Current Biology, Volume *26*

Supplemental Information

Expansion of Signal Transduction Pathways in Fungi by Extensive Genome Duplication

Luis M. Corrochano, Alan Kuo, Marina Marcet-Houben, Silvia Polaino, Asaf Salamov, José M. Villalobos-Escobedo, Jane Grimwood, M. Isabel Álvarez, Javier Avalos, Diane Bauer, Ernesto P. Benito, Isabelle Benoit, Gertraud Burger, Lola P. Camino, David Cánovas, Enrique Cerdá-Olmedo, Jan-Fang Cheng, Angel Domínguez, Marek Eliáš, Arturo P. Eslava, Fabian Glaser, Gabriel Gutiérrez, Joseph Heitman, Bernard Henrissat, Enrique A. Iturriaga, B. Franz Lang, José L. Lavín, Soo Chan Lee, Wenjun Li, Erika Lindquist, Sergio López-García, Eva M. Luque, Ana T. Marcos, Joel Martin, Kevin McCluskey, Humberto R. Medina, Alejandro Miralles-Durán, Atsushi Miyazaki, Elisa Muñoz-Torres, José A. Oguiza, Robin A. Ohm, María Olmedo, Margarita Orejas, Lucila Ortiz-Castellanos, Antonio G. Pisabarro, Julio Rodríguez-Romero, José Ruiz-Herrera, Rosa Ruiz-Vázquez, Catalina Sanz, Wendy Schackwitz, Mahdi Shahriari, Ekaterina Shelest, Fátima Silva-Franco, Darren Soanes, Khajamohiddin Syed, Víctor G. Tagua, Nicholas J. Talbot, Michael R. Thon, Hope Tice, Ronald P. de Vries, Ad Wiebenga, Jagjit S. Yadav, Edward L. Braun, Scott E. Baker, Victoriano Garre, Jeremy Schmutz, Benjamin A. Horwitz, Santiago Torres-Martínez, Alexander Idnurm, Alfredo Herrera-Estrella, Toni Gabaldón, and Igor V. Grigoriev

Supplemental Figures and Tables

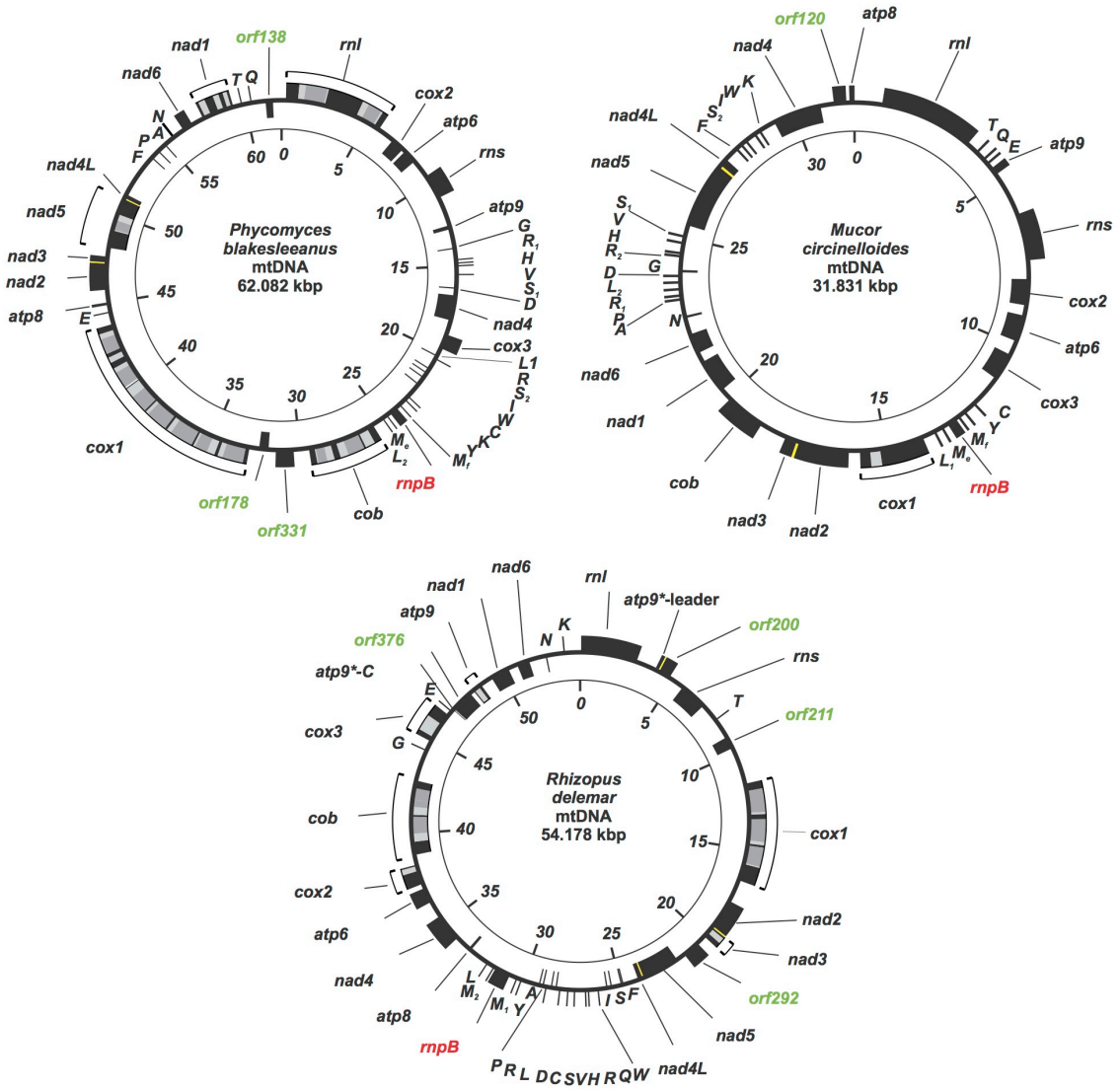

Figure S1 (related to Figure 1). Gene maps of Mucoromycotina mitochondrial DNAs.
Black boxes represent genes. Transfer RNAs are indicated by the one-letter code of their cognate amino acid. Genes with several copies are distinguished by subscript numbers. Me, Mf, genes coding for elongator and initiator tRNAMet, respectively. Gene-name colors depict taxonomic gene distribution. Black, common genes found in most animal, fungal, and protist mtDNAs; grey portions represent introns, with intronic open reading frames (ORFs) shown in dark grey. Red, expanded structural RNA gene set mostly present in mtDNAs of protists and plants and rarely in fungi. Green, hypothetical protein-coding genes that are unique to a given organism; the number specifies the amino acid count in the ORF. Genes on the outer circle are transcribed in a clockwise direction, those on the inner circle are transcribed counter clockwise. The innermost circle serves as size marker.

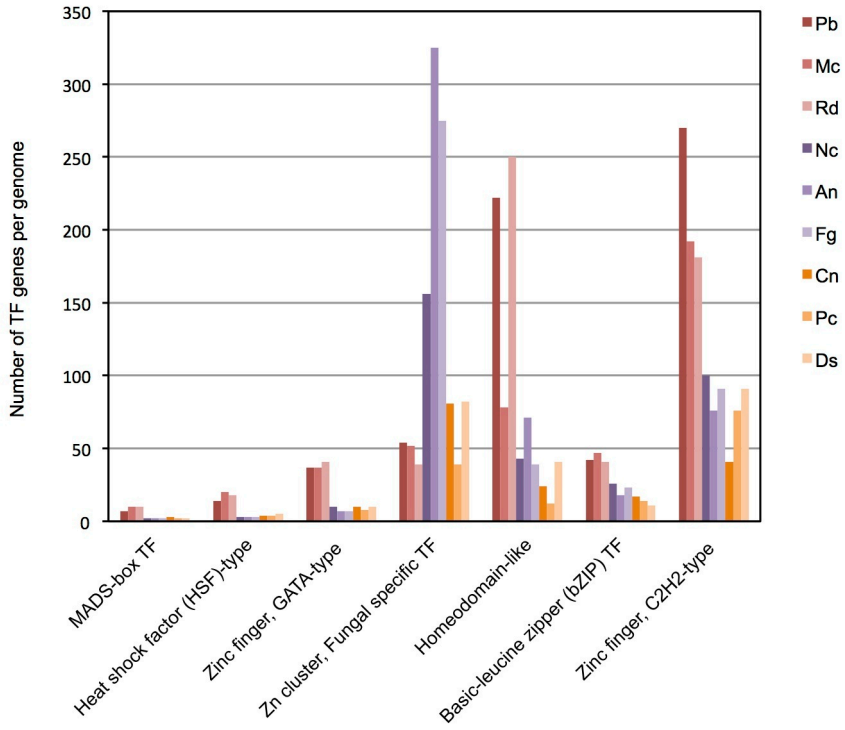

Figure S2 (related to Figure 2). Expanded and under-represented families of transcription factors in *P. blakesleeanus*, *M. circinelloides* and other fungi.
The plot shows the number of transcription factors (TF) for each family in representative fungi of the Mucoromycotina, Ascomycota and Basidiomycota fungi. Pb, *Phycomyces blakesleeanus*; Mc, *Mucor circinelloides*; Rd, *Rhizopus delemar*; Nc, *Neurospora crassa*; An, *Aspergillus nidulans*; Fg, *Fusarium graminearum*; Cn, *Cryptococcus neoformans*; Pc, *Phanerochaete chrysosporium*; Ds, *Dichomitus squalens*.

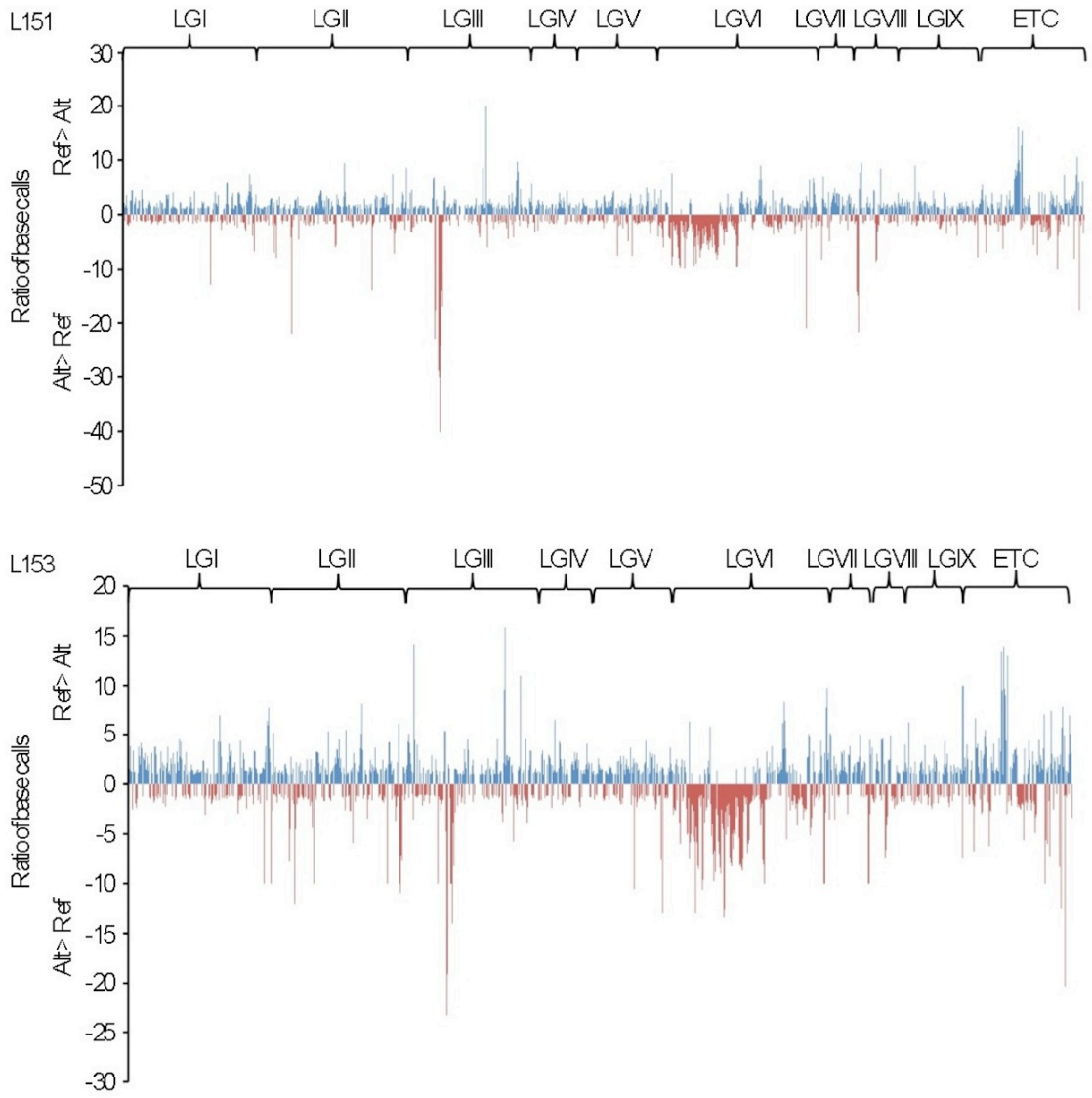

Figure S3 (related to Figure 4). Heterozygosity in the *madI* strains L151 and L153.
The graphs show the distribution of heterozygous SNPs as the ratio of the number of reads specifying the most numerically prevalent base call to the number of reads specifying the alternate base call. The horizontal axis corresponds to the distribution along linkage groups and is not to scale. The vertical axis shows the ratio of the alternate base calls with positive (blue) values indicating that the reference genome base call is more frequent and negative (red) values indicating that the alternate base call is more frequent. LG= Linkage Group, ETC = scaffolds not assigned to linkage groups.

Table S1 (related to Figure 1). Main features of the *P. blakesleeanus* and *M. circinelloides* genomes.

S1A. Main features of the *P. blakesleeanus* and *M. circinelloides* genome assemblies

Nuclear Genome Assembly	***Phycomyces***	***Mucor***
Main genome scaffold total	80	26
Main genome contig total	350	26
Main genome scaffold sequence total	53.9 Mbp	36.6 Mbp
Main genome contig sequence total	53.4 Mbp	36.6 Mbp
Estimated % sequence bases in gaps	1.1%	0.0%
Main genome scaffold N50 / L50	11/1.5 Mbp	4/4.3 Mbp
Main genome contig N50 / L50	41/370.4 kbp	4/4.3 Mbp
Number of scaffolds >50 Kbp	51	15
% main genome in scaffolds >50	99.4%	99.7%
% ESTs aligned to scaffolds	98%	95%

S1B. Genomic libraries included in the *P. blakesleeanus* and *M. circinelloides* whole genome assembly and their respective assembled sequence coverage levels.

	Library Type	**Average insert size, pb**	**Number of reads**	**Sequence coverage (x)**
Phycomyces	Sanger, 3kb	2,834	329,772	2.59
	Sanger, 8kb	6,474	400,608	4.29
	Sanger, fosmid1	35,155	27,264	0.26
	Sanger, fosmid2	35,230	46,080	0.35
	Total		803,724	7.49
Mucor	Sanger, 3kb	2,527	216,830	3.60
	Sanger, 8kb	6,516	340,704	5.02
	Sanger, fosmid	35,117	82,175	0.87
	Total		639,709	9.49

S1C. *P. blakesleeanus* and *M. circinelloides* filtered models classified by gene prediction method.

Prediction method	***Phycomyces* gene models**	***Mucor* gene models**
Total	16,528	11,719
Ab initio	8,494 (51%)	4,839 (41%)
Protein-based	7,261 (44%)	4,901 (41%)
EST-based	773 (5%)	1,979 (17%)

S1D. Properties of *P. blakesleeanus*, *M. circinelloides*, and *R. delemar* gene models.

Property or number	***Phycomyces***	***Mucor***	***Rhizopus***
Avg. gene length	1628 nt	1429 nt	1212 nt
Avg. transcript length	1128 nt	1176 nt	1028 nt
Avg. protein length	353 aa	379 aa	343 aa
Avg. exon length	249 nt	311 nt	310 nt
Avg. intron length	143 nt	93 nt	81 nt
Avg. exon frequency/gene	4.5	3.8	3.3
# multiexon genes	14640 (89%)	9701 (83%)	13540 (78%)
Mean CDS GC content	43.3%	46.3%	40.2%
Mean intron GC content	29.3%	40.1%	30.0%
# genes with similarity to protein in nr	11116 (67%)	9597 (82%)	12303 (70%)
# genes with similarity to gene in same genome	12517 (76%)	8366 (72%)	13776 (79%)
# genes with similarity to gene in other two Mucoromycotina	9632 (58%)	8923 (76%)	11658 (67%)
# genes with EST coverage	5461 (33%)	3897 (33%)	ND
# genes with Pfam domain	7191 (44%)	7524 (64%)	7483 (43%)
# genes with signal peptide	1818 (11%)	1498 (13%)	1546 (9%)
# genes with transmembrane domain	2682 (16%)	2101 (18%)	2133 (12%)
# genes with EC number	2328 (14%)	2596 (22%)	2784 (16%)
# genes with GO term	6408 (39%)	6543 (56%)	8452 (48%)

S1E. Repetitive sequence content of the *P. blakesleeanus* and *M. circinelloides* genomes. Values in parentheses represent the percentage of the genome occupied by the respective repeat family.

Transposable Element Family	***Phycomyces***	***Mucor***
Class I	38 (5.4%)	17 (5.3%)
LTR	9	5
LINE	12	8
Class I Unclassified	17	4
Class II	161 (16.7%)	34 (1.7%)
CMC-EnSpm	31	0
TC1 Mariner	29	3
MuLE	19	0
PIF-Harbinger	9	0
Helitron	4	1
Merlin	1	0
Crypton	1	0
Class II Unclassified	67	30
Other Unclassified Repeats	36 (6.0%)	23 (10.4%)
Total Number of Families	235 (28.1%)	129 (17.4%)

S1F. rRNA genes in the genomes of *P. blakesleeanus*, *M. circinelloides*, *R. delemar*, and *N. crassa*.

	5S	**18S**	**5.8S/ITS**	**28-26S**
P. blakesleeanus	8	6	7	5
M. circinelloides	5	3	3	3
R. delemar	2	2	3	2
N. crassa	79	>40	43	≈ 52

S1G. EST matches to TE families in the *P. blakesleeanus* and *M. circinelloides* genomes.

Repeat Family	***Phycomyces***	***Mucor***
Class I		
LTR	7/9	2/5
LINE	6/12	5/8
Class I Unclassified	11/17	2/4
Class II		
CMC-EnSpm	24/31	0
TC1 Mariner	11/29	0/3
MuLE	6/19	0
PIF-Harbinger	3/9	0
Helitron	2/4	1/1
Merlin	0/1	0
Crypton	0/1	0
Class II Unclassified	17/67	4/30
Other Unclassified Repeats	19/36	8/23
Total	106/235 (45%)	22/129 (17%)

Table S2 (related to Figure 1). Analysis of duplications in fungal genomes.

S2A. Analysis of duplicated genes in fungal genomes. Number of proteins in gene families (540 MCL gene clusters) and fraction of gene families with more members than average for a selection of fungal genomes. Full species names are provided in the section "Clustering of fungal genes in families" of the Supplemental Experimental Procedures.

		Number of proteins per family	Families with more members than average (%)
Mucoromycotina	*P. blakesleeanus*	3.0	54.3
	M. circinelloides	3.2	59.8
	R. delemar	3.6	68.0
	L. corymbifera	2.9	50.2
Mortierellomycotina	*M. alpina*	2.4	43.1
Ascomycota	*S. cerevisiae*	2.0	25.4
	S. pombe	1.9	22.4
	N. crassa	1.9	14.6
	A. nidulans	2.0	16.9
	A. niger	2.1	22.4
	S. sclerotiorum	1.8	13.3
	C. globosum	1.8	13.1
	F. graminearum	2.1	18.3
	P. chrysogenum	2.1	18.9
	M. graminicola	1.9	16.3
	C. heterostrophus	2.0	18.9
	C. grayi	1.9	19.6
	X. parietina	1.8	12.2
	T. melanosporum	1.8	15.9
Basidiomycota	*C. neoformans*	1.7	9.3
	U. maydis	1.7	7.4
	T. mesenterica	1.7	6.9
	S. roseus	1.8	11.1
	P. graminis	1.9	20.2
	M. globosa	1.6	8.3
	P. chrysosporium	2.0	17.0
	W. cocos	2.0	15.7
	T. versicolor	2.0	14.8
	D. squalens	1.9	20.2
	H. annosum	1.9	12.8
	L. bicolor	2.2	22.2
	S. commune	2.0	17.8
	S. lacrymans	1.9	12.6
Chytridiomycota	*B. dendrobatidis*	1.9	16.7

S2B. Predicted duplicated regions (paralogons) in fungal genomes. Weighted number of detected duplicated regions, with number of genes in each duplicated region divided by 3 - the minimum number of orthologs in the paralogon, the P-values for real and maximal observed number of weighed duplicated regions in 1000 simulations, the total length of duplicated regions, and the fraction of duplicated regions to the total genome length using gene-based synteny and gene boundaries to extract coordinates. To identify duplicated regions we requested that at least 10% of genes in duplicated regions were homologous (see Supplemental Experimental Procedures). This value resulted in an overestimation of duplicated DNA for some genomes but allowed the best prediction in a set of duplicated regions in the *S. cerevisiae* genome.

		Number of duplicated regions	P-value		Length of duplicated regions (Mb)	Fraction of duplicated regions in the genome (%)
			Real	Max		
Mucoromycotina	*P. blakesleeanus*	123.0	<0.001	7	7.939	14.70
	M. circinelloides	59.7	<0.001	9	2.857	7.80
	R. delemar	624.0	<0.001	8	35.593	77.20
	L. corymbifera	169.3	<0.001	12	8.276	24.60
Mortierellomycotina	*M. alpina*	309.0	<0.001	4	16.564	43.20
Ascomycota	*S. cerevisiae*	156.7	<0.001	11	8.469	70.20
	N. crassa	1.0	0.130	5	0.018	0.04
	A. nidulans	11.3	<0.001	6	0.445	1.50
	A. niger	12.3	<0.001	7	0.765	2.20
	S. sclerotiorum	8,3	<0.001	5	0.279	0.70
	C. globosum	9.0	<0.001	4	0.437	1.30
	F. graminearum	8.7	<0.001	4	0.567	1.60
	P. chrysogenum	17.3	0.001	87	1.441	4.60
	M. graminicola	6.0	<0.001	5	0.648	1.60
	C. heterostrophus	133.7	<0.001	6	5.314	14.60
	C. grayi	48.3	<0.001	4	0.909	2.30
	X. parietina	8.7	<0.001	5	0.628	2.00
	T. melanosporum	3.7	<0.001	2	0.372	0.30
Basidiomycota	*C. neoformans*	5.7	0.003	8	0.458	2.40
	U. maydis	2.0	0.002	3	0.206	1.00
	T. mesenterica	3.3	0.008	5	0.264	0.90
	S. roseus	6.3	0.002	7	0.311	1.50
	P. graminis	252.7	<0.001	5	10.862	12.30
	M. globosa	0	1.000	5	0	0
	P. chrysosporium	10.0	<0.001	9	0.537	1.50
	W. cocos	16.3	<0.001	6	1.281	2.50
	T. versicolor	32.0	<0.001	6	1.378	3.10
	D. squalens	16.0	<0.001	5	0.756	1.80
	H. annosum	35.3	<0.001	5	1.425	4.20
	L. bicolor	180.3	<0.001	5	9.446	15.60
	S. commune	46.0	<0.001	5	1.821	4.70
	S. lacrymans	21.3	<0.001	6	1.007	2.40
Chytridiomycota	*B. dendrobatidis*	22.7	<0.001	6	0.552	2.30

S2C. Predicted duplicated genes in fungal genomes. The table shows the number of duplicated genes in each genome, the fraction of duplicated genes relative to the total number of genes in the genome, the average number of duplicated genes in duplicated regions (paralogons), and the fraction of lineage-species genes in paralogons (*i.e.* genes present only in one species).

		Number of duplicated genes	Fraction of duplicated genes (%)	Average duplicated genes in duplicated regions	Lineage-specific genes (%)
Mucoromycotina	*P. blakesleeanus*	738	4.5	4	39.5
	M. circinelloides	358	3.1	4	17.2
	R. delemar	3744	21.4	13	19.2
	L. corymbifera	1016	8.2	5	27.6
Mortierellomycotina	*M. alpina*	1854	12.7	6	51.3
Ascomycota	*S. cerevisiae*	940	16.0	8	36.6
	N. crassa	6	0.1	3	46.1
	A. nidulans	68	0.6	3	5.0
	A. niger	74	0.7	3	16.9
	S. sclerotiorum	50	0.3	4	85.5
	C. globosum	54	0.5	5	46.4
	F. graminearum	52	0.4	3	43.6
	P. chrysogenum	104	0.9	4	40.0
	M. graminicola	36	0.3	4	75.0
	C. heterostrophus	802	6.0	10	45.8
	C. grayi	290	2.5	3	30.7
	X. parietina	52	0.5	3	25.8
	T. melanosporum	22	0.3	4	31.8
Basidiomycota	*C. neoformans*	34	0.5	4	65.6
	U. maydis	12	0.2	3	61.5
	T. mesenterica	20	0.2	3	63.0
	S. roseus	38	0.7	4	55.1
	P. graminis	1516	7.4	7	73.3
	M. globosa	0	0	0	-
	P. chrysosporium	60	0.6	3	40.8
	W. cocos	98	0.8	4	30.3
	T. versicolor	192	1.3	4	34.8
	D. squalens	96	0.8	3	24.9
	H. annosum	212	1.6	5	43.6
	L. bicolor	1082	4.7	7	67.9
	S. commune	276	1.9	5	59.7
	S. lacrymans	128	1.0	4	38.9
Chytridiomycota	*B. dendrobatidis*	136	1.5	4	78.0

S2D. Duplicated genes in phylogenies. Number of duplicated genes in phylogenies between Mucoromycotina fungi (M), and non-Mucoromycotina fungi (nonM) after individually comparing each Mucoromycotina genome with each of 28 genomes from Dikarya. Mucoromycotina genomes are statistically significantly enriched in 2(M):1(nonM) triplets relative to 2(nonM):1(M) triplets (Pearson chi-square<0.001).

	2(M):1(nonM)	2(nonM):1(M)
P. blakesleeanus	4181	1139
M. circinelloides	6210	1115
R. delemar	7637	893
L. corymbifera	5928	1076

Supplemental Experimental Procedures

Fungal strains, cultures, and crosses

The strains of *P. blakesleeanus* and *M. circinelloides* are listed below. Cultures of *P. blakesleeanus* were inoculated on minimal medium agar plates and grown at 22 °C in the dark for two days, exposed to light during 30 min or kept in the dark as control [S1]. To isolate RNAs from sporangiophores, cultures were initiated as described above and were grown for two days in the dark at 22 ºC. Then mycelium was exposed to light for two minutes to induce sporangiophore initiation and returned to the dark for another day. Sporangiophores were then either exposed to light for 30 minutes or kept in the dark, removed, and stored at -80 ºC before RNA extraction. Cultures of *M. circinelloides* were inoculated on cellophane sheets on solid YNB medium and grown for 18 hours at 26 °C in the dark, exposed to light for 20 minutes or kept in the dark as control [S2]. Crosses were performed with *P. blakesleeanus* [S3], and the efficiency of phototropism was measured by the bending of sporangiophores to unilateral white light. We selected a light intensity that allowed us to distinguish the phototropic responses of the wild type from the *madA* and *madI* mutants using the detailed characterization of phototropism in these strains performed by Álvarez et al. [S4].

Strains used in this work.

The *P. blakesleeanus mad* mutations were all induced using chemical mutagenesis.

Species and name	**Genotype**
Phycomyces blakesleeanus	
NRRL1555/ FGSC 10004	Wild type (−)
UBC21/ FGSC 10459	Wild type (+)
A893/ FGSC 25052	*madA403* (−)
A905/ FGSC 10435	*madC406* (−)
A909 / FGSC 25053	*madJ407* (−)
B2/ FGSC 10430	*madC452* (−)
B16	*madD462* (−)
C6 / FGSC 10450	*carRA12 madF48* (−)
C21/ FGSC 10434	*madA7 pde*-1 (−)
C47 / FGSC 25066	*madA35* (−)
C68 / FGSC 10447	*madD59* (−)
C107	*madD99 madG131* (−)
C109 / FGSC 10432	*madB101* (−)
C110 / FGSC 10449	*madE102* (−)
C149 / FGSC 10448	*madD120* (−)
C307 / FGSC 10451	*madG131* (−)
L51 / FGSC 25074	*madA7 madB103* (−)
L151 / FGSC 10453	*madI714* (−)
L153 / FGSC 10454	*madI716* (−)
L157	*madE720* (−)
L161	*madF724* (−)
L163	*madE726* (−)
Mucor circinelloides	
CBS277.49	Wild type

DNA and RNA purification

Genomic DNA from *P. blakesleeanus* or *M. circinelloides* was purified from mycelia [S5, S6]. RNA was isolated from *P. blakesleeanus* or *M. circinelloides* using the Perfect RNA eukaryotic mini kit (Eppendorf) or the RNeasy Plant Mini Kit (Qiagen). Poly A+ RNA was isolated from total RNA using the Absolutely mRNA Purification kit (Stratagene).

Sequencing and assembly
The genomes of *P. blakesleeanus* and *M. circinelloides* were sequenced using Sanger sequencing on ABI 3730XL capillary machines, using three libraries with different insert length that were sequenced from both ends (Table S1). Totals 639,709 reads for *M. circinelloides* (9.49x coverage) and 803,624 reads (7.49x coverage) for *P. blakesleeanus* were collected. Reads were assembled using a modified version of Arachne [S7] v.20071016 with parameters maxcliq1=100, correct1_passes=0 and BINGE_AND_PURGE=True. The resulting whole genome shotgun assemblies were then used as the basis for finishing the genomes. The genomes of the *P. blakesleeanus mad* mutant strains were sequenced using Illumina technology. cDNA synthesis and cloning was a modified procedure based on the "SuperScript plasmid system with Gateway technology for cDNA synthesis and cloning" (Invitrogen). Expressed Sequence Tags (ESTs) were processed through the JGI EST pipeline.

Genome sequence improvement
To perform genome sequence improvement, the *M. circinelloides* and *P. blakesleeanus* whole genome shotgun assemblies were broken down into scaffolds and each scaffold reassembled with phrap. These scaffold were then improved using our Phred/Phrap/Consed pipeline. Initially all low quality regions and gaps were targeted with computationally selected sequencing reactions completed with 4:1 BigDye terminator: dGTP chemistry (Applied Biosystems). These automated rounds included walking on plasmid subclones using custom primers. Following completion of the automated rounds, a trained finisher manually inspected each assembly. This examination included a visual examination of subclone paired ends and visual inspection of high quality discrepancies and all remaining low-quality areas. Further reactions were then manually selected to improve the genome sequence. These reactions included additional custom primer walks on plasmid subclones and fosmids. The reactions were completed using 4:1 BigDye terminator:dGTP chemistry. Smaller repeats in the sequence were resolved by transposon-hopping 8 kb plasmid clones. Fosmid clones were shotgun sequenced and finished to fill large gaps and resolve larger repeats.

cDNA library construction and sequencing
Poly A+ RNA was isolated from total RNA using the Absolutely mRNA Purification kit (Stratagene). cDNA synthesis and cloning was a modified procedure based on the "SuperScript plasmid system with Gateway technology for cDNA synthesis and cloning" (Invitrogen). 1-2 µg of poly A+ RNA, reverse transcriptase SuperScript II (Invitrogen) and oligo dT-NotI primer (5' GACTAGTTCTAGATCGCGAGCGGCCGCCCT15VN 3') were used to synthesize first strand cDNA. Second strand synthesis was performed with *E. coli* DNA ligase, polymerase I, and RNaseH followed by end repair using T4 DNA polymerase. The SalI adaptor (5' TCGACCCACGCGTCCG and 5' CGGACGCGTGGG) was ligated to the cDNA, digested with NotI (New England Biolabs), and subsequently size selected by gel electrophoresis (1.1% agarose). The cDNA inserts were directionally ligated into the SalI and NotI digested vector pCMVsport6 (Invitrogen). The ligation was transformed into ElectroMAX T1 DH10B cells (Invitrogen).

Library quality was first assessed by randomly selecting 24 clones and PCR amplifying the cDNA inserts with the primers M13-F (5' GTAAAACGACGGCCAGT) and M13-R (5' AGGAAACAGCTATGACCAT) to determine the fraction of insertless clones. Colonies from each library were plated onto agar plates (254 mm plates from Teknova) at a density of approximately 1000 colonies per plate. Plates were grown at 37 °C for 18 hours, then individual colonies were picked and each used to inoculate a well containing LB media with appropriate antibiotic in a 384 well plate (Nunc). Clones in 384 well plates were grown at 37 °C for 18 hours. Plasmid DNA for sequencing was produced by rolling circle amplification [S8] (TempliPhi, GE Healthcare). Subclone inserts were sequenced from both ends using primers complimentary to the flanking vector sequence (Fw: 5'ATTTAGGTGACACTATAGAA, Rv: 5' TAATACGACTCACTATAGGG) and Big Dye terminator chemistry then run on ABI 3730 instruments (Applied Biosystems).

EST sequence processing and assembly
Expressed Sequence Tags (ESTs) were processed through the JGI EST pipeline. ESTs were generated in pairs, a 5' and a 3' end read from each cDNA clone. To trim vector and adaptor sequences, common sequence patterns at the ends of ESTs were identified and removed using an internally developed tool. Insertless clones were identified if either of the following criteria were met: >200 bases of vector sequence at the 5' end or less than 100 bases of non-vector sequence remained. ESTs were then trimmed for quality using a sliding window trimmer (window = 11 bases). Once the average quality score in the window was below the threshold (Q15) the EST was split and the longest remaining sequence segment was retained as the trimmed EST. EST sequences with less than 100 bases of high quality sequence were removed. ESTs were evaluated for the presence of polyA or polyT tails (which, if present, were removed) and the EST reevaluated for length, removing ESTs with less than 100 bases remaining. ESTs consisting of more than 50% low complexity sequence were also removed from the final set of "good ESTs". In the case of resequencing the same EST, the longest high quality EST was retained. Sister ESTs (end pair reads) were categorized as follows: if one EST was insertless or a contaminant then by default

the second sister was categorized as the same. However, each sister EST was treated separately for complexity and quality scores. Finally, EST sequences were compared to the GenBank nucleotide database in order to identify contaminants; non-desirable ESTs such as those matching non-cellular and rRNA sequences were removed.

For clustering, ESTs were evaluated with malign, a kmer based alignment tool, which clusters ESTs based on sequence overlap (kmer = 16, seed length requirement = 32 alignment ID>=98%). Clusters of ESTs were further merged based on sister ESTs using double linkage. Double linkage requires that 2 or more matching sister ESTs exist in both clusters to be merged. EST clusters were then each assembled using CAP3 [S9] to form consensus sequences. Clusters may have more than one consensus sequence for various reasons to include; the clone has a long insert, clones are splice variants or consensus sequences are erroneously not assembled. Cluster singlets are clusters of one EST, whereas CAP3 singlets are single ESTs which had joined a cluster but during cluster assembly were isolated into a separate singlet consensus sequence. ESTs from each separate cDNA library were clustered and assembled separately and subsequently the entire set of ESTs for all cDNA libraries were clustered and assembled together. For cluster consensus sequence annotation, the consensus sequences were compared to SwissProt using BLASTx.

Genome annotation and sequence analysis

Genome assemblies of *P. blakesleeanus* NRRL1555 and *M. circinelloides* CBS277.49 were annotated using the JGI Annotation Pipeline and a subset of genes were curated manually. This led to a filtered set of 16,528 *P. blakesleeanus* and 11,719 *M. circinelloides* gene models with their properties and support by different lines of evidence, summarized in Table S1. Repetitive DNA was annotated using the TEdenovo pipeline and manual curation. All predicted gene models were functionally annotated by the JGI Annotation Pipeline against highly curated databases. Automated mtDNA annotation was performed with MFannot (http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl) and validated manually, in particular regarding intron-exon boundaries and termini of ribosomal RNA genes.

The genome assemblies of *P. blakesleeanus* NRRL1555 and *M. circinelloides* CBS277.49 were annotated using the JGI Annotation Pipeline, which combines several gene predictors: A) protein-based gene models were predicted using FGENESH+ [S10] and GeneWise [S11] seeded by BLASTx alignments of genomic sequence against sequences from the NCBI non-redundant protein set nr, B) ab initio gene models were predicted using GeneMark [S12] and FGENESH, the latter trained on the set of putative full-length genes and reliable protein-based models, and C) transcriptome-based gene models were derived by assembling transcript sequences which were then modelled on genomic sequence. GeneWise models were completed using scaffold data to find start and stop codons. Transcriptome alignments to the genome were used to verify, complete, and extend the gene models. Because multiple gene models per locus were often generated, a single representative gene model for each locus was chosen based on protein similarity and transcriptome support, and used for further analysis. This led to a filtered set of 16,528 *P. blakesleeanus* and 11,719 *M. circinelloides* gene models with their properties and support by different lines of evidence summarized in Table S1. A fraction of the gene models were manually curated by scientist experts in each gene category. The manually curated gene models can be accessed at the JGI genome database.

All predicted gene models were functionally annotated by the JGI Annotation Pipeline using InterProScan [S13] and hardware-accelerated double-affine Smith-Waterman alignments (http://www.timelogic.com/) against highly curated databases such as SwissProt [S14], KEGG [S15], and Pfam [S16]. KEGG hits were used to map EC numbers [S17], and InterPro, KEGG, and SwissProt hits were used to map GO terms [S18]. In addition, predicted proteins were annotated according to KOG classification [S19]. Protein targeting predictions were made with signalP [S20] and TMHMM [S21]. Finally, all proteins of each single genome were aligned by BLASTp to proteins in the GenBank nr database and to each other; after the latter procedure the alignment scores were used as a distance metric for clustering by MCL (http://www.micans.org/mcl/) into a first draft of 2,117 *P. blakesleeanus* and 2,066 *M. circinelloides* multigene families. The same method was used to group the *P. blakesleeanus* and *M. circinelloides* proteins with each other as well as with those of *R. delemar* strain 99-880, which were downloaded from the Broad Institute (http://www.broadinstitute.org/annotation/genome/rhizopus_oryzae/) on April 21, 2006. For comparative purposes, the Broad Institute's *R. delemar* supercontigs and genes were imported into the JGI database and were functionally annotated by the JGI Annotation Pipeline in the manner described for *P. blakesleeanus* and *M. circinelloides*.

Based on results from BLAST on the 18 genomes used in the phylome analyses, we divided the Mucoromycotina genes into the following categories: ancestral, fungal-specific, Mucoromycotina-specific or species-specific (Figure 1). Mucoromycotina species contained an average of 5014 genes that had homologs in the two non-fungal out-groups, deeming them of ancestral origin. This number is double the average found in the other fungal species (2320 on average), likely due to the whole genome duplications that occurred in this group. A similar trend is observed in genes of fungal origin (2737 genes versus 1381). Very few of those genes are

found in all of the fungal species: this is mainly due to the presence of microsporidian genomes in our dataset and partly because most of the widespread genes were also found in the outgroups. An average of 840 genes were shared between all four Mucoromycotina species but this number increased to 2568 if we consider all genes that were at least in two members of this group. Surprisingly, despite the fact that *R. delemar* underwent a species-specific whole genome duplication, *P. blakesleeanus* is the Mucoromycotina species with the largest number of species-specific genes (5183). On the contrary, *M. circinelloides* only contained 1867 genes that could only be found in this species.

Genome resequencing

For each strain, genomic DNA (1µg) was sheared into ~250bp fragments using the Covaris E210 (Covaris). The DNA fragments were subjected to end repair, A-tailing, and ligation of Illumina compatible adapters (IDT). The final library was enriched with 10 cycles of PCR (NEB). The prepared libraries were quantified using KAPA Biosystem's next-generation sequencing library qPCR kit and run on a Roche LightCycler 480 real-time PCR instrument. The quantified libraries were then prepared for sequencing on the Illumina GA sequencing platform utilizing a paired-end cluster generation kit, v4, and Illumina's cBot instrument to generate clustered flowcells for sequencing. Sequencing of the flowcells was performed on the Illumina GAIIx sequencer using SBS sequencing kits, v4, following a 2x76 run recipe. The libraries were sequenced on Illumina GA II sequencers generating 75bp paired end reads. These reads were aligned to the reference genome and putative SNPs and small indels were called using maq-0.7.1 [S22]. Putative structural variants were called using BreakDancer [S23], filtering for a confidence score of >90.

Identification and analysis of repetitive elements

Repetitive DNA was annotated with REPET [S24, S25] using the TEdenovo pipeline included in REPET to detect and classify repeat families. The repeat family consensus sequences were manually curated by performing BLAST searches to the RepBase protein database [S26] and manually validating the repeat classifications. The REPET pipeline TEannot was used to annotate the repeat families. TE consensus sequences were aligned to the EST consensus with megaBLAST, using an e-value threshold of 10^{-5}. rRNA genes were identified with the program RNAmmer [S27].

***P. blakesleeanus* RNAseq and analysis**

Libraries for RNAseq were prepared with RNA from mycelia and sporangiophores of the wild-type and the *madA madB* mutant strains using the SOLiD whole transcriptome kit protocol (Applied Biosystems) [S28]. Three biological replicates for each condition and two technical replicates of the experiment were performed. Sequencing was performed on a SOLiD 4 ABI sequencer. 50-base-pair-long reads were obtained and mapped to the *P. blakesleeanus* genome V2 using bowtie [S29], searching for end-to-end hits with at most three mismatches. Alignment results were recorded in BAM format for further downstream analysis. Read counts per gene were calculated for each library using a shell script in the statistical software R, and collapsed into a table considering both technical replicas.

Only genes with at least three reads per million were considered for differential expression analyses; this was done using the edgeR package [S30]. Normalization of the data was performed using the Cox-Reid adjusted likelihood method. We determined that the WTMD3, WTSD3, WTSL2 and ΔL51SD2 libraries had greater dispersion compared to their respective biological replicas, so it was decided not to consider these data for the final analysis. For determining differential expression between the different comparisons we used the generalized linear model likelihood ratio test. False discovery rates (FDR) were calculated using the Benjamini & Hochberg procedure [S31]. Genes with a FDR<0.05 were considered differentially expressed.

The protein sequences were obtained from the gene catalog of *P. blakesleeanus* V2 genome in FASTA format. BLAST comparison with this file was performed against the non-redundant database (nr) of NCBI using Blast2GO version 2.6.0 (http://www.blast2go.org/), with an E-value $\leq 1 \times 10^{-3}$. GO term assignation was performed using an E-value $\leq 1 \times 10^{-3}$, an annotation score ≤ 40, a GO weight of 5 [S32]. The annotation was improved using the protein domains databases InterProScan (http://www.ebi.ac.uk/Tools/pfa/iprscan/) and ANNEX [S33]. All GOs ancestors for each gene were obtained using an R script.

Using the GOseq package of Bioconductor, an enrichment analysis was performed with the hypergeometric distribution method. GO terms with FDR ≤ 0.05 were considered significantly enriched in each comparison. The RNAseq results have been submitted to the GEO database with accession GSE64369.

Microarray analysis of *M. circinelloides* expression

Total RNAs (0.2 µg) from each sample were amplified, labelled and hybridized to a custom Agilent Microarray (Agilent, ID G2509F). RNA labelling was performed according to the Low Input Quick Amp Labeling kit (Agilent). Images of the microarrays were acquired using a G2505C Scanner (Agilent). Raw data were obtained from images using Feature Extraction Software v. 10.7.3.1 (Agilent). All statistical and differential expression

analyses were carried out with the Limma package from Bioconductor (http://www.bioconductor.org/). Gene expression was considered statistically significant when FDR<0.05.

Clustering of fungal genes in families

The Markov clustering method (MCL) algorithm [S34] was used to cluster proteins from the following 34 fungi with complete genomes. Four Mucoromycotina: *Phycomyces blakesleeanus*, *Mucor circinelloides*, *Rhizopus delemar*, *Lichtheimia corymbifera*; 1 Mortierellomycotina: *Mortierella alpina*. 14 Ascomycota: *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Neurospora crassa*, *Aspergillus nidulans*, *Aspergillus niger*, *Sclerotinia sclerotiorum*, *Chaetomium globosum*, *Fusarium graminearum*, *Penicillium chrysogenum*, *Mycosphaerella graminicola*, *Cochliobolus heterostrophus*, *Cladonia grayi*, *Xanthoria parietina*, *Tuber melanosporum*. 14 Basidiomycota: *Cryptococcus neoformans*, *Ustilago maydis*, *Tremella mesenterica*, *Sporobolomyces roseus*, *Puccinia graminis*, *Malassezia globosa*, *Phanerochaete chrysosporium*, *Wolfiporia cocos*, *Trametes versicolor*, *Dichomitus squalens*, *Heterobasidion annosum*, *Laccaria bicolor*, *Schizophyllum commune*, *Serpula lacrymans*. One Chytridiomycota: *Batrachochytrium dendrobatidis*.

Detection of duplicated regions (paralogons)

Several methods have been developed for the identification of duplicated regions (paralogons) [S35] but, until recently, there was no way to confirm their accuracy. A manually curated set of duplicated regions in the *S. cerevisiae* genome is considered as the standard set for ancient duplicated regions [S36]. We implemented a simple approach for detecting such regions with two parameters: Nmin (minimal number of duplicated gene pairs) and Fhom (minimal fraction of homologous genes in the region). Because possible duplications in the Mucoromycotina may be even more ancient we removed the constraint of ordered genes in the algorithm. When tested on *S. cerevisiae* best accuracy was achieved at Nmin=3 and Fhom=10%, with 58 of the 67 paralogons predicted correctly, which amounts to sensitivity Sn=91% and specificity Sp=87%. At the level of detecting duplicated pairs of genes, the method predicts in total 515 pairs in paralogons, 436 of them correctly (Sn=91%, Sp=85%). We applied the above method to detect paralogons in the fungal genomes described in the previous section. We also computed potential paralogons in 1000 simulated genomes with randomly shuffled order of genes in scaffolds/chromosomes.

Detection of whole genome duplication (WGD) signatures

For each pair of genomes we concentrated only on families consisting of three members where genes with two copies duplicated after the split of the two compared species resulting in "2:1 after split" triplets. Our rationale is that the number of such triplets should be significantly larger in one genome relative to another after the genome had undergone a WGD event. For every pair of compared genomes we first selected the triplets of homologous genes (with BLASTp score at least 100 and alignment covering at least 80% of the compared genes length) and which have no other homologs in both genomes. Then for each selected triplet we constructed a phylogenetic tree and selected triplets for which duplication occurred after the split between the two genomes, *i.e.* duplicated genes are closer to each other than to the homolog in another genome.

Phylome reconstruction

The phylome - *i.e.* the complete collection of phylogenetic trees for each gene in a genome - was reconstructed for the genomes of the four Mucoromycotina species: *R. oryzae*, *P. blakesleeanus*, *M. circinelloides* and *L. corymbifera*. A fifth phylome was reconstructed using the related species, *Mortierella alpina*, as seed. All phylomes were reconstructed in the context of 13 other fungal species: two additional early diverging fungi (*Batrachochytrium dendrobatidis*, *Homoloaphlyctis polyrhiza*), three Microsporidia (*Nematocida parisii*, *Encephalitozoon cuniculi*, *Nosema ceranae*) and six Dikarya species (*Neurospora crassa*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Puccinia graminis*, *Ustilago maydis*, and *Cryptococcus neoformans*). *Nematostella vectensis* (a sea anemone) and *Monosiga brevicollis* (a choanoflagellate) were used as outgroups.

The phylome was reconstructed using an automated pipeline described previously [S37]. Briefly, for each protein encoded in the three seed genomes a Smith-Waterman search was performed against the proteome database. Results were then filtered using an e-value cut-off of $E<10^{-5}$ and requiring a continuous overlapping region of 0.5 over the query sequence. A limit of 150 BLAST hits for each protein was used. Homologous sequences were then aligned using three different programs: MUSCLE v3.8 [S38], MAFFT [S39] and kalign [S40]. Alignments were performed in forward and reverse direction and the six resulting alignments were combined with M-COFFEE [S41]. TrimAl v1.3 [S42] was then used to trim the alignment using a consistency-score cut-off of 0.1667 and a gap-score cut-off of 0.9. The alignments were then used to reconstruct maximum likelihood trees. The model best fitting the data was selected as follows: for each evolutionary model (JTT, LG, WAG, Blosum62, MtREV, VT and Dayhoff) a Neighbour Joining (NJ) tree was reconstructed as implemented in BioNJ [S43] then the likelihood of this topology was computed, allowing branch-length optimization as implemented in PhyML v3.0 [S44]; the model best fitting the data was determined by the AIC criterion [S45].

Four rate categories were used and invariant positions were inferred from the data. Branch support was computed using an aLRT (approximate likelihood ratio test) based on a chi-square distribution. The trees and alignments reconstructed for the three phylomes can be found at phylomeDB [S46] (http://phylomedb.org), with the phylomeIDs 252, 253, 254, 255 and 256.

Species tree reconstruction
Two methodologies were used to reconstruct the species trees encompassing the 18 species used in the phylome reconstruction. First, a gene concatenation analysis was done based on 49 proteins encoded in genes that were found in single copy in at least 15 out of the 18 species. The concatenated alignment contained 25,023 amino acids. A ML tree was reconstructed using PhyML [S44] with LG model [S47] selected, four rate categories were used and invariant positions were inferred from the data. The second species tree was reconstructed using a super-tree approach. Duptree [S48] was used on the 13,734 trees reconstructed in the *R. delemar* phylome. This super-tree approach aims to find the species tree that minimizes the number of duplications inferred when reconciling genes trees with the species tree. The topologies obtained by both methods were identical and only the ML phylogeny is shown in Fig. 1C.

Mapping duplication rates
Single gene trees were scanned using ETE v2 [S49]. For each tree, duplications were found using a species-overlap algorithm. The duplications were mapped onto the species trees assuming that they occurred at the common ancestor of all the species involved in the duplication node. The number of duplications was then divided by the number of trees that could potentially contain a duplication at that given point. Trees that contained species-specific expansions of more than five members for the seed species were omitted.

Identification of G-protein coupled receptors
A specialized hidden Markov model specific to identify transmembrane (TM) regions of GPCRs, GPCRHMM [S50], was applied. This technique predicts if a given sequence is or is not a GPCR based on different TM topology features, such as loop length and amino acid composition. In a sensitivity-selectivity test, the sensitivity of GPCRHMM was reported to be about 15% higher than that of any other predictor tested, at comparable false positive rates.

To reduce the chances of wrongly classifying proteins as GPCRs, the PHOBIUS method, another TM predictor [S51], was applied to estimate independently the number of TM regions for each putative GPCR. Only in cases where PHOBIUS also predicted seven TM regions was the entry retained. This procedure resulted in the removal of about half the sequences from the initial set of candidates.

The resulting list was classified with three different tools. The first classification used BLASTclust with the following options: 50% coverage, identity cutoff 30% (http://toolkit.tuebingen.mpg.de/). In parallel, the list was searched with a GPCR classification tool [S52], and an alignment and a phylogenetic tree were also constructed directly from the list of candidate sequences. The GPCR-2L procedure may have been too stringent for fungal sequences, excluding, for example, Class X defined in *Trichoderma* spp. Furthermore, a class of GPCR-like proteins with similarity to the PQ-loop-domain containing protein Stm1 of *Schizosaccharomyces pombe* was not detected by GPCRHMM. Stm1 is a predicted 7TM protein by both PHOBIUS and GPCRHMM, however, it fails to meet the criteria for GPCR prediction by GPCRHMM. There is functional evidence that Stm1 is indeed a GPCR, based on a screening assay in yeast [S53, 54]. The homologs of Stm1 were identified by BLASTp in the Mucoromycotina databases. The list in Data S1 combines the results of these analyses, annotated to include the smaller subset identified by GPCR-2L.

Identification of protein kinases
In order to conduct as unbiased a search as possible, pre-calculated HMM profiles (http://www.compbio.dundee.ac.uk/kinomer) [S55] were used to identify and classify all kinases in the predicted proteomes belonging to one of 10 families. A protein was considered to belong to a family for which its raw bit score was the best (often there is multiple family classification) and it is at least ≥ 20.

Identification of cell wall biosynthesis proteins
The study was performed manually using a number of programs. PSORTII (http://www.genscript.com/psort/psort2.html) was used for the prediction of proteins with extracellular location and the presence of a signal peptide. Protein location was further confirmed with an extension of PSORTII (http://www.genscript.com/psort/wolf_psort.html). The presence of a signal peptide was confirmed by SignalP (http://www.cbs.dtu.dk/services/SignalP/). Proteins containing a putative extracellular location and a signal peptide were further analyzed for the presence of serine/threonine residues, and putative glycosylation sites using PROSITE (http://us.expasy.org/tools/scanprosite/). The presence of a GPI motif was analyzed with the big-PI Fungal Predictor algorithm (http://mendel.imp.univie.ac.at/gpi/fungi_server.html). The presence of amino

acid repeats in the proteins was analyzed with the program RADAR (http://www.ebi.ac.uk/Radar/). BLAST analyses were run at the SIB Blast Network service (http://us.expasy.org/tools/blast/) and at http://www.ebi.ac.uk/blastall/. Identification of important motifs in the analyzed proteins was carried out using the program SAPS (http://www.ebi.ac.uk/Tools/seqstats/saps/).

Identification of transcription factors

The genome-wide assignments of TFs were made by scanning the InterProScan [S56] predictions for all known DNA-binding domains (DBDs) assigned to transcriptional regulation. The library of the corresponding DBDs was collected from different sources including Pfam [S57] and DBD [S58] databases and literature. The InterProScan predictions were downloaded from the JGI database (http://jgi.doe.gov). The *P. blakesleeanus* and *M. circinelloides* genomes contain 879 and 650 genes for transcription factors (TFs), respectively (about 5% of the protein-coding genes), with an abundance of C_2H_2 Zn finger TFs. The 879 TFs in the *P. blakesleeanus* genome are assigned to 49 families of DNA-binding domains as based on InterProScan predictions.

Identification of TRAFAC class GTPases and regulators

The sequences were identified by BLASTp and PSI-BLAST searches of the database of predicted *P. blakesleeanus* proteins, *R. delemar* proteins and the nr database at NCBI (the remaining fungi). For *P. blakesleeanus* and *R. delemar*, the genome was further checked by tBLASTn to identify possible homologs that had been annotated incorrectly or not at all. For the remaining fungal genomes analysed, tBLASTn searches were conducted only in cases of suspicious absence of a homolog in the protein database (leading in several cases to identification of an un-annotated gene), hence it is possible that some additional un-annotated genes have been missed. Assignment of mutual orthologs is based mainly on reciprocal BLAST comparisons and in a few cases should be verified by phylogenetic analysis (accession numbers of individual GTPases from Dikarya are available upon request). Unless stated otherwise, the names assigned to groups of orthologs follow the nomenclature of the corresponding *S. cerevisiae* genes. Probable or known function of individual GTPases is assigned by consulting the literature dealing with characterisation of orthologs in fungal or other species. Fungal sequences falling into various categories of known regulators of the Ras GTPase superfamily were identified and analysed by an approach similar to that used for the analysis of fungal GTPases. In addition, for identification of proteins containing poorly conserved regulatory domains, we also used HMMER searches with profile HMMs built from multiple alignments retrieved from Pfam (http://www.sanger.ac.uk/Software/Pfam/) or SMART (http://smart.embl-heidelberg.de/) collections. The number of TRAFAC class GTPases identified in the *P. blakesleeanus* genome (148) is lower than the number for the *R. delemar* genome (192) but higher than the number for any Dikarya genome analysed, 70-133. The total number of Ras superfamily proteins encoded by the *P. blakesleeanus* and *R. delemar* genomes (213 and 255, respectively) is higher than in the Dikarya (from 77 to 100). The higher number of GTPases and Ras superfamily proteins in the Mucoromycotina fungi compared to Dikarya is due to both retention of a higher number of ancestral genes and more extensive lineage-specific gene duplications.

Supplemental references

S1. Idnurm, A., Rodríguez-Romero, J., Corrochano, L.M., Sanz, C., Iturriaga, E.A., Eslava, A.P., and Heitman, J. (2006). The *Phycomyces madA* gene encodes a blue-light photoreceptor for phototropism and other light responses. Proc Natl Acad Sci USA *103*, 4546-4551.

S2. Silva, F., Torres-Martínez, S., and Garre, V. (2006). Distinct *white collar-1* genes control specific light responses in *Mucor circinelloides*. Mol Microbiol *61*, 1023-1037.

S3. Chaudhary, S., Polaino, S., Shakya, V.P., and Idnurm, A. (2013). A new genetic linkage map of the zygomycete fungus *Phycomyces blakesleeanus*. PLoS One *8*, e58931.

S4. Alvarez, M.I., Eslava, A.P., and Lipson, E.D. (1989). Phototropism mutants of *Phycomyces blakesleeanus* isolated at low light intensity. Exp Mycol *13*, 38-48.

S5. Weinkove, D., Poyatos, J.A., Greiner, H., Oltra, E., Avalos, J., Fukshansky, L., Barrero, A.F., and Cerdá-Olmedo, E. (1998). Mutants of *Phycomyces* with decreased gallic acid content. Fungal Genet Biol *25*, 196-203.

S6. Sambrook, J., and Russell, D.W. (2001). Molecular Cloning. A Laboratory Manual, (Cold Spring Harbor: Cold Spring Harbor Laboratory Press).

S7. Jaffe, D.B., Butler, J., Gnerre, S., Mauceli, E., Lindblad-Toh, K., Mesirov, J.P., Zody, M.C., and Lander, E.S. (2003). Whole-genome sequence assembly for mammalian genomes: Arachne 2. Genome Res *13*, 91-96.

S8. Detter, J.C., Jett, J.M., Lucas, S.M., Dalin, E., Arellano, A.R., Wang, M., Nelson, J.R., Chapman, J., Lou, Y., Rokhsar, D., et al. (2002). Isothermal strand-displacement amplification applications for high-throughput genomics. Genomics *80*, 691-698.
S9. Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Res *9*, 868-877.
S10. Salamov, A.A., and Solovyev, V.V. (2000). Ab initio gene finding in *Drosophila* genomic DNA. Genome Res *10*, 516-522.
S11. Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. Genome Res *14*, 988-995.
S12. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y.O., and Borodovsky, M. (2008). Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res *18*, 1979-1990.
S13. Zdobnov, E.M., and Apweiler, R. (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics *17*, 847-848.
S14. Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. (2005). The Universal Protein Resource (UniProt). Nucleic Acids Res *33*, D154-159.
S15. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res *27*, 29-34.
S16. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., et al. (2004). The Pfam protein families database. Nucleic Acids Res *32*, D138-141.
S17. Bairoch, A. (2000). The ENZYME database in 2000. Nucleic Acids Res *28*, 304-305.
S18. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29.
S19. Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., et al. (2004). A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol *5*, R7.
S20. Nielsen, H., Brunak, S., and von Heijne, G. (1999). Machine learning approaches for the prediction of signal peptides and other protein sorting signals. Protein Eng *12*, 3-9.
S21. Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol *305*, 567-580.
S22. Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res *18*, 1851-1858.
S23. Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods *6*, 677-681.
S24. Quesneville, H., Bergman, C.M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., and Anxolabehere, D. (2005). Combined evidence annotation of transposable elements in genome sequences. PLoS Comput Biol *1*, 166-175.
S25. Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. (2011). Considering transposable element diversification in de novo annotation approaches. PLoS One *6*, e16526.
S26. Jurka, J. (2000). Repbase update: a database and an electronic journal of repetitive elements. Trends Genet *16*, 418-420.
S27. Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T., and Ussery, D.W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res *35*, 3100-3108.
S28. Vega-Arreguín, J.C., Ibarra-Laclette, E., Jiménez-Moraila, B., Martínez, O., Vielle-Calzada, J.P., Herrera-Estrella, L., and Herrera-Estrella, A. (2009). Deep sampling of the *Palomero* maize transcriptome by a high throughput strategy of pyrosequencing. BMC Genomics *10*, 299.
S29. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.
S30. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.
S31. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc Ser B *57*, 289-300.
S32. Götz, S., García-Gómez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talón, M., Dopazo, J., and Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res *36*, 3420-3435.

S33. Myhre, S., Tveit, H., Mollestad, T., and Laegreid, A. (2006). Additional gene ontology structure for improved biological reasoning. Bioinformatics *22*, 2020-2027.
S34. Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res *30*, 1575-1584.
S35. McLysaght, A., Hokamp, K., and Wolfe, K.H. (2002). Extensive genomic duplication during early chordate evolution. Nat Genet *31*, 200-204.
S36. Gordon, J.L., Byrne, K.P., and Wolfe, K.H. (2009). Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome. PLoS Genet *5*, e1000485.
S37. Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L.P., Denisov, I., Kormes, D., Marcet-Houben, M., and Gabaldón, T. (2011). PhylomeDB v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. Nucleic Acids Res *39*, D556-560.
S38. Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.
S39. Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res *33*, 511-518.
S40. Lassmann, T., and Sonnhammer, E.L. (2005). Kalign--an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics *6*, 298.
S41. Wallace, I.M., O'Sullivan, O., Higgins, D.G., and Notredame, C. (2006). M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res *34*, 1692-1699.
S42. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972-1973.
S43. Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol *14*, 685-695.
S44. Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol *59*, 307-321.
S45. Akaike, H. (1975). Information theory and an extension of the maximum likelihood principle. In 2nd International Symposium on Information Theory, B.N. Petrov and F. Csáki, eds. (Budapest, Hungary: Akadémia Kiado), pp. 267–281.
S46. Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L.P., Marcet-Houben, M., and Gabaldón, T. (2014). PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res *42*, D897-902.
S47. Le, S.Q., and Gascuel, O. (2008). An improved general amino acid replacement matrix. Mol Biol Evol *25*, 1307-1320.
S48. Wehe, A., Bansal, M.S., Burleigh, J.G., and Eulenstein, O. (2008). DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics *24*, 1540-1541.
S49. Huerta-Cepas, J., Dopazo, J., and Gabaldón, T. (2010). ETE: a python environment for tree exploration. BMC Bioinformatics *11*, 24.
S50. Wistrand, M., Kall, L., and Sonnhammer, E.L. (2006). A general model of G protein-coupled receptor sequences and its application to detect remote homologs. Protein Sci *15*, 509-521.
S51. Kall, L., Krogh, A., and Sonnhammer, E.L. (2004). A combined transmembrane topology and signal peptide prediction method. J Mol Biol *338*, 1027-1036.
S52. Xiao, X., Wang, P., and Chou, K.C. (2011). GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions. Mol Biosyst *7*, 911-919.
S53. Chung, K.S., Won, M., Lee, J.J., Ahn, J., Hoe, K.L., Kim, D.U., Song, K.B., and Yoo, H.S. (2007). Yeast-based screening to identify modulators of G-protein signaling using uncontrolled cell division cycle by overexpression of Stm1. J Biotechnol *129*, 547-554.
S54. Chung, K.S., Won, M., Lee, S.B., Jang, Y.J., Hoe, K.L., Kim, D.U., Lee, J.W., Kim, K.W., and Yoo, H.S. (2001). Isolation of a novel gene from *Schizosaccharomyces pombe*: *stm1*+ encoding a seven-transmembrane loop protein that may couple with the heterotrimeric Galpha 2 protein, Gpa2. J Biol Chem *276*, 40190-40201.
S55. Miranda-Saavedra, D., and Barton, G.J. (2007). Classification and functional annotation of eukaryotic protein kinases. Proteins *68*, 893-914.
S56. Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236-1240.

S57. Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2014). Pfam: the protein families database. Nucleic Acids Res *42*, D222-230.

S58. Wilson, D., Charoensawan, V., Kummerfeld, S.K., and Teichmann, S.A. (2008). DBD--taxonomically broad transcription factor predictions: new content and functionality. Nucleic Acids Res *36*, D88-92.