicates (Fig. 4). Assuming that silent substitutions are not subjected to selection and that their number increases linearly with time, *hydB* and *hydD*, and a not yet fully characterized *hyd* gene (*hydX*) have experienced a phase of accelerated evolution, as confirmed from d_n/d_s ratios ≥ 1, which is an indicative of relaxed or positive selection (Nei & Kumar, 2000). A decline in the d_n/d_s ratio reflects a gradual increase in the magnitude of selective constraints (purifying selection). Notably, *hydE*, which is the closest duplicate to *hydB*, had a d_n/d_s ratio < 1. This suggests that the two recently duplicated *hydB* and *hydE* genes were diverging at different rates and under different selection pressures, as predicted by the neofunctionalization model.

Changes in gene expression are thought to be a major reason for the functional divergence and retention of duplicated genes (Ohno, 1970; Prince & Pickett, 2002). We asked whether the divergence in expression levels of the *P. involutus hyd* genes have increased with gene sequence divergence, that is, evolutionary time. This question has been examined for members in gene families in several eukaryotic model organisms using data from microarray experiments. These studies have provided various and in some cases contradictory results (Wagner, 2000; Makova & Li, 2003; Blanc & Wolfe, 2004; Haberer *et al.*, 2004). We found that the expression divergence of the *P. involutus hyd* paralogs increased with time, but their association with protein sequence divergence (d) or d_s or d_n, was weak and not statistically significant. Thus, the evolution of coding regions and mRNA expression patterns appear to be uncoupled among the *P. involutus hyd* genes. A reason for not observing such correlations can be that the number of gene pairs compared was low and that the data set contains both old and young duplicates. Studies of gene duplicates in yeast, human and *Arabidopsis* suggest that evolution in expression patterns are rapid and correlate with sequence divergence only during a brief period of time after duplication (Gu *et al.*, 2002; Makova & Li, 2003; Blanc & Wolfe, 2004).

Gene duplicates that are evolving according to the neofunctionalization and subfunctionalization models can be expected to accumulate mutations in their promoter sequences that can lead to shifts in their expression levels and tissue specificity (Prince & Pickett, 2002; Duarte *et al.*, 2006). Notably, such shifts were detected in the expression of several *P. involutus hyd* genes. These shifts involved increased (*hydE*) and reduced levels (*hydD* and *hydF*) of expression, as well as increased tissue specificity (*hydD*) (Figs 5 and 6; Table 2). According to the so-called DDC (duplication–degeneration–complementation) model of subfunctionalization, degenerative promotor mutations can alter the level of expression of daughter genes to the point where both copies are needed to supply enough protein products (Force *et al.*, 1999). Although, the DDC model can explain the shift in expression levels of the *P. involutus hyd* genes, the model cannot explain the altered tissue specificity of *hydD*. The subfunctionalization model predicts that the expression pattern of an ancestral gene is partitioned between its daughters. However, the expression pattern of the presumably ancestral *hydA* gene was not divided between *hydD* and other *P. involutus hyd* genes. Moreover, the expression pattern of *hydD* has not evolved according to the neofunctionalization model. Thus, the *hydD* expression pattern is not entirely novel, as other *hyd* paralogs were also expressed in the ECM root tissue. The above data suggests that the expression patterns of the *P. involutus hyd* gene family have evolved according to more complex combinations of the neofunctionalization and subfunctionalization models. Mixtures of these models have recently been noted in studies of duplicates in yeast, human, mouse and *Arabidopsis* (He & Zhang, 2005; Huminiecki & Wolfe, 2004; Duarte *et al.*, 2006).

One concern in microarray experiments is that cross-hybridization between closely related paralogs and reporters may affect the result. We have previously shown, using similar array hybridization conditions referred to in this investigation, that the signals decrease rapidly when the sequence similarity drops below 90–95% (Le Quéré *et al.*, 2006). Considering that the sequence identity between the *hyd* paralogs analysed was below 88% (see the Supplementary Material, Table S2), we do not feel that the expression data of these genes have been distorted by cross-hybridization.

Acknowledgements

This study was supported by grants from the Swedish Research Council. B.R. and P.S. were supported by grants from the Research School in Genomics and Bioinformatics. 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.

References

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

Binder M, Bresinsky A. 2002. Derivation of a polymorphic lineage of Gasteromycetes from boletoid ancestors. *Mycologia* **94**: 85–98.

Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* **16**: 1679–1691.

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.

Campanella JJ, Bitincka L, Smalley J. 2003. MatGAT: an application that generates similarity/identity matrices using protein or DNA sequences. *BMC Bioinformatics* **4**: 29.

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.

Cogoni C. 2001. Homology-dependent gene silencing mechanisms in fungi. *Annual Review of Microbiology* **55**: 381–406.

Creevey CJ, McInerney JO. 2003. CRANN: detecting adaptive evolution in protein-coding DNA sequences. *Bioinformatics* **19**: 1726.

Duarte JM, Cui L, Wall PK, Zhang Q, Zhang X, Leebens-Mack J, Ma H, Altman N, dePamphilis CW. 2006. Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of *Arabidopsis*. *Molecular Biology and Evolution* **23**: 469–478.

Duplessis S, Courty PE, Tagu D, Martin F. 2005. Transcript patterns associated with ectomycorrhiza development in *Eucalyptus globulus* and *Pisolithus microcarpus*. *New Phytologist* **165**: 599–611.

Duplessis S, Sorin C, Voiblet C, Palin B, Martin F, Tagu D. 2001. Cloning and expression analysis of a new hydrophobin cDNA from the ectomycorrhizal basidiomycete Pisolithus. *Current Genetics* **39**: 335–339.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**: 1792–1797.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**: 1531–1545.

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

Galtier N, Gouy M, Gautier C. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Computer Applications in the Biosciences* **12**: 543–548.

Gu Z, Nicolae D, Lu HH, Li WH. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends in Genetics* **18**: 609–613.

Haberer G, Hindemitt T, Meyers BC, Mayer KF. 2004. Transcriptional similarities, dissimilarities, and conservation of *cis*-elements in duplicated genes of *Arabidopsis*. *Plant Physiology* **136**: 3009–3022.

Hakanpaa J, Paananen A, Askolin S, Nakari-Setala T, Parkkinen T, Penttila M, Linder MB, Rouvinen J. 2004. Atomic resolution structure of the HFBII hydrophobin, a self-assembling amphiphile. *Journal of Biological Chemistry* **279**: 534–539.

He X, Zhang J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**: 1157–1164.

Huminiecki L, Wolfe KH. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Research* **14**: 1870–1879.

Huson DH. 1998. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**: 68–73.

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

Johansson T, Le Quéré A, Ahrén D, Söderström B, Erlandsson R, Lundeberg J, Uhlén M, Tunlid A. 2004. Transcriptional responses of *Paxillus involutus* and *Betula pendula* during formation of ectomycorrhizal root tissue. *Molecular Plant–Microbe Interactions* **17**: 202–215.

Kershaw MJ, Talbot NJ. 1998. Hydrophobins and repellents: Proteins with fundamental roles in fungal morphogenesis. *Fungal Genetics and Biology* **23**: 18–33.

Kershaw MJ, Thornton CR, Wakley GE, Talbot NJ. 2005. Four conserved intramolecular disulphide linkages are required for secretion and cell wall localization of a hydrophobin during fungal morphogenesis. *Molecular Microbiology* **56**: 117–125.

Kovach WL. 1998. *MVSP: a multivariate statistical package for Windows*, version 3.0. Petraeth, UK: Kovach Computing Services.

Lafontaine I, Fischer G, Talla E, Dujon B. 2004. Gene relics in the genome of the yeast *Saccharomyces cerevisiae*. *Gene* **335**: 1–17.

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, Canbäck B, Hedh J, Erland S, Johansson T, Tunlid A. 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 Interactions* **18**: 659–673.

Le Quéré A, Astrup EK, Rajashekar B, Schutzendubel A, Canbäck B, Johansson T, Tunlid A. 2006. Screening for rapidly evolving genes in the ectomycorrhizal fungus *Paxillus involutus* using cDNA microarrays. *Molecular Ecology* **15**: 535–550.

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

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

Makova KD, Li WH. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Research* **13**: 1638–1645.

Mankel A, Krause K, Kothe E. 2002. Identification of a hydrophobin gene that is developmentally regulated in the ectomycorrhizal fungus *Tricholoma terreum*. *Applied and Environmental Microbiology* **68**: 1408–1413.

Nei M, Kumar S. 2000. *Molecular evolution and phylogenetics*. New York, NY, USA: Oxford University Press.

Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. *Annual Review of Genetics* **39**: 121–152.

Ohno S. 1970. *Evolution by gene duplication*. Berlin, Germany: Springer Verlag.

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

Prince VE, Pickett FB. 2002. Splitting pairs: the diverging fates of duplicated genes. *Nature Review Genetics* **3**: 827–837.

Segers GC, Hamada W, Oliver RP, Spanu PD. 1999. Isolation and characterization of five different hydrophobin-encoding cDNAs from the fungal tomato pathogen *Cladosporium fulvum*. *Molecular General Genetics* **261**: 644–652.

Swofford DL. 1998. *PAUP: phylogenetic analysis using parsimony (and other methods)*, version 4. Sunderland, MA, USA: Sinauer Associates.

Tagu D, Nasse B, Martin F. 1996. Cloning and characterization of hydrophobins-encoding cDNAs from the ectomycorrhizal basdiomycete *Pisolithus tinctorius*. *Gene* **168**: 93–97.

Wagner A. 2000. Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist–selectionist debate. *Proceedings of the National Academy of Sciences, USA* **97**: 6579–6584.

Wessels JGH. 1997. Hydrophobins: proteins that change the nature of the fungal surface. *Advances in Microbial Physiology* **38**: 1–45.

Wessels JGH, De Vries OMH, Asgeirsdottir SA, Schuren FHJ. 1991. Hydrophobin genes Involved in formation of aerial hyphae and fruit bodies in *Schizophyllum*. *Plant Cell* **3**: 793–799.

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS. 2001. Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* **8**: 625–637.

Wösten HA. 2001. Hydrophobins: multipurpose proteins. *Annual Review of Microbiology* **55**: 625–646.

Wright DP, Johansson T, Le Quéré A, Söderström B, Tunlid A. 2005. Spatial patterns of gene expression in the extramatrical mycelium and mycorrhizal root tips formed by the ectomycorrhizal fungus *Paxillus involutus* in association with birch (*Betula pendula* Roth.) seedlings in soil microcosms. *New Phytologist* **167**: 579–596.

Xia X, Xie Z. 2001. DAMBE: software package for data analysis in molecular biology and evolution. *Journal of Heredity* **92**: 371–373.

Supplementary Material

The following supplementary material is available for the article online:

Fig. S1 Comparison of the primary structure of the hydrophobins hydA to hydG from the *Paxillus involutus* ATCC 200175.

Fig. S2 Phylogenetic relationship between *hydA* to *hydG* from the *Paxillus involutus* ATCC 200175 as revealed by the SplitsTree method.

Table S1 Primers used for PCR amplification and DNA sequencing of hydrophobin genes in various *Paxillus* strains

Table S2 Pairwise nucleotide (upper triangle matrix) and amino acid (lower triangle matrix) identity of *hydA* to *hydG* from the *Paxillus involutus* ATCC 200175

Table S3 Hydrophobin genes analysed from different strains of *Paxillus*

This material is available as part of the online article from http://www.blackwell-synergy.com/doi/abs/10.1111/j.1469-8137.2007.02022.x

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

List of Thesis

Published at Department of Ecology, Microbial Ecology, Lund University

1. **BIRGIT NORDBRING-HERTZ**, Nematode-trapping organs in the fungus *Arthrobotrys oligospora*. Formation, structure and function. 1974-03-08.

2. **BENGT SÖDERSTRÖM**, Fungi in Swedish coniferous soils. Fungal biomass and activity and microfungal species composition. 1978-04-18.

3. **ERLAND BÅÅTH**, Soil fungi in mature and clear-cut coniferous forests. 1980-11-20.

4. **TIIU KAURI**, The dynamics of bacterial populations in a forest soil. 1982-05-18.

5. **HANS-BÖRJE JANSSON**, Attraction of nematodes to nematophagous fungi. 1982-10-21.

6. **BJÖRN LUNDGREN**, Bacteria in a pine forest soil. Biomass and activity as affected by environmental factors. 1983-05-25.

7. **ANDERS NORDGREN**, Soil microbial effects of smelter induced heavy metal contamination. 1986-10-10.

8. **STEFAN OLSSON**, Effects of root exudation on growth of bacteria and fungal pathogens in the rhizosphere. 1988-02-05.

9. **CARIN DACKMAN**, Fungal parasites of cyst nematodes. 1988-12-16.

10. **SUSANNE ERLAND**, Effects of liming on pine ectomycorrhiza. 1990-12-14.

11. **KRISTINA ARNEBRANT**, Effects of forest fertilization on soil microorganisms. 1991-05-24.

12. **YVONNE PERSSON**, Mycoparasitism by the nematode-trapping fungus *Arthrobotrys oligospora*. 1991-10-18.

13. **EVA FRIMAN**, The adhesive traps of two nematophagous fungi. 1993-05-26.

14. **ÅSA FROSTEGÅRD**, Phospholipid fatty acid analyses to detect changes in soil microbial community structure. 1995-01-20.

15. **HELENA ÅKESSON**, Infection of barley by *Bipolaris sorokiniana*: toxin production and ultrastructure. 1995-05-05.

16. **SOLBRITT ANDERSSON**, Ectomycorrhizal activity as affected by soil liming. 1996-05-10.

17. **STEFAN ROSÉN**, Fungal Lectins: Molecular structure and function of a member of a novel lectin family. 1996-06-07.

18. **LOTTA PERSMARK**, Ecology of nematophagous fungi in agricultural soils. 1997-03-21.

19. **PÅL-AXEL OLSSON**, The external mycorrhizal mycelium-growth and interactions with saprophytic microorganisms. 1998-01-29.

20. **JOHAN ÅHMAN**, Extracellular serine proteases as virulence factors in nematophagous fungi. Molecular charterization and functional analysis of the PII protease in *Arthrobotrys oligospora*. 2000-03-10.

21. **SHAHID MAHMOOD**, Ectomycorrhizal community structure and function in relation to forest residue harvesting and wood ash applications. 2000-05-16.

22. **DACE APOGA**, Infection biology of the plant pathogenic fungus *Bipolaris sorokiniana*. 2000-06-10.

23. **KARIN TORNBERG**, Wood-Decomposing Fungi: Soil colonization, effects on indigenous bacterial community in soil and hydroxyl radical production. 2001-11-30.

24. **DAG AHRÉN**, Genomic diversity and evolution of parasitism in nematode-trapping fungi. 2002-01-18.

25. **INGRID VAN AARLE**, The ecophysiology of arbuscular mycorrhizal fungi: Phosphatase activity associated with extraradical and intraradical mycelium. 2002-06-03.

26. **KATARINA SÖDERBERG**, Bacterial activity and community composition in the rhizosphere: influence of plant species, root age, nitrogen addition and mycorrhizal status. 2003-03-07.

27. **DAVID HAGERBERG**, The growth of external ectomycorrhizal mycelia in the field in relation to host nutrient status and local addition of mineral sources. 2003-03-28.

28. **MARIE PETTERSSON**, Factors affecting rates of change in soil bacterial communities. 2004-02-26.

29. **ANTOINE LE QUÉRÉ**, Genome and transcriptome analyzes in the ectomycorrhizal fungus *Paxillus involutus*. 2004-05-18.

30. **LARS OLA NILSSON**, External mycelia by mycorrhizal fungi - responses to elevated N in forest ecosystems. 2004-10-15.

31. **FREDRIK DEMOLING**, Nutrient limitation of bacterial growth in soil. 2007-02-16.

32. **MARGARETA THOLANDER**, Transcriptome analyses of the nematode-trapping fungus *Monacrosporium haptotylum*. 2007-04-19.

33. **BALAJI RAJASHEKAR**, Evolutionary genomics of symbiotic fungi. 2008-03-28.

Organization
LUND UNIVERSITY
Department of Ecology
Microbial Ecology
Ecology Building
SE-223 62, Lund, Sweden

Document name
DOCTORAL THESIS

Date of issue
28th March 2008

Sponsoring organization
1. The Swedish Research Council
2. The Research School in Genomics and Bioinformatics, Göteborg

Author
Balaji Rajashekar

Title and subtitle
Evolutionary Genomics of Symbiotic Fungi

Abstract
Ectomycorrhizae is a mutualistic association between roots of woody plants and a diverse range of soil fungi. The fungi exchange soil derived mineral nutrients for photosynthetic sugars from the host plant. The mycorrhizal symbioses are commonly found in all forest ecosystems and have a major ecological and economical importance. I have used comparative genomics, DNA microarrays and computational approaches to gain insights into the evolution of the ectomycorrhizal symbiosis in two fungi *Laccaria bicolor* (Basidiomycetes; Agaricales) and *Paxillus involutus* (Basidiomycetes; Boletales).

L. bicolor is the first symbiotic fungus to have its genome sequence determined. The genome assembly contains 65 million base pairs with about ~20,000 predicted protein-encoding genes. Here, I report the analysis of *L. bicolor* genome and its comparison with the genomes of four other basidiomycetes including the saprotrophic species *Coprinopsis cinerea* and *Phanerochaete chrysosporium*, the human pathogen *Cryptococcus neoformans* and the plant pathogen *Ustilago maydis*. The compared genomes cover about 550 million years of evolution. A total of 58,030 protein sequences from these five basidiomycetes were clustered into 7352 protein families. The evolution of protein families were analysed for accelerated rates of gain and loss along specific branches of a phylogenetic tree using a stochastic birth and death model. Analysis of the genome sequence of *L. bicolor* in comparison to other analysed basidiomycetes revealed large genome size, large number of protein families, larger size of protein families, many lineage specific and expanded families, and large number of recent duplicates. The evolution of two large and expanded protein families in *L. bicolor* having significant homology to protein kinases and Ras GTPases superfamilies were analysed in more detail. The analyses showed these families to contain many paralogs that have arisen through recent duplication events. The comparative analyses of gene families showed that the evolution of symbiosis in *L. bicolor* has been associated with the expansion of large multigene families. The functions of many of these families are unknown but many of them are differentially expressed during symbiosis.

In the second part of my thesis, I have analysed duplicated and rapidly evolving genes that could be associated with symbiotic adaptations in the ectomycorrhizal fungus *P. involutus*. Strains of *P. involutus* forming ectomycorrhiza showing various degree of host-specificity were analysed by comparative genomic hybridizations using a cDNA microarray representing 1076 putative unique genes. Approximately 17% of the genes investigated on the array were detected as rapidly and presumably non-neutrally evolving within *Paxillus*. Among these genes, there were several hydrophobins. Hydrophobins are small, secreted hydrophobic cell surface proteins having several roles in growth and development of fungi. The evolutionary mechanisms responsible for generating sequence and expression divergence among members of the hydrophobin multigene family in *P. involutus* were examined in more detail.

Key words: Basidiomycetes, Comparative Genomics, Ectomycorrhizae, Evolution, Fungi, Hydrophobin, *Laccaria bicolor*, *Paxillus involutus*, Protein family, Symbiosis

Classification system and/or index termes (if any):		Language: English
Supplementary bibliographical information:		ISBN: 978-91-7105-267-4
ISSN and key title:	Number of pages: 128	Price
Recipient's notes	Security classification	

Distribution by
Balaji Rajashekar
Department of Ecology, Microbial Ecology, Ecology Building, SE-223 62, Lund, Sweden

Signature _____ Date 21 November 2007