

Dandruff-associated *Malassezia* genomes reveal convergent and divergent virulence traits shared with plant and human fungal pathogens

Jun Xu*, Charles W. Saunders*, Ping Hu*, Raymond A. Grant*, Teun Boekhout^{†‡}, Eiko E. Kuramae^{†§}, James W. Kronstad[¶], Yvonne M. DeAngelis*, Nancy L. Reeder*, Kevin R. Johnstone*, Meredith Leland*, Angela M. Fieno*, William M. Begley*, Yiping Sun*, Martin P. Lacey*, Tanuja Chaudhary*, Thomas Keough*, Lien Chu^{||}, Russell Sears^{***††}, Bo Yuan^{****}, and Thomas L. Dawson, Jr.^{*§§}

*Procter & Gamble Co., Miami Valley Innovation Center, Cincinnati, OH 45253-8707; [†]Centraalbureau voor Schimmelcultures, Fungal Biodiversity Centre, 3584 CT, Utrecht, The Netherlands; [‡]Division of Acute Medicine and Infectious Diseases and Eijkman–Winkler Centre for Microbiology, Infectious Diseases and Inflammation, University Medical Centre, 3508 GA, Utrecht, The Netherlands; [§]Netherlands Institute of Ecology, Centre for Terrestrial Ecology, 6666 ZG, Heteren, The Netherlands; [¶]The Michael Smith Laboratories, University of British Columbia, 2185 East Mall, Vancouver, BC, Canada V6T 1Z4; ^{||}Integrated Genomics, 2201 West Campbell Park Drive, Suite 15, Chicago, IL 60612-3547; and ^{**}Department of Biomedical Informatics and Pharmacology, Ohio State University, Columbus, OH 43210

Edited by Joan Wennstrom Bennett, Rutgers, The State University of New Jersey, New Brunswick, NJ, and approved October 1, 2007 (received for review July 18, 2007)

Fungi in the genus *Malassezia* are ubiquitous skin residents of humans and other warm-blooded animals. *Malassezia* are involved in disorders including dandruff and seborrheic dermatitis, which together affect >50% of humans. Despite the importance of *Malassezia* in common skin diseases, remarkably little is known at the molecular level. We describe the genome, secretory proteome, and expression of selected genes of *Malassezia globosa*. Further, we report a comparative survey of the genome and secretory proteome of *Malassezia restricta*, a close relative implicated in similar skin disorders. Adaptation to the skin environment and associated pathogenicity may be due to unique metabolic limitations and capabilities. For example, the lipid dependence of *M. globosa* can be explained by the apparent absence of a fatty acid synthase gene. The inability to synthesize fatty acids may be complemented by the presence of multiple secreted lipases to aid in harvesting host lipids. In addition, an abundance of genes encoding secreted hydrolases (e.g., lipases, phospholipases, aspartyl proteases, and acid sphingomyelinases) was found in the *M. globosa* genome. In contrast, the phylogenetically closely related plant pathogen *Ustilago maydis* encodes a different arsenal of extracellular hydrolases with more copies of glycosyl hydrolase genes. *M. globosa* shares a similar arsenal of extracellular hydrolases with the phylogenetically distant human pathogen, *Candida albicans*, which occupies a similar niche, indicating the importance of host-specific adaptation. The *M. globosa* genome sequence also revealed the presence of mating-type genes, providing an indication that *Malassezia* may be capable of sex.

Despite their widespread occurrence and association with multiple common skin disorders, remarkably little is known about these fungi other than their unusual dependence on external lipids for *in vitro* growth (1). To understand the role of *Malassezia* in dandruff, we have focused on *Malassezia globosa* and *Malassezia restricta*, because they are the most commonly isolated species from human scalp (3, 11) and are associated with prevalent skin disorders, including dandruff (4).

We describe the genome sequence, secretory proteome, and *in vivo* expression of selected genes of *M. globosa* and a partial genome sequence and secretory proteome of *M. restricta*. This report highlights a conflict between phylogeny and host-specific adaptation, because genome- and proteome-based evidence indicates that the basidiomycete *Malassezia* species share similar sets of extracellular hydrolases with the phylogenetically distant *Candida albicans* that occupies an overlapping skin-related niche. In contrast, *Malassezia* is closely related to the plant pathogen *Ustilago maydis*, but these fungi are easily distinguished by differences in their sets of extracellular hydrolases and their host preferences.

Results and Discussion

Whole-genome shotgun sequencing (7× coverage) was performed on *M. globosa* type strain CBS 7966 [supporting information (SI) Table 2]. The assembled genome size of 9 Mb closely matches the value reported by Boekhout *et al.* (12). The genome size is among the smallest of the sequenced genomes of free-

fungal genomics | fungal proteomics | seborrheic dermatitis | skin | fungal mating

For >100 years, *Malassezia* have been associated with dandruff (1), an easily recognizable skin flaking condition occurring in 30–95% of humans (2). The presence of *Malassezia* is not sufficient to cause either dandruff or the more extreme skin condition, seborrheic dermatitis; in fact, many people harbor *Malassezia* without showing symptoms. However, *Malassezia* must have an essential role in these conditions, because scalp-flaking symptoms are improved by treatment with a variety of antifungal materials that remove *Malassezia* (1, 3–5). These fungi also cause systemic infections in neonates (6–8) and the skin disease pityriasis versicolor (1, 3, 5), as well as many common veterinary disorders, such as canine dermatitis and otitis externa (9). *Malassezia* are also thought to contribute to the common skin disease atopic eczema by host sensitization to fungal protein allergens (10).

Author contributions: J.X. and C.W.S. contributed equally to this work; C.W.S. and T.L.D. designed research; J.X., C.W.S., R.A.G., T.B., Y.M.D., N.L.R., K.R.J., M.L., A.M.F., W.M.B., Y.S., M.P.L., T.C., T.K., L.C., and T.L.D. performed research; J.X., C.W.S., P.H., E.E.K., R.S., and B.Y. analyzed data; and J.X., C.W.S., P.H., R.A.G., T.B., J.W.K., and T.L.D. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The sequences reported in this paper have been deposited in the DNA Data Bank of Japan/European Molecular Biology Laboratory/GenBank database [accession nos. AAYY00000000 (*M. globosa*) and AAXK00000000 (*M. restricta*)]. The versions described in this paper are the first versions, AAYY01000000 and AAXK01000000.

^{††}Present address: Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720.

^{††}Present address: Department of Computer Science, Institute of Biocomputing and Systems Biology, Shanghai Jiaotong University, Shanghai 200290, China.

^{§§}To whom correspondence should be addressed. E-mail: dawson.tl@pg.com.

This article contains supporting information online at www.pnas.org/cgi/content/full/0706756104/DC1.

© 2007 by The National Academy of Sciences of the USA

Table 1. Comparison of enzyme classes among several fungal species

Enzyme	Gene copies (predicted secreted proteins)				
	<i>M. globosa</i>	<i>U. maydis</i>	<i>C. albicans</i>	<i>C. neoformans</i>	<i>S. cerevisiae</i>
Lipase	14(13)	7(2)	14(12)	3(1)	3(0)
Phospholipase C	6(6)	0(0)	0(0)	0(0)	0(0)
Phospholipase D	2(0)	3(0)	2(0)	4(0)	2(0)
Phospholipase B	1(0)	1(1)	7(5)	1(1)	4(4)
Acid sphingomylinase	4(4)	1(1)	4(3)	1(1)	0(0)
Aspartyl protease	18(15)	10(7)	15(15)	9(6)	8(7)
Cysteine protease	7(1)	9(0)	9(1)	9(1)	8(1)
Metalloprotease	31(3)	35(7)	42(3)	44(5)	38(5)
Serine protease	10(2)	17(9)	15(10)	18(10)	13(8)
Glycosyl hydrolase	15(3)	65(36)	45(27)	52(27)	38(24)
Cutinase	0(0)	4(4)	0(0)	0(0)	0(0)
Pectin lyase	0(0)	1(0)	0(0)	0(0)	0(0)
Pectin esterase	0(0)	1(1)	0(0)	0(0)	0(0)
Polyketide synthase	1(0)	3(0)	0(0)	0(0)	0(0)
Total gene number	4,285	6,902	7,677	6,810	6,604

The number of genes within the indicated haploid genomes that encode a selected set of enzymes are shown to illustrate differences in coding capacity. The entry is highlighted in pink for the species with the largest number of genes for each category. The number of proteins that are predicted to be secreted, on the basis of a predicted signal sequence, using SignalP3.0 (37), are listed in parentheses for each category. The phospholipase C entry refers to the secretory enzyme similar to *Pseudomonas aeruginosa* phospholipase C and not the more commonly discussed cytoplasmic phosphoinositol-specific phospholipase C. Detailed methods for the construction of the table are described in [SI Text](#). A comparison of *M. globosa* and *C. albicans* aspartyl proteases is depicted in [SI Table 13](#).

a *MAT* locus of ≈ 170 kb in *M. globosa*. There was no indication of alternate mating-type loci in the genome of the sequenced strain CBS 7966 of *M. globosa*, again suggesting that this strain is haploid, and that mating-type switching does not occur. Several fragments of putative mating-related genes were also found in *M. restricta*: *pral* (MRE_contig_0425), *bE1* (MRE_contig_2721), and *bW1* (MRE_contig_3923 and _1863).

The 170-kb *MAT* region of *M. globosa* does not contain the high density of repetitive elements or possess orthologs of the other genes found at the 527-kb *MAT* locus of *U. hordei* (20). Furthermore, a comparison of the 82 genes predicted for the *MAT* locus of *M. globosa* with the genes found at the *MAT α* locus of another basidiomycete fungus, *Cryptococcus neoformans*, revealed that these loci share orthologs of the pheromone receptor gene, as expected, and both contain an ortholog of the *CIDI*

gene thought to play a role in the cell cycle (21). The significance of the shared possession of this gene at the *MAT* locus is not clear but may have implications for the evolution of basidiomycete mating type loci. The *MAT* locus of *C. neoformans* is remarkable, because it encodes components of the pheromone response pathway (21); this is not the case for the *MAT* locus of *M. globosa*.

The presence of candidate genes for mating in the *Malassezia* genomes is interesting, because mating has never been observed in *Malassezia* species. The presence of these genes suggests that the fungus may be capable of sexual development and that the characterization of the *MAT* region in other strains may identify one or more additional mating specificities for use in functional mating tests. We used the genome sequence to investigate the possibility of sexual competency in more detail by searching for orthologs of genes known to function in pheromone signaling,

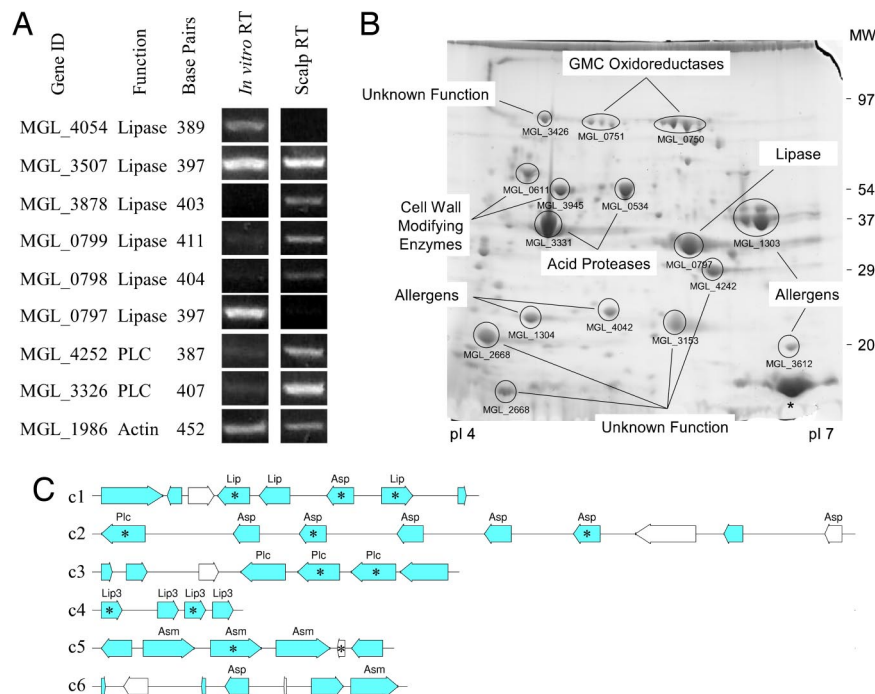


Fig. 2. Secreted proteins: gene expression and gene clusters. (A) Transcripts for lipase and phospholipase genes were detected from *M. globosa* cells isolated from scalp and from cells cultured *in vitro*. Isolated RNA was used for first-strand cDNA synthesis with reverse transcriptase and random hexamers and then PCR-amplified by using gene-specific primers, as described (17). (B) The 2D gel was stained with Coomassie blue. Multiple spots within a circle are different isoforms of the same protein. The dark spot in the bottom right corner of the gel (*) was resistant to multiple attempts at digestion and could not be identified. See also SI Tables 9 and 10. (C) Cyan symbols, proteins predicted to be secreted. Open symbols, proteins predicted not to be secreted. * in the middle of the gene indicates the protein was found in the extracellular fraction in the proteomics experiments. Lip, lipase family LIP; Lip3, lipase family LIP3; Plc, phospholipase C; Asp, aspartyl protease; Asm, acid sphingomyelinase. C1–6 refer to gene cluster names, with the contig numbers, gene numbers, and predicted gene functions listed in SI Table 12.

meiosis and sporulation in other fungi. This analysis revealed that the *M. globosa* genome encodes candidate orthologs for many of these functions. For example, orthologs were found for components of the *Saccharomyces cerevisiae* pheromone responsive MAPK module encoded by *STE11*, *STE7*, and *FUS3*, and for the *GPA1*- and *STE4*-encoded components of the heterotrimeric G protein complex that couples pheromone receptor signaling with the MAPK module (22). Overall, this analysis supports the idea that *M. globosa* is capable of sexual reproduction. The possibility of a sexual cycle in *Malassezia* yeasts is intriguing, because it may have a profound effect on the distribution of virulence-related genes among populations (23). Furthermore, the presence of a sexual state may open possibilities for classical genetic experimentation. This situation is reminiscent of the discovery of mating in *C. albicans* as a consequence of genomic evidence for mating type genes (24), and it is interesting that *C. albicans* mating is enhanced on skin (25).

There are important differences in gene content that may be linked to the adaptation of *M. globosa* and *U. maydis* to their specific host environments (Fig. 3C). Overall, *U. maydis* contains 2,536 genes with no apparent homologs in *M. globosa* (at $e < 10^{-5}$). The most striking difference is the presence of more carbohydrate-hydrolyzing proteins (glycosyl hydrolases) encoded by *U. maydis* than by *M. globosa* (Table 1). In particular, xylanase, cellulase, pectin esterase, polygalacturonase, pectin lyase, α -galactosidase, and α -amylase genes are present in *U. maydis* but not in *M. globosa*. *M. globosa* has only 325 genes that are missing in *U. maydis*, and most of these genes have unknown functions. The most prominent difference in genes with assignable function is that *M. globosa* encodes more lipases, aspartyl proteases, phospholipase C enzymes, and acid sphingomyelinases, as described above (Table 1). Many of these *M. globosa*

genes are clustered in the genome (Fig. 2C and SI Table 12). It is remarkable that clusters of genes for secreted proteins are also found in the *U. maydis* genome, although these proteins appear to function in plant-specific interactions, such as defense suppression and tumor induction (15). In general, the different sets of secreted proteins likely reflect the specificity in the host molecules targeted by the fungi.

The abundance of genes for phospholipases, lipases, aspartyl proteases, and acid sphingomyelinases in *M. globosa* is also seen in *C. albicans* (26) (Table 1), a distantly related (SI Fig. 4) human pathogenic and commensal fungus that also colonizes skin and mucosal surfaces (27). In contrast, these specific genes are not prevalent in the genome of the fermentative yeast *S. cerevisiae* (Table 1). The *C. albicans* genes encoding aspartyl proteases and phospholipase B are implicated in virulence in animal models (27, 28). The prevalence of similar gene families in two fungal species, *M. globosa* and *C. albicans*, that are opportunists on human skin may point to similar roles in colonization and virulence, although *C. albicans* does not provoke the specific skin disorders associated with *Malassezia* species.

Lipases, phospholipases, and allergens have been proposed to be important in *Malassezia* virulence, although definitive genetic tests have not been made, because gene disruption has not yet been demonstrated in *Malassezia* species. The presence of phospholipase activity correlates with the virulence of *Malassezia pachydermatis* in dogs (29). There is a model linking *Malassezia* lipase activity to the pathogenesis of dandruff and seborrheic dermatitis (17). *Malassezia* antigens may play a role in atopic eczema (1), and the *M. globosa* genome encodes proteins similar to 12 of 13 *Malassezia sympodialis* and *Malassezia furfur* proteins identified as reacting with IgE from atopic eczema patients (10), extending previous observations based on PCR analysis for a

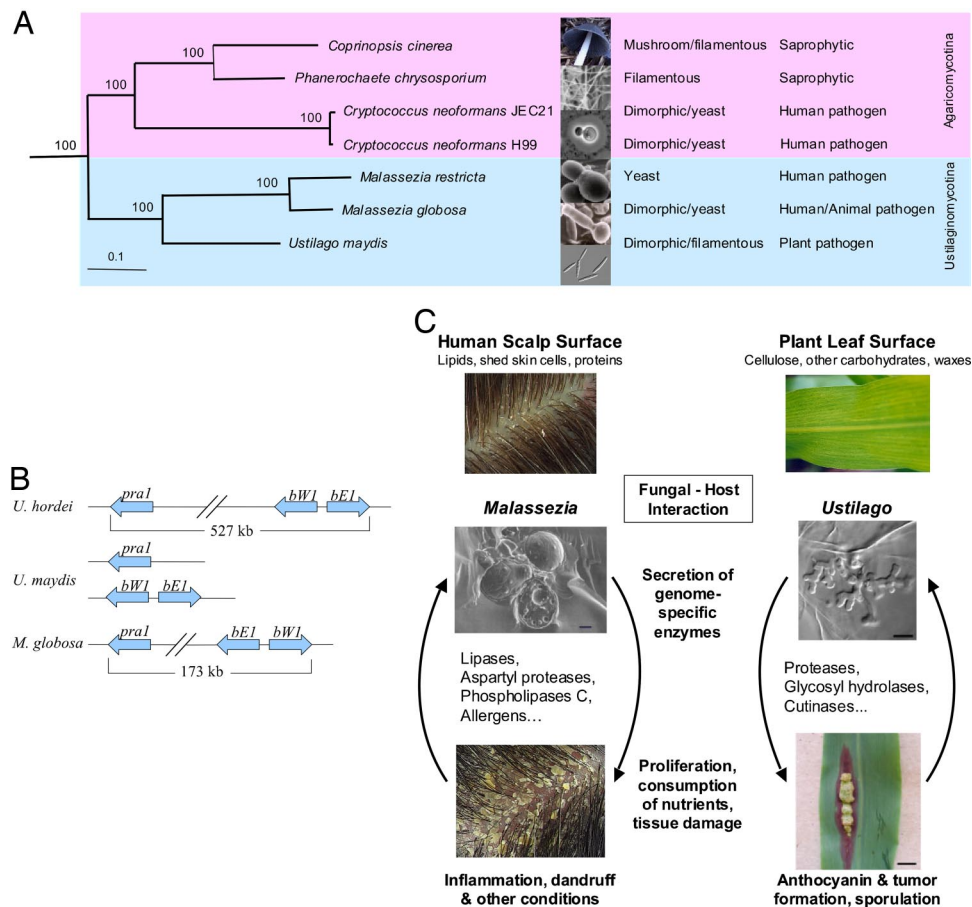


Fig. 3. Similarities of phylogeny, mating type gene organization, and pathogenesis between *Malassezia* and *Ustilago*. (A) The phylogenetic tree based on 28 concatenated single-copy orthologous proteins present in 34 fungal genomes was prepared by using Maximum Likelihood (PHYML) by using JTT amino acid model substitution (SI Table 11). This tree depicts a small number of fungi; a more inclusive tree is shown (SI Fig. 4). (B) Schematic representation of the mating loci of *U. hordei*, *U. maydis*, and *M. globosa*. *pra1*: pheromone receptor 1; *bW1*: homeodomain transcription factor b west 1; *bE1*: homeodomain transcription factor b east 1. *M. globosa* mating-related genes: putative pheromone, MGL0963; *pra1*, MGL0964; *bE1*, MGL0884; *bW1*, MGL0883. (C) Illustration of the adaptations by *Malassezia* and *Ustilago* to take advantage of distinct host environments for survival. The model emphasizes the most prominent protein families based on the available data. The downward-pointing arrows represent the damage to the host substrate inflicted by secreted proteins, and the upward-pointing arrows represent the damaged host substrate that provides resources to support the growth and persistence of the fungal population. [Scale bars: 2 μ m (*Malassezia*), 10 μ m (*Ustilago*), and 1 cm (tumor formation).]

subset of these allergen genes (30). Our discoveries of multicopy genes, *in vitro* protein production and secretion, and (in some cases) gene expression on the scalp provide support for the importance of several *M. globosa* gene families. These include the gene families whose importance has already been suggested, such as lipase, phospholipase C, and allergens, but also newly discovered families, such as aspartyl protease, glucose methanol choline oxidoreductase, and acid sphingomyelinase. These discoveries should facilitate the testing of the role of these protein families in skin diseases such as dandruff and seborrheic dermatitis.

Our report of *Malassezia* genome sequences is part of a growing understanding of the adaptations of human fungal pathogens. Because *Cryptococcus* and *Malassezia* are the only two basidiomycete fungi to infect humans commonly, there is interest in further comparison of their virulence traits. Of additional interest is the study of adaptation to animal skin that has apparently emerged independently in three lineages: the basidiomycete *Malassezia* spp. (Ustilaginomycotina), the ascomycete *C. albicans* (Saccharomycotina), and the true dermatophyte ascomycetes such as *Trichophyton* and *Microsporum* spp. (Pezizomycotina).

Summary

Overall, our genomic comparisons provide insights into the mechanisms by which fungi adapt to the mammalian skin environment. The phylogenetic position of *Malassezia* indicates they are closely related to plant pathogens, implying an ancestral shift from plant to animal host preference. This report describes the features of the *M. globosa* genome and some characteristics that may have facilitated this proposed host shift. The *M. globosa* genome is among the smallest of genomes of free-living fungi (SI Tables 2–4 and 14). One key insight is the apparent absence of a fatty acid synthase gene and compensation by the abundance of genes encoding hydrolases that could provide fatty acids. An additional feature of this adaptation to human skin is the presence of a set of secreted hydrolases similar to the set produced by the distantly related human pathogen *C. albicans*. Moreover, the secreted hydrolases are quite different between *M. globosa* and the phylogenetically related plant pathogen *U. maydis*.

Because *Malassezia*-associated skin diseases are relatively accessible for study, these diseases could become models to study the interaction of host and pathogen. With this report of the genome sequence of *M. globosa* and partial genome sequence of

M. restricta, coupled with the human genome sequence, this interaction can be studied at the molecular level, an approach formerly impossible because of the shortage of sequence information for *Malassezia* spp. For example, there are only four protein-coding genes or gene segments from each of *M. globosa* and *M. restricta* in the current National Center for Biotechnology Information nr database. A deeper understanding of these interactions may well lead to an understanding of the conditions that lead to skin disease and new treatment paradigms to alleviate the major impact that *Malassezia* species have on human and animal health.

Materials and Methods

Strains. *M. globosa* CBS 7966 and *M. restricta* CBS 7877 were obtained from the Centraalbureau voor Schimmelcultures Fungal Biodiversity Centre.

EST Sequencing, Genome Sequencing, and Annotation. An EST library was created from RNA isolated from *M. globosa* cells grown in mDixon medium (17). *M. globosa* DNA was sequenced by using a whole-genome shotgun approach (7 $\times$), and 2,000 finishing reactions were performed to fill gaps and fix low-quality regions. The genome was assembled by using Phrap. Gene models were constructed by combining sequencing information from 1,536 EST sequences and *in silico* predictions from TWIN-SCAN (31), GeneID (32), and GeneWise (33) (*SI Text*). Putative gene function assignments were obtained through homology searches against the National Center for Biotechnology Information nr, PFAM (34), and TIGRFAM (35) databases. *M. restricta* DNA was sequenced to 1 $\times$ coverage and assembled by using Phrap. Gene models were constructed by aligning predicted *M. globosa* proteins and proteins from the nr database to the *M. restricta* genome by using EXONERATE (36). More details are provided in *SI Text*.

Prediction of Secreted Proteins. SignalP 3.0 (37) was used to predict N-terminal secretion signals. The intersection (347 genes) be-

tween the hidden Markov and neural network prediction was used to indicate candidates for secreted proteins.

Isolation of Cell Surface-Associated Proteins. Cells grown in mDixon media (17) were pelleted by centrifugation at 4,000 $\times$ g and washed three times with TE buffer [50 mM Tris-HCl (pH 8.0)/5 mM EDTA containing protease inhibitors (Roche, Indianapolis, IN)]. The cell pellet was resuspended in TE buffer, and cells were washed further by pelleting through a cushion of 20% sucrose at 10,000 $\times$ g. The washed cell pellets were resuspended in a buffer containing 100 mM EDTA, 300 mM 2-mercaptoethanol (one part cells to two parts buffer) and then gently mixed for 1 h at room temperature. Cells were removed by centrifugation, and the extracted proteins in the supernatant were dialyzed [Snake-Skin dialysis tubing with 7000 molecular weight cutoff (Pierce, Rockford, IL)] for 24 h at 4°C against 0.1% acetic acid, lyophilized, and solubilized in isoelectric focusing sample buffer. Proteins were focused by using pH 4–7 strips (GE Healthcare, Piscataway, NJ) and separated on 10–20% gradient gels (Invitrogen, Carlsbad, CA). Details of the 2D gel electrophoresis procedure and the identification of proteins by mass spectrometry have been described (17). The amino acid sequences were used to identify the corresponding genes.

We thank Gil Cloyd, Melissa Monich, Bob LeBoeuf, Jeff Hamner, Eric Armstrong, Bob Lindenschmidt, Chris Kelling, Pete Ellingson, John McIver, Charlie Bascom, Kenny Morand, Mike Manhart, Cathy Oppenheimer, and Noelle Fitzgerald for support; Eveline Guého for photographs of *Malassezia* (Fig. 3A); Jana Klose for photographs of *Ustilago*; George Smulian, Fred Dietrich, Pamela Trotter, Aditya Gupta, Jacques Guillot, Carlos Echavarrri, and Jim Schwartz for technical discussions; Lisa Bosch and Diane Vonderheide for help with manuscript preparation; Vincent Robert for use of parts of the Biolumics software; Integrated Genomics for *M. globosa* sequencing reactions; and Agencourt (Beverly, MA) for sequencing the *M. restricta* genomic DNA and the EST library. Eiko Kuramae was supported by a KNAW renewal fund grant.

- Chen TA, Hill PB (2005) *Vet Dermatol* 16:4–26.
- Schwartz JR, Cardin CW, Dawson TL (2004) in *Textbook of Cosmetic Dermatology*, eds Baran R, Maibach HI (Martin Dunitz, Ltd, London), pp 259–272.
- Ashbee HR, Evans EG (2002) *Clin Microbiol Rev* 15:21–57.
- Gupta AK, Batra R, Bluhm R, Boekhout T, Dawson TL, Jr (2004) *J Am Acad Dermatol* 51:785–798.
- Midgley G (2000) *Med Mycol* 38 Suppl 1:9–16.
- Batra R, Boekhout T, Gueho E, Cabanes FJ, Dawson TL, Jr, Gupta AK (2005) *FEMS Yeast Res* 5:1101–1113.
- Chang HJ, Miller HL, Watkins N, Arduino MJ, Ashford DA, Midgley G, Agüero SM, Pinto-Powell R, von Reyn CF, Edwards W, et al. (1998) *N Engl J Med* 338:706–711.
- Crespo Erchiga V, Guého E (2005) in *Topley and Wilson's Microbiology and Microbial Infections, Medical Mycology*, eds Merz WG, Hay RJ (Hodder Arnold, London), pp 202–243.
- Outerbridge CA (2006) *Clin Tech Small Anim Pract* 21:128–134.
- Schmid-Grendelmeier P, Scheynius A, Cramer R (2006) *Chem Immunol Allergy* 91:98–109.
- Gemmer CM, DeAngelis YM, Theelen B, Boekhout T, Dawson TL, Jr (2002) *J Clin Microbiol* 40:3350–3357.
- Boekhout T, Kamp M, Gueho E (1998) *Med Mycol* 36:365–372.
- Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, Mohr C, Pohlmann R, Luedi P, Choi S, et al. (2004) *Science* 304:304–307.
- Hermida L, Brachat S, Voegeli S, Philippsen P, Primig M (2005) *Nucleic Acids Res* 33:D348–D352.
- Kämper J, Kahmann R, Bolker M, Ma LJ, Brefort T, Saville BJ, Banuett F, Kronstad JW, Gold SE, Muller O, et al. (2006) *Nature* 444:97–101.
- Gaitatzis N, Kunze B, Muller R (2001) *Proc Natl Acad Sci USA* 98:11136–11141.
- DeAngelis YM, Saunders CW, Johnstone KR, Reeder NL, Coleman CG, Kaczvinsky JR, Jr, Gale C, Walter R, Mekel M, Lacey MP, et al. (2007) *J Invest Dermatol* 127:2138–2146.
- Begerow D, Stoll M, Bauer R (2006) *Mycologia* 98:906–916.
- Kronstad JW, Staben C (1997) *Annu Rev Genet* 31:245–276.
- Bakkeren G, Jiang G, Warren RL, Butterfield Y, Shin H, Chiu R, Linning R, Schein J, Lee N, Hu G, et al. (2006) *Fungal Genet Biol* 43:655–666.
- Lengeler KB, Fox DS, Fraser JA, Allen A, Forrester K, Dietrich FS, Heitman J (2002) *Eukaryot Cell* 1:704–718.
- Bardwell L (2005) *Peptides* 26:339–350.
- Heitman J (2006) *Curr Biol* 16:R711–R725.
- Hull CM, Johnson AD (1999) *Science* 285:1271–1275.
- Lachke SA, Lockhart SR, Daniels KJ, Soll DR (2003) *Infect Immun* 71:4970–4976.
- Braun BR, van Het Hoog M, d'Enfert C, Martchenko M, Dungan J, Kuo A, Inglis DO, Uhl MA, Hogues H, Berriman M, et al. (2005) *PLoS Genet* 1:36–57.
- Calderone RA, Fonzi WA (2001) *Trends Microbiol* 9:327–335.
- Naglik JR, Challacombe SJ, Hube B (2003) *Microbiol Mol Biol Rev* 67:400–428, table of contents.
- Cafarchia C, Otranto D (2004) *J Clin Microbiol* 42:4868–4869.
- Andersson A, Scheynius A, Rasool O (2003) *Med Mycol* 41:479–485.
- Korf I, Flicek P, Duan D, Brent MR (2001) *Bioinformatics* 17(Suppl 1):S140–S148.
- Guigo R (1998) *J Comput Biol* 5:681–702.
- Birney E, Clamp M, Durbin R (2004) *Genome Res* 14:988–995.
- Eddy SR (1998) *Bioinformatics* 14:755–763.
- Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, White O (2001) *Nucleic Acids Res* 29:41–43.
- Slater GS, Birney E (2005) *BMC Bioinformatics* 6:31.
- Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) *J Mol Biol* 340:783–795.

Section II. Supplementary Tables

Table 2. *Malassezia* genome features

<i>M. globosa</i> Nuclear Genome Information		
	Size (bp)	8.9M
	G+C Content (%)	52
	Protein Coding Gene Number	4,285
	Percent Coding	69
	Single Exon Gene Number	3,113
	Percent Genes with Introns	27
	Average Gene Length (bp)	1,484
	Average Coding Length (bp)	1,447
	Proteins mapped to gene ontology	2,943
	Proteins with match in NCBI nr database	4,066
	Proteins best match <i>U. maydis</i> sequence	3,354 (82%)
	tRNA Genes	82
Exon		
	Average Exon Length (bp)	975
	Total Exon Number	6,377
	Exons per Gene	1.5
	Exons per Multi-Exon Gene	2.8
	G+C Content (%)	53
Intron		
	Average Intron length (bp)	76
	Total Intron Number	2,092
	Introns per Gene	0.5
	Introns per Multi-Exon Gene	1.8
	G+C Content (%)	47
Intergenic Region		
	Average intergenic length (bp)	593
	G+C Content (%)	49
Assembly statistics		
	Number of reads	142,966
	Number of reads incorporated into assembly	127,504
	Coverage	~7
	Number of contigs	67
	Number of scaffolds	54
	N50 contig (kb)	655
	N50 scaffold (kb)	660
<i>M. restricta</i> Genome Information		
	Number of reads	17,914
	Number of contigs	3,812
	Number of nucleotides in assembly	4.6M
	N50 contig (kb)	1.4
	Coverage	~1
	G+C Content (%)	55
	Gene Fragment Number	7,057
	Gene Fragments w/o homology to <i>M. globosa</i> genes	4

SI Table 3. Genome sizes reported by the fungal genome sequencing projects in NCBI

Genome	Subgroup	Genome Size (Mb)
<i>Encephalitozoon cuniculi</i> GB-M1	Microsporidia	2.5
<i>Antonospora locustae</i>	Microsporidia	2.9
<i>Pneumocystis carinii</i>	Ascomycota	8
<i>Eremothecium gossypii</i> ATCC 10895	Ascomycota	8.74
<i>Malassezia globosa</i> CBS 7960	Basidiomycota	8.9
<i>Pichia angusta</i> RB11	Ascomycota	9.5
<i>Kluyveromyces thermotolerans</i> CBS 6340	Ascomycota	10.6
<i>Kluyveromyces lactis</i> NRRL Y-1140	Ascomycota	10.69
<i>Kluyveromyces waltii</i> NCYC 2644	Ascomycota	10.91
<i>Saccharomyces bayanus</i> 623-6C	Ascomycota	11.54
<i>Saccharomyces bayanus</i> MCYC 623	Ascomycota	11.54
<i>Saccharomyces paradoxus</i> NRRL Y-17217	Ascomycota	11.75
<i>Pichia guilliermondii</i> ATCC 6260	Ascomycota	12
<i>Saccharomyces cerevisiae</i> RM11-1a	Ascomycota	12
<i>Saccharomyces cerevisiae</i> YJM789	Ascomycota	12
<i>Saccharomyces cerevisiae</i> S288c	Ascomycota	12.07
<i>Saccharomyces mikatae</i> IFO 1815	Ascomycota	12.12
<i>Debaryomyces hansenii</i> CBS767	Ascomycota	12.22
<i>Candida glabrata</i> CBS 138	Ascomycota	12.28
<i>Saccharomyces servazzii</i> CBS 4311	Ascomycota	12.3
<i>Schizosaccharomyces pombe</i> 972h-	Ascomycota	12.5
<i>Saccharomyces kluyveri</i> NRRL Y-12651	Ascomycota	12.6
<i>Zygosaccharomyces rouxii</i> CBS 732	Ascomycota	12.8
<i>Pichia farinosa</i> CBS 7064	Ascomycota	13.9
<i>Candida albicans</i> WO-1	Ascomycota	14
<i>Kluyveromyces marxianus</i> CBS 712	Ascomycota	14
<i>Schizosaccharomyces japonicus</i> yFS275	Ascomycota	14
<i>Schizosaccharomyces octosporus</i> Beij. 1897	Ascomycota	14
<i>Candida tropicalis</i> CBS 94	Ascomycota	15
<i>Candida tropicalis</i> MYA-3404	Ascomycota	15
<i>Pichia stipitis</i> CBS 6054	Ascomycota	15.4
<i>Candida albicans</i> SC5314	Ascomycota	16
<i>Clavispora lusitaniae</i> ATCC 42720	Ascomycota	16
<i>Lodderomyces elongisporus</i> NRRL YB-4239	Ascomycota	16
<i>Kazachstania exigua</i> CBS 379	Ascomycota	18
<i>Cryptococcus neoformans</i> WM276	Basidiomycota	18.37
<i>Cryptococcus neoformans</i> var. <i>neoformans</i> B-3501A	Basidiomycota	18.52
<i>Cryptococcus neoformans</i> var. <i>neoformans</i> JEC21	Basidiomycota	19.05
<i>Cryptococcus neoformans</i> R265	Basidiomycota	20

<i>Cryptococcus neoformans</i> var. <i>grubii</i> H99	Basidiomycota	20
<i>Ustilago maydis</i> 521	Basidiomycota	20
<i>Yarrowia lipolytica</i> CLIB122	Ascomycota	20.5
<i>Batrachochytrium dendrobatidis</i> JEL423	Chytridomycota	23.7
<i>Ajellomyces capsulatus</i> G186AR	Ascomycota	24
<i>Ajellomyces capsulatus</i> G217B	Ascomycota	24
<i>Ascosphaera apis</i> USDA-ARSEF 7405	Ascomycota	24
<i>Ajellomyces capsulatus</i> NAm1 NAm I	Ascomycota	28
<i>Ajellomyces dermatitidis</i> ATCC 26199	Ascomycota	28
<i>Coccidioides immitis</i> RS	Ascomycota	28.78
<i>Coccidioides immitis</i> H538.4	Ascomycota	29
<i>Coccidioides immitis</i> RMSCC 2394	Ascomycota	29
<i>Coccidioides immitis</i> RMSCC 3703	Ascomycota	29
<i>Coccidioides posadasii</i> C735	Ascomycota	29
<i>Coccidioides posadasii</i> CPA 0001	Ascomycota	29
<i>Coccidioides posadasii</i> CPA 0020	Ascomycota	29
<i>Coccidioides posadasii</i> CPA 0066	Ascomycota	29
<i>Coccidioides posadasii</i> RMSCC 1037	Ascomycota	29
<i>Coccidioides posadasii</i> RMSCC 1038	Ascomycota	29
<i>Coccidioides posadasii</i> RMSCC 1040	Ascomycota	29
<i>Coccidioides posadasii</i> RMSCC 2133	Ascomycota	29
<i>Coccidioides posadasii</i> RMSCC 3488	Ascomycota	29
<i>Coccidioides posadasii</i> RMSCC 3700	Ascomycota	29
<i>Coccidioides posadasii</i> Silveria	Ascomycota	29
<i>Aspergillus fumigatus</i> Af293	Ascomycota	30
<i>Phanerochaete chrysosporium</i> RP-78	Basidiomycota	30
<i>Uncinocarpus reesii</i> 1704	Ascomycota	30
<i>Aspergillus nidulans</i> FGSC A4	Ascomycota	31
<i>Lentinula edodes</i> L-54	Basidiomycota	33
<i>Trichoderma reesei</i> QM9414	Ascomycota	33
<i>Aspergillus niger</i> CBS 513.88	Ascomycota	33.9
<i>Podospora anserina</i> S mat+	Ascomycota	34
<i>Aspergillus clavatus</i> NRRL 1	Ascomycota	35
<i>Aspergillus terreus</i> ATCC 20542	Ascomycota	35
<i>Aspergillus terreus</i> NIH2624	Ascomycota	35
<i>Neosartorya fischeri</i> NRRL 181	Ascomycota	35
<i>Aspergillus flavus</i> NRRL3357	Ascomycota	36
<i>Aspergillus oryzae</i> RIB40	Ascomycota	37.2
<i>Chaetomium globosum</i> CBS 148.51	Ascomycota	36
<i>Coprinopsis cinerea</i> (<i>Coprinus cinereus</i>) okayama7#130	Basidiomycota	37.5
<i>Botryotinia fuckeliana</i>	Ascomycota	38
<i>Botryotinia fuckeliana</i> B05.10	Ascomycota	38
<i>Sclerotinia sclerotiorum</i> 1980	Ascomycota	38
<i>Gibberella zeae</i> PH-1	Ascomycota	40

<i>Magnaporthe grisea</i> 70-15	Ascomycetes	40
<i>Magnaporthe grisea</i> 70-15	Ascomycetes	40
<i>Nectria haematococca</i> MPVI	Ascomycetes	40
<i>Rhizopus oryzae</i> RA 99-880	Zygomycota	40
<i>Neurospora crassa</i> OR74A	Ascomycetes	43
<i>Gibberella moniliformis</i> 7600	Ascomycetes	46
<i>Phakopsora pachyrhizi</i>	Basidiomycetes	50
<i>Puccinia graminis</i> f. sp. <i>tritici</i> CRL 75-36-700-3	Basidiomycetes	81.5

Note that the three smallest genomes are from organisms incapable of growth outside of host cells (34-36). Among Basidiomycota genome sequences in this table, the *M. globosa* genome is two-fold smaller than all others. There is a recent report (37) of another Basidiomycota genome of similar size to that of *M. globosa*, i.e., the *Sporobolomyces roseus* genome is 10 Mb. We have not included that genome in this table because the sequence is neither published nor deposited in GenBank, as of this writing.

SI Table 4. The genome size and protein-coding capacity of 19 sequenced fungal genomes

Genome	Approximate gene number	Genome size (Mb)	Reference
<i>Encephalitozoon cuniculi</i>	1997	2.9	Katinka, M. D., <i>et al.</i> (34)
<i>Malassezia globosa</i>	4288	8.9	This study
<i>Eremothecium gossypii</i>	4726	8.74	http://agd.vital-it.ch/Ashbya_gossypii/index.html
<i>Schizosaccharomyces pombe</i>	5001	12.5	http://www.sanger.ac.uk/Projects/S_pombe/
<i>Candida glabrata</i>	5215	12.28	http://cbi.labri.fr/Genolevures/elt/CAGL
<i>Kluyveromyces lactis</i>	5327	10.69	http://cbi.labri.fr/Genolevures/elt/KLLA
<i>Saccharomyces cerevisiae</i>	5869	12	http://www.yeastgenome.org/
<i>Debaryomyces hansenii</i>	6319	12.22	http://cbi.labri.fr/Genolevures/elt/DEHA
<i>Yarrowia lipolytica</i>	6436	20.5	http://cbi.labri.fr/Genolevures/elt/YALI
<i>Candida albicans</i>	6679	16	http://www.candidagenome.org/
<i>Ustilago maydis</i>	6904	20	http://mips.gsf.de/genre/proj/ustilago/
<i>Cryptococcus neoformans</i>	7302	14	http://www.broad.mit.edu/annotation/genome/cryptococcus_neoformans/Home.html
<i>Neurospora crassa</i>	9356	43	http://mips.gsf.de/genre/proj/ncrassa/
<i>Aspergillus terreus</i>	10406	35	http://www.broad.mit.edu/annotation/genome/aspergillus_terreus/Home.html
<i>Aspergillus nidulans</i>	10701	31	http://www.broad.mit.edu/annotation/genome/aspergillus_nidulans/Home.html
<i>Aspergillus oryzae</i>	12074	37.2	http://www.bio.nite.go.jp/dogan/MicroTop?GENOME_ID=ao
<i>Magnaporthe grisea</i>	12841	40	http://www.broad.mit.edu/annotation/genome/magnaporthe_grisea/Home.html
<i>Coprinopsis cinerea</i>	13544	37.5	http://www.broad.mit.edu/annotation/genome/coprinus_cinereus/Home.html
<i>Fusarium graminearum</i>	13969	40	http://mips.gsf.de/genre/proj/fusarium/

Most of the gene numbers are derived by downloading proteome sequences from the reference web sites and counting the number of proteins. The data for *E. gossypii* were directly obtained through the web page listed in the reference URL because the proteome download link was not working at the time this paper was written. The *Encephalitozoon cuniculi* gene count was obtained from the published report (34) because there is no direct link to download the proteome.

SI Text

Section I

***M. globosa* DNA sequencing and assembly.** *M. globosa* CBS 7966 was sequenced using a whole genome shotgun approach (WGS) approach. Genomic DNA was prepared from laboratory-cultured *M. globosa* cells. A plasmid library, with inserts around 2 kb, and a cosmid library, with inserts about 40 kb, were constructed. Pair reads were obtained and assembled using Phrap (1). The contigs were edited using Consed (2), and 2,000 finishing reactions were performed based on the recommendations of the autofinish function of Consed (3). The contigs were linked into super contigs based on pair-reads information. The detailed assembly stats are listed in SI Table 2.

The *M. globosa* genome was sequenced to about 7X coverage. Assuming that the sequencing was random, this should cover >99% of the genome. The genome size of 8.9 Mb, derived from assembled sequences, correlates well with a prior estimate of 8.8-8.9 Mb (4). This indicates the sequencing is reasonably complete. Of the 15,462 unplaced reads, 11,978 appear to have low sequence quality (PHRED20 bases < 200). The remaining 3,484 reads are either still of not sufficient quality to be incorporated into the assembly or there is strong evidence indicating these reads are contaminating DNA from other species, based on blastn searching against the nr database. *M. globosa* has an estimated 8 chromosomes with sizes of 660, 830, 900, 970, 1,080, 1,190, 1,360, and 1,870 kb (5). The assembly has 15 scaffolds larger than 100 kb (comprising about 8.75 Mb) and 20 larger than 10 kb (comprising about 8.9 Mb). Smaller contigs were also included in the assembly, although in some cases we have no evidence whether they belong to the genome. Since the scaffold number is larger than the estimated

chromosome number and only 6 of the scaffold ends have putative telomere sequences, this indicates there are still clone gaps and sequencing gaps in our assembly. We estimate there are around 12 clone gaps (1.5/chromosome) to reduce the top 20 scaffolds to 8 chromosomes, and at least 20 sequence gaps exist (12 clone gaps + 8 existing sequence gaps, or 2.5/chromosome). The largest scaffold has a size of 1,772 kb, which correlates well with the largest estimated chromosome size of 1,870 kb.

Listed below are all the scaffolds that are longer than 10 kb as well as their member contigs and predicted gene count for each scaffold and contig.

Scaffold ID	Contig ID	Total bases	Total # predicted genes
mgl_sf_1		1772005	1,171
	mgl_sf_1.1	1406286	884
	mgl_sf_1.2	365719	287
mgl_sf_2		1012602	760
	mgl_sf_2.1	1012602	760
mgl_sf_3		867662	632
	mgl_sf_3.1	866456	631
	mgl_sf_3.2	1206	1
mgl_sf_4		765572	640
	mgl_sf_4.1	765572	640
mgl_sf_5		659885	521
	mgl_sf_5.1	139126	112
	mgl_sf_5.2	520759	409
mgl_sf_6		654612	465
	mgl_sf_6.1	654612	465
mgl_sf_7		632991	457
	mgl_sf_7.1	632991	457
mgl_sf_8		545971	381
	mgl_sf_8.1	537776	380
	mgl_sf_8.2	8195	1
mgl_sf_9		457438	332
	mgl_sf_9.1	457438	332
mgl_sf_10		452823	305
	mgl_sf_10.1	452823	305
mgl_sf_12		243067	162
	mgl_sf_12.1	243067	162
mgl_sf_11		194297	130
	mgl_sf_11.1	194297	130
mgl_sf_13		190001	119

	mgl_sf_13.1	189045	118
	mgl_sf_13.2	956	1
mgl_sf_14		177679	121
	mgl_sf_14.1	177679	121
mgl_sf_15		124299	81
	mgl_sf_15.1	124299	81
mgl_sf_16		18655	10
	mgl_sf_16.1	16903	8
	mgl_sf_16.2	1752	2
mgl_sf_17		24914	11
	mgl_sf_17.1	24914	11
mgl_sf_18		31127	21
	mgl_sf_18.1	31127	21
mgl_sf_19		48901	26
	mgl_sf_19.1	48901	26
mgl_sf_1001		12582	6
	mgl_sf_1001.1	688	1
	mgl_sf_1001.2	2931	1
	mgl_sf_1001.4	8963	4

***M. restricta* DNA sequencing and assembly.** Genomic DNA was prepared from laboratory-cultured *M. restricta* CBS 7877 cells, and a WGS library containing inserts around 2 kb was constructed. Pair reads were obtained and assembled using Phrap. The detailed assembly stats are listed in SI Table 2.

EST sequencing. An EST library was constructed from cultured *M. globosa* cells. 1,536 sequences were obtained and assembled into 950 contigs using Phrap.

***M. globosa* genome annotation.** The assembled EST sequences were aligned back to the genome using EST_GENOME (6). A set of training genes was manually curated based on the alignments. The NCBI nr database was searched against the assembled contigs using tblastn (7). Protein sequences with bit score (>200) and with protein identity > 60% were used to align to the genome using GeneWise (8). The alignments were then manually curated to create additional training sequences. Based on the curated training sequences, gene finding models were created for both GeneId (9) and TwinScan

(10). Gene models from the two predictions were combined together. Discrepancies between the predictions were resolved by choosing the prediction with the better homology score with known proteins and by manual inspection. Putative function was assigned based on homology searches against PFAM (11) and TIGRFAM (12) databases using HMMER (13), COG and KOG (14) databases using XUGNITOR (15), and the NCBI nr database using blastp. Transfer RNAs were identified using tRNAScan-SE (16). Manual curation was performed using ARTEMIS (17).

***M. restricta* genome annotation.** *M. restricta* DNA was sequenced to 1x coverage and assembled using Phrap. Gene models were constructed by aligning predicted *M. globosa* proteins and proteins from the NCBI nr database to the *M. restricta* genome using EXONERATE (18). The genome regions longer than 300 bases that were not mapped to *M. globosa* were searched against the NCBI nr database using blastx. Additional gene models were created using EXONERATE and protein sequences from the NCBI nr database that best matched to each region and showed a bit score greater than 100.

Metabolic pathways. Metabolic capabilities were investigated by searching the *M. globosa* proteome against the KEGG (19) pathway database. The hits against proteins annotated to a particular enzyme activity were overlaid on the KEGG pathway map using internally developed software. Each pathway was then manually inspected.

Comparison to known ribosomal DNA sequences. As a test of the identity of the strains whose genomes we sequenced, we compared the genome sequences to the ribosomal DNA sequences previously deposited in GenBank. The ribosomal DNA region of the *M. globosa* 7966 genome shows 100% sequence identity with the ribosomal DNA sequences generated from the same strain and already deposited in GenBank

(AY743630, AY743604, AY387228). The *M. restricta* 7877 genome ribosomal DNA sequence also matches previously deposited ribosomal DNA sequences (AY743607, AY387239) from that strain.

Synteny Identification. The *M. globosa* proteome was searched against the *U. maydis* proteome using blastp with an e-value cutoff of 10^{-5} . Adjacent reciprocal best matched genes in *M. globosa* that are also adjacent in *U. maydis* and have the same relative orientation were identified. In addition, DAGchainer (20) was used to identify syntenic arranged regions between *M. globosa* and *U. maydis* using parameters -A 3 -D 3000 to specify a minimum synteny group of three and a maximum gap of 3000 base pairs.

Comparison of enzyme classes among several fungal species. A list of PFAM profiles (see below) representing the activities of interest was searched against the proteome sequences of *M. globosa*, *U. maydis*, *C. albicans*, *C. neoformans* and *S. cerevisiae* using hmmpfam (13) trusted cutoff option (--cut_tc). For acid sphingomyelinase activity, we used acid sphingomyelinases from *C. albicans* to search against the genomes using blastp, and the results were manually inspected. Candidate PKS genes were identified by manual inspection of a list of proteins containing the PKS domain (CD00833) (21). Secretory phospholipases C (PLC) were identified by using *Pseudomonas aeruginosa* secretory phospholipase C (NP_252009.1) to search against the genomes using blastp and manual inspection. This class of PLCs was also found in several other fungi not listed in the table (22). The signal peptides were predicted using the intersection between the hidden Markov and neural network prediction from SignalP 3.0 (23). The sources of the genome sequence were listed in SI Table 4.

PFAM profiles used for classification.

PFAM name	Class name
-----------	------------

LIP	Lipase
Lipase	
Lipase3_N	
Lipase_2	
Lipase_3	
Lipase_chap	
Lipase_GDSL	
PLD_envelope	Phospholipase D
PLDc	
PLA2_B	Phospholipase B
Asp	Aspartyl protease
Peptidase_A17	
Peptidase_A21	
Peptidase_A22B	
Peptidase_A24	
Peptidase_A25	
Peptidase_A3	
Peptidase_A4	
Peptidase_A6	
Peptidase_A8	
Peptidase_C1	Cysteine protease
Peptidase_C10	
Peptidase_C11	
Peptidase_C12	
Peptidase_C13	
Peptidase_C14	
Peptidase_C15	
Peptidase_C16	
Peptidase_C1_2	
Peptidase_C2	
Peptidase_C21	
Peptidase_C23	
Peptidase_C24	
Peptidase_C25	
Peptidase_C25_C	
Peptidase_C26	
Peptidase_C27	
Peptidase_C28	
Peptidase_C3	
Peptidase_C30	
Peptidase_C31	
Peptidase_C32	
Peptidase_C33	
Peptidase_C34	
Peptidase_C36	
Peptidase_C37	
Peptidase_C39	
Peptidase_C4	
Peptidase_C41	
Peptidase_C42	
Peptidase_C47	
Peptidase_C48	
Peptidase_C5	

Peptidase_C50	
Peptidase_C53	
Peptidase_C54	
Peptidase_C55	
Peptidase_C57	
Peptidase_C58	
Peptidase_C6	
Peptidase_C69	
Peptidase_C7	
Peptidase_C8	
Peptidase_C9	
Peptidase_M1	Metalloprotease
Peptidase_M10	
Peptidase_M10_C	
Peptidase_M11	
Peptidase_M13	
Peptidase_M13_N	
Peptidase_M14	
Peptidase_M15	
Peptidase_M15_2	
Peptidase_M15_3	
Peptidase_M16	
Peptidase_M16_C	
Peptidase_M17	
Peptidase_M17_N	
Peptidase_M18	
Peptidase_M19	
Peptidase_M2	
Peptidase_M20	
Peptidase_M22	
Peptidase_M23	
Peptidase_M24	
Peptidase_M26_C	
Peptidase_M26_N	
Peptidase_M27	
Peptidase_M28	
Peptidase_M29	
Peptidase_M3	
Peptidase_M32	
Peptidase_M35	
Peptidase_M36	
Peptidase_M3_N	
Peptidase_M4	
Peptidase_M41	
Peptidase_M42	
Peptidase_M43	
Peptidase_M44	
Peptidase_M48	
Peptidase_M49	
Peptidase_M4_C	
Peptidase_M50	
Peptidase_M55	
Peptidase_M56	

Peptidase_M6	
Peptidase_M61	
Peptidase_M7	
Peptidase_M74	
Peptidase_M8	
Peptidase_M9	
Peptidase_M9_N	
Zn_peptidase	
Zn_peptidase_2	
Peptidase_S10	Serine protease
Peptidase_S11	
Peptidase_S13	
Peptidase_S15	
Peptidase_S21	
Peptidase_S24	
Peptidase_S28	
Peptidase_S29	
Peptidase_S3	
Peptidase_S30	
Peptidase_S31	
Peptidase_S32	
Peptidase_S37	
Peptidase_S39	
Peptidase_S41	
Peptidase_S48	
Peptidase_S49	
Peptidase_S49_N	
Peptidase_S51	
Peptidase_S55	
Peptidase_S58	
Peptidase_S6	
Peptidase_S64	
Peptidase_S66	
Peptidase_S7	
Peptidase_S8	
Peptidase_S9	
Peptidase_S9_N	
Alpha-amylase	Glycosyl hydrolase
Cellulase	
Glyco_hydro_10	
Glyco_hydro_11	
Glyco_hydro_12	
Glyco_hydro_14	
Glyco_hydro_15	
Glyco_hydro_16	
Glyco_hydro_17	
Glyco_hydro_18	
Glyco_hydro_1	
Glyco_hydro_20b	
Glyco_hydro_20	
Glyco_hydro_25	
Glyco_hydro_26	
Glyco_hydro_28	

Glyco_hydro_2_C	
Glyco_hydro_2	
Glyco_hydro_2_N	
Glyco_hydro_30	
Glyco_hydro_31	
Glyco_hydro_32C	
Glyco_hydro_32N	
Glyco_hydro_35	
Glyco_hydro_38C	
Glyco_hydro_38	
Glyco_hydro_39	
Glyco_hydro_3_C	
Glyco_hydro_3	
Glyco_hydro_42C	
Glyco_hydro_42	
Glyco_hydro_42M	
Glyco_hydro_43	
Glyco_hydro_45	
Glyco_hydro_46	
Glyco_hydro_47	
Glyco_hydro_48	
Glyco_hydro_49	
Glyco_hydro_4	
Glyco_hydro_52	
Glyco_hydro_53	
Glyco_hydro_56	
Glyco_hydro_57	
Glyco_hydro_59	
Glyco_hydro_61	
Glyco_hydro_62	
Glyco_hydro_63	
Glyco_hydro_65C	
Glyco_hydro_65m	
Glyco_hydro_65N	
Glyco_hydro_67C	
Glyco_hydro_67M	
Glyco_hydro_67N	
Glyco_hydro_68	
Glyco_hydro_6	
Glyco_hydro_70	
Glyco_hydro_71	
Glyco_hydro_72	
Glyco_hydro_76	
Glyco_hydro_77	
Glyco_hydro_79n	
Glyco_hydro_7	
Glyco_hydro_81	
Glyco_hydro_85	
Glyco_hydro_88	
Glyco_hydro_8	
Glyco_hydro_92	
Glyco_hydro_98C	
Glyco_hydro_98M	

Glyco_hydro_9	
Melibiase	
RolB_RolC	
Cutinase	Cutinase
CBM_14	Chitinase
ChiC	
Chitin_bind_4	
ChitinaseA_N	
Glyco_hydro_19	
Pec_lyase_C	Pectin lyase
Pec_lyase_N	
Pectate_lyase	
Pectate_lyase_2	
Pectinesterase	Pectinesterase

RT-PCR. Primers were designed using software Primer3 (24), and the primer sequences are described in the following table. The 5' end of the sequence is on the left.

Gene ID	Forward primer	Reverse primer
MGL_4054	GAGGGACGCCAGTCACTAGA	TTGAAGGGGAAACGTAGCAC
MGL_3507	ACACTTTGGACTCGCTCGAT	CCATTGTTGCACCACTCAAC
MGL_3878	TAGGTGATGCCGAGATTGTG	GGCATAGCAAAAGCAAAAGC
MGL_0799	AGCAGCCGTTCTAGCTCAAG	AAAGACGCAAGCAGACCAAT
MGL_0798	GCGATGGAAGCCTTAAACAA	GTTTGCAAAGGTTGGGTTTC
MGL_0797	GACAGCACGGAAAATGGTCT	CATCCATACGCAGCTCAATG
MGL_4252	TTCCGGACATTCTAAGGTG	GCTTCAGAGAGTCCGTCAGG
MGL_3326	CTGTTGAGCGCCTTTGATT	TTCTTACCCTGCGTCGAGAT
MGL_1986	GGTTACCCATTCACGACGAC	GCAAGAATCGAACCACCAAT

Phylogenetic reconstruction

Genomes and KOG assignment. Proteins of *Aspergillus fumigatus*, *A. nidulans*, *Botrytis cinerea*, *C. albicans*, *C. glabrata*, *C. guilliermondi*, *C. lusitaniae*, *Chaetomium globosum*, *Coccidioides immitis*, *Coprinopsis cinerea* (*Coprinus cinereus*), *Cryptococcus neoformans* JEC21, *C. neoformans* H99, *Debaryomyces hansenii*, *Eremothecium* (*Ashbya*) *gossypii*, *Fusarium graminearum*, *Kluyveromyces lactis*, *Magnaporthe grisea*, *M. globosa*, *M. restricta*, *Neurospora crassa*, *Phanerochaete chrysosporium*, *Rhizopus oryzae*, *Saccharomyces bayanus*, *S. castelli*, *S. kluyverii*, *S. kudriavzevii*, *S. mikatae*, *S.*

paradoxus, *Sclerotinia sclerotiorum*, *Stagonospora nodorum*, *Ustilago maydis* and *Yarrowia lipolytica* were assigned for orthologies using the STRING program (25). Subsequently, we selected only the KOGs represented by a single protein in order to avoid problems in orthology assessment due to paralogy in the KOGs represented by a higher number of protein families.

Alignment, GBlocks, and phylogenetic analysis. Each KOG protein shared among the 34 genomes and represented by a single protein (SI Table 11) was aligned by Clustal X (26). The phylogenetic informative blocks were selected by GBlocks program (27) using the parameters of 50% of minimum number of sequences for a conserved position and not allowing gap positions. Then, all proteins were concatenated, and phylogenetic analysis was done using Maximum Likelihood (PHYML) (28). The amino acid model substitution was JTT (29). The number of substitution rate categories was two. The model of rate heterogeneity was Gamma distribution rates with four categories.

RT-PCR methods. For the human scalp samples, RNA was isolated directly from cells that were present on a rayon swab wiped across human scalp according to the method of DeAngelis *et al.* (30). The samples shown are derived from one subject but representative of five subjects. For *in vitro* samples, cells were grown in mDixon medium. Isolated RNA was used for first-strand cDNA synthesis with reverse transcriptase and random hexamers and then PCR amplified using gene-specific primers as described (30). The human subjects were not measured for the presence of dandruff but were selected for having a relatively high level of *Malassezia* cells, as monitored by PCR (31). There were no bands observed in negative control experiments in the absence of reverse transcriptase or template (not shown). *M. globosa* genomic DNA was used as a positive control for primer design and PCR efficiency (not shown).

Proteomics Methods

Detailed proteomics conditions for SI Table 9.

Media Type

Rich Media *M. globosa* was grown in mDixon media (30).

Defined Media Final concentrations are listed in parentheses.

<u>Major components</u>	<u>Trace amounts</u>	<u>Amino acids</u>
Sorbitol (5%)	H ₃ BO ₃ (0.0005%)	Tryptophan (0.002%)
Tween 40 (1%)	MnCl ₂ ·4H ₂ O (0.001%)	Histidine (0.002%)
Glycerol (0.2%)	ZnCl ₂ (0.004%)	Arginine (0.002%)
Oleic Acid (0.2%)	NaMoO ₄ ·2H ₂ O (0.0004%)	Methionine (0.002%)
Tween 60 (0.4%)	FeCl ₃ ·6H ₂ O (0.0009%)	Tyrosine 0.003%)
Urea (2%)	CuSO ₄ ·5H ₂ O (0.004%)	Leucine (0.003%)
KH ₂ PO ₄ (1.5%)	Thiamin (0.0009%)	Isoleucine (0.003%)
NaSO ₄ (0.4%)	Riboflavin (0.0005%)	Lysine (0.003%)
KCl (0.7%)	Pyridoxin (0.0005%)	Phenylalanine (0.005%)
MgSO ₄ (0.2%)	Calcium Pantothenate (0.0002%)	L-Glutamine (0.01%)
CaCl ₂ (0.1%)	p-Aminobenzoic Acid (0.0005%)	L-Asparagine (0.01%)
	Choline Chloride (0.002%)	Valine (0.02%)
	Folic Acid (0.0009%)	Threonine (0.02%)
	α-Biotin (0.0009%)	Serine (0.04%)
	Adenine Sulfate (0.002%)	
	Uracil (0.002%)	

Extraction Type

Pierce PagePrep: Isolation of Proteins Secreted into the Media. Fifty milliliters of cultured media (rich or defined) was combined with 50 ml of DMSO, to which 0.6 ml of PagePrep Resin (Pierce Biotechnology) were added. Proteins were allowed to bind for 1 hr at room temperature with gentle agitation on a rocker. The resin was pelleted by centrifugation at 3,500 x *g*, and the supernatant was discarded. The resin was washed with 50:50 DMSO:water and centrifuged as above. Secreted proteins were eluted using

the elution buffer supplied in the PagePrep kit (one part resin to two parts elution buffer; 15 min gentle agitation). The resin was pelleted by centrifugation and discarded. The eluant was removed and prepared for 1D SDS PAGE.

EDTA: Extraction of Cell Wall-Associated Proteins. The procedure used was a modification of that described by Koyama *et al.* (32). Cells grown in rich or defined media were pelleted by centrifugation at 4,000 x *g* and washed three times with TE buffer [50 mM Tris-HCl (pH 8.0), 5 mM EDTA containing protease inhibitors (Roche)]. The cell pellet was resuspended in TE buffer, and cells were washed further by pelleting through a cushion of 20% sucrose at 10,000 x *g*. The washed cell pellets were resuspended in a buffer containing 100 mM EDTA, 300 mM 2-mercaptoethanol (one part cells to two parts buffer) and then gently mixed for 1 hr at room temperature. Cells were removed by centrifugation, and the extracted proteins in the supernatant were dialyzed [SnakeSkin dialysis tubing with 7000 MWCO (Pierce)] for 24 hr at 4°C against 0.1% acetic acid, lyophilized, and solubilized in either 1D SDS PAGE sample buffer (Invitrogen) or isoelectric focusing sample buffer.

SDS: Extraction of Cell Wall-Associated Proteins. Cells grown in rich or defined media were pelleted by centrifugation at 4,000 x *g* and washed three times with TE buffer [50 mM Tris-HCl (pH 8.0), 5 mM EDTA containing protease inhibitors (Roche)]. The cell pellet was resuspended in two volumes of 2X SDS PAGE sample buffer [(Invitrogen) 0.5 M Tris-HCl, pH 6.8, 5% glycerol, 10% SDS, 2% 2-mercaptoethanol, 0.1% bromophenol blue], plus 5 mM EDTA and protease inhibitors and incubated for 1 hr at room temperature with gentle agitation. The cells were pelleted as above and discarded, and the proteins in the supernatant were prepared for 1D SDS PAGE.

Gel Type

1D SDS PAGE was performed using 10-20% gels (Invitrogen). For 2D gel electrophoresis, proteins were focused using pH 4-7 strips (GE Healthcare) and separated on 10-20% gradient gels (Invitrogen). All gels were stained with Coomassie blue.

Details of these procedures have been given previously (33).

Identification of Proteins from Gels

Details for the identification of proteins by MALDI and electrospray mass spectrometry have been described (30, 33). For Edman sequencing, proteins separated by 1D SDS-PAGE were electroblotted to a PVDF membrane (Fluorotrans, Pall Life Sciences) and stained with Coomassie blue. Protein bands were excised from the blot for NH₂-terminal sequencing on a Procise 494 HT Protein Sequencer (Applied Biosystems) (30).

Section IV

Genome characterization: repeated elements and telomeres. The relatively small size of the *Malassezia (M.) globosa* genome reflects a low frequency of introns (SI Table 2) and repetitive elements, as well as a small number of genes (SI Table 4). Only 27% of all *M. globosa* predicted genes contain introns, with an average of 1.8 introns per intron-containing gene; these values are quite similar to those seen with *U. maydis* (43). Only 0.5% of the genome is comprised of simple repeats or regions of low complexity. The shortage of repeats probably is not due to the repeat-induced point mutation (RIP) system (first described in *Neurospora crassa*), because there is no apparent homolog of the *rid* methyltransferase (44), and the AT dinucleotide bias associated with RIP is not seen (45). There is also no sign in the *M. globosa* genome of components of the RNAi machinery that have been implicated in controlling transposable elements (46).

Six of the fifteen super-contigs contained telomere-like sequences at one end, with repeated copies of the TTAGTG sequence that is similar but not identical to the telomere sequences of other fungi (47).

Gene copy numbers within the human pathogen *Aspergillus fumigatus* genome. *A. fumigatus* is a common human fungal pathogen, although, unlike *Malassezia*, *A. fumigatus* is not commonly associated with skin. With its larger size (9,600 genes), the *A. fumigatus* genome contained a larger number of gene copies than any of the selected fungal genomes for 9 of the 14 genes listed in Table 1. *A. fumigatus* encodes 15 lipase genes, one more than found in *M. globosa* and *C. albicans*. In addition, *A. fumigatus*

encodes three phospholipase C, five phospholipase D, zero phospholipase B, two acid sphingomyelinase, and 10 aspartyl protease genes.

Features present in the *M. restricta* genome sequence but missing from the *M.*

***globosa* genome sequence.** Predicted proteins from *M. globosa* were aligned to the *M. restricta* genome using the Exonerate (18) protein2genome model. The *M. restricta* DNA sequences that were not aligned with the *M. globosa* proteome and were longer than 300 bp were searched against the NCBI nr database using Blastx. The best hits for each region from the nr database were aligned back to the genome using Exonerate to produce a set of *M. restricta* genes that may be missing from *M. globosa*. Each gene was manually inspected and searched against the *M. globosa* nucleotide sequence using tblastn. Using this method, we identified three genes that were found in *M. restricta* and not present in the *M. globosa* sequences. One gene matched nicotinate-nucleotide diphosphorylase, an enzyme involved in niacin synthesis. A second gene matched a NADH dehydrogenase subunit 5, and this gene was not found in the *U. maydis* genome as well. The third gene matched a hypothetical protein UM06042 (accession EAK86926) from *U. maydis*. Further work is required to learn whether these genes are truly absent from the *M. globosa* genome or if the genes are present in the genome and missed in the sequencing reactions.

There is no fungal-like fatty acid synthase (FAS) in *M. globosa*. First, there is little similarity between fungal FAS and any *M. globosa* gene. The fungal FASs have strong similarity to each other that can be easily identified using a BLAST search. Shown below is a BLAST alignment between a putative *U. maydis* FAS (UM10339, accession XP02971) with the beta (NP_012739.1) and alpha (NP_015093.1) subunits of *S.*

cerevisiae FAS as an example (Alignments 1 and 2). Using a *S. cerevisiae* FAS sequence, we were able to identify FAS or FAS fragments in all the free-living fungi currently with genome sequences available in the GenBank (SI Table 8). When using the putative *U. maydis* FAS as a query sequence against the *M. globosa* genome, the best match is *M. globosa* putative PKS MGL_1024, with a blast bit score of 62.4 (Alignment 3, shown below). This represents a much worse match than is observed between fungal FAS's.

Second, *M. globosa* lacks a gene with the domain architecture of a typical fungal FAS (Fig. 1). The *M. globosa* putative PKS contains domains in the same order as a typical fungal PKS and different from the order of a typical fungal FAS.

Third, it is unlikely that we simply failed to find *M. globosa* FAS gene segments in the course of the seven-fold sequencing coverage of the *M. globosa* genome. For an 11 kb gene, if one assumes random sequencing, the probability that an existing FAS would not be covered is about 10^{-70} . When we used the *U. maydis* putative FAS gene in a tblastn search against all the *M. globosa* sequencing reads and assembled genome, no other plausible matches were found. Taken together, we conclude there is no fungal-like FAS in *M. globosa*.

Alignment 1: Blast Alignment between *U. maydis* putative FAS and *S. cerevisiae* FAS beta subunit.

```

Score = 1383 bits (3580), Expect = 0.0
Identities = 800/2046 (39%), Positives = 1162/2046 (56%), Gaps = 84/2046 (4%)

Query: 103  LAAKFLGFLASRATLKTSSHELNLLQHAWAHFHDEFALTRGVSVDLVAIDSDLRTPV-- 160
           L  KFLG+++S              +L      F  + +L      +H L A +  + T +
Sbjct: 60   LVGKFLGYVSSLVEPSKVGQFDQVLNLCLEFENCYLEGN-DIHALAAKLLQENDTTLVK 118

Query: 161  ----LRSYFEAYVQLESAGQAKPSLITPKLFSLSAQNKAEVHAVFGGQGNNEAYFKELRT 216
           +++Y  A +  +      K +      LF      +  A++ A+FGGQGN + YF+ELR
Sbjct: 119  TKELIKNYITARIMAKRPFDKKSN---SALFRAVGEGNAQLVAIFGGQGNNTDDYFEELRD 175

Query: 217  LFDTYRPLVQPLLESIQPVLDKLVQDAVKQGFASYIYHGMVDVLAWLTGAAPPAVLYLAS 276
           L+ TY  LV  L++      L +L++  +      +  G+++L WL  + P  YL S
Sbjct: 176  LYQTYHVLVGDLIKFSATLSELIRTTLDA--EKVFTQGLNILEWLENPSNTPDKDYLLS 233

```

Query: 277 VPLSLPLIGLTQITQFLVSIRASGLTPAEFRS QIKSATGHSQGIISAVVLSASDSLESFY 336
 +P+S PLIG+ Q+ ++V+ + G TP E RS +K ATGHSQG+++AV ++ +DS ESF+
 Sbjct: 234 IPISCPGIVIGLAHYVVTAKLLGFTPGELRSYLKATGHSQGLVAVAIETDSWESFF 293

Query: 337 VNVAKALELLFHIGKRGQEFFPTLAIDPALVKDTEGGEGETPMLS SVSGLEQTVLEKHI 396
 V+V KA+ +LF IG R E +P ++ P++++D++E EG P+PMLS+S L Q ++ ++
 Sbjct: 294 VSVRKAITVLFFIGVRCYEAYPNTSLPPSILEDSLENNEGVPSPMLSISNLTQEQQVDYV 353

Query: 397 QKTNSHLAESQAVSISLFGNRPFSFIVTGYPKALVGLVGLSLRAIRADTGADQKIPFSKRK 456
 KTNSHL + V ISL NG ++ +V+G P++L GL +LR +A +G DQ +IPFS+RK
 Sbjct: 354 NKTNSHLPAGKQVEISLVNGAKNLVVGPPQSLYGLNLTLRKAKAPSGLDQSRIPFSERK 413

Query: 457 AVFRMNFLPVGPYPYHSSYLIGATEKMLQEDYAGRTIVWDASELKT TVKNTEDSSDLASYS 516
 F FLPV P+HS L+ A++ ++ +D + ++A +++ V +T D SDL S
 Sbjct: 414 LKFSNRFLPVASPFHSHLLVPASD-LINKDLVKNNVSFNAKDIQIPVYDTFDGSDLRVLS 472

Query: 517 SEDILKSIADQIFVKHIYWSKVS DLPKSATHLIDFXXXXXXXXXXXXXARILEGRGVRIIT 576
 I + I D I + W + ATH++DF R +G GVR++
 Sbjct: 473 GS-ISERIVDCIIRLPVKWETTTQF--KATHILDFGPGGASGLGLVTLHRNKDGTGVRVIV 529

Query: 577 -----PSGTHARSTELYNAST--VKFEERWEDVWSPRIKTKNG- IKIDTPFSRTTGR 626
 P + E+++ ++ +K W + + P+LIK K+G I ++T FS+ GR
 Sbjct: 530 AGTLDINPDDDYGFKQEIFDVT SNGLKKNPNWLEEYHPKLIKNSGKIFVETKFSKLIGR 589

Query: 627 GILMVGGMTPTTVSANINAAVLNAGYVIELAGGGLHSEAGLRKRVNEILAQXXXXXXXXXX 686
 L+V GMTP TVS + AA NAGY IELAGGG S AG+ ++ +++Q
 Sbjct: 590 PPLLVPGMTPTCTVSPDFVAATTNAGYTIELAGGGYFSAAGMTAAIDSVVSQIEKGSTFGI 649

Query: 687 XXXFIDQRRFGWQYPLWGQMRREGFPCKGFTVAAGIPSTEKAKEIIDGLKAAGIEYVGFK 746
 +++ W PL ++R +G+P + T+ AG+PS E A E I+ L G++Y+G K
 Sbjct: 650 NLIYVNPFMLQWGIPLIKELRSKGYP IQFLTIGAGVPSLEVASEYIETL---GLKYLGLK 706

Query: 747 PGTADGLRAVAAAIAANPDFKIICQWTGGRAGGHHSTEDFHQPILQTYASLRKQKNLILV 806
 PG+ D + V IA A+P+F I QWTGGR GGHHS ED H P+LQ Y+ +R+ N++L+
 Sbjct: 707 PGSIDAISQVINIAKAHPNFPIALQWTGGRGGGHHSFEDAHTPMLQMYSKIRRHPIMLI 766

Query: 807 GSGFGDAEGVWPYLTGEWSVKLGYPMPFSAFLFASWMLVAKEAATADAVKQLLVDTPG 866
 GSGFG A+ +PYLTGEWS K Y PMPF FLF S +++AKE T+ K+ + G
 Sbjct: 767 FSGFGSADDTYPYLTGEWSTKFDYPMPFDGFLGSRVMIKEVKTSPDAKKCIAACTG 826

Query: 867 CGDDKWQD TYGKGAGGTITCNSELGESISYVHNRAALWRQLDDQLFSLPREKQRVWLSE 926
 DDKW+ TY K GG +T SE+GE I + R LW++ D+ +F+LP+ K L
 Sbjct: 827 VPDDKWEQTYKKPTGGIVTVRSEMGEPIHKIATRGMVLWKEFDETFNLPKNKLVP TLEA 886

Query: 927 KKS WLIERLNKDFQKPWF AAKLDGTVLMDVADMTYEEVTRRMLELLFVGHQSRWIDPSLK 986
 K+ ++I RLN DFQKPWFA ++G D+A MTYEEV +R++EL+F+ + W D + +
 Sbjct: 887 KRDIISRLNADFQKPWFAT-VNGQA-RDLATMTYEEVAKRLVELMFIIRSTNSWFDV TWR 944

Query: 987 KLFGDWCRRVEERFAGIETAGKESLLQSYSVLDKEPSKFVD AFFAEYTGAK EIRLAED 1046
 GD+ RRV EERF T K SL+QSYS+LDK P + ++ F Y A+E L A+D
 Sbjct: 945 TFTGDFLRRVEERF----TKSKTSLIQSYSLDK-PDEAIEKVF NAYPAAREQFLNAQD 999

Query: 1047 VGYFLTICQRRGQKPVFPIPELGDNFQTLFKKDSLWQSEDLDAVVGKDAQKVIILQGSVA 1106
 + +FL++CQ QKVPF+P L F+ FKKDSLWQSE L+AVV +D Q+ IL G VA
 Sbjct: 1000 IDHFLSMCQNPQM KPVFPVLD RRF EIFFKKDSLWQSEHLEAVVDQDVQRTCILHGPVA 1059

Query: 1107 VQHCTKANVPAGEMLGDI EKQLVAKVLAKYYDNDESKVPVADYVSNDPLAIDEADVYARY 1166
 Q + P ++ I + K+L +YY +DESK+P +Y E+ V +
 Sbjct: 1060 AQFTKVIDEPIKSIMDGIHDGHIKLLHQYYGDDESKIPAVEYFGG-----ESPVDVQS 1113

Query: 1167 GVEASIEGNLKT LKIGSQLPDTTEWLEALATRRTSWLRALLRNVSVVQGKNYADNPMLRM 1226
 V++S + + D W +ALA +W A + Q K + NP+ ++
 Sbjct: 1114 QVDSSSVSEDSAVFKATSSTDEESWFKALAGSEINWRHASFLCSFITQDKMFVSNPIRKV 1173

Query: 1227 LAPRKNQTAKIAYAADGEPISLTLFGAARSYGGHDDGFKALEITKSSDSAKIQVTLFEER 1286
 P + +I+ +G S T+ + G K I K IQ+ + E R
 Sbjct: 1174 FKPSQGMVVEIS---NGNTSSKTVVTLSEPVQGE---LKPTVILKLLKENIIQMEMIENR 1227

Query: 1287 L--GESIPLHFYFTYRPDQPFASIHEVMNGRNDRIKQFYWLWFHXXXXXXXXXXXXXTQ- 1343
 G+ + L + + PD FA I EVM RN RIK+ YW+LW +
 Sbjct: 1228 TMDGKPVSLPLLNFNPNGFAPISEVMEDRNQRIKEMYWKLWIDEFPNLD FDRDVIK 1287

Query: 1344 --FESPVKSLDETAIRTCRIVENRGEVFTKGQ-----APMDFAIVAGWEAIMQALMASC 1396
 FE K + + F V N E F APMDFAIV GW AI++A+ +

Sbjct: 1288 KDFEITAKEVYD-----FTHAVGNNCEDFVSRPDRTMLAPMDFAIVVGWRAIIKAIFPNT 1342

Query: 1397 -DADLLSLVHLSNSFKVIDDARPLRAGDVCTATAHASSIKISDTGKSVSVTGTVLRKDAS 1455
 D DLL LVHLSN +K+I A+PL+ GDV + TA S+ TGK V V GT+ R

Sbjct: 1343 VDGDLLKLVHLSNGYKMIPGAKPLQVGDVSTTAVIESVVNQPTGKIVDVVGTLSRNGKP 1402

Query: 1456 GDXXXXXXXXXXXXXXXXXXXXXXXXXETSQNEFTLEIKTKADATVLLAKEWFDWVAPQPL 1515
 +T + + + IKT D VL +KEWF

Sbjct: 1403 ----VMEVTSSFFYRGNYDFENTFQKTVEPVYQMHIKTSKDIAVLRKEWFLDDEDFD 1458

Query: 1516 APGTVLHFNIQSLLKYKDSTSYSDVEVTGGAYVRDYKDDLIKVGIEIYTADAISYGNPVI 1575
 L F ++ + +K++ +S V+ G V + +++G ++Y A A +GNPV+

Sbjct: 1459 LLNKTLTFETETEVTFKNANIFSSVKCFGIKVELPTKETVEIGIVDYEAGA-SHGNPVV 1517

Query: 1576 EYIKRNGGASTGGVAPLETGYTLGASGVPTTFVAPATNEPYSNASGDFNPIHINPYFSSF 1635
 +++KRNG V LE + T P+TNEPY+ SGD NPIH++ +F+S+

Sbjct: 1518 DFLKRNGSTLEQKVN-LENPIPIAVLDSYT----PSTNEPYARVSGDLNPIHVSRLFASY 1572

Query: 1636 ASLPGTITHGLFSSAASRKFEVVAEGYPDRCLSFEANFVGMVLPGDELTVKLRHIAMR 1695
 A+LPGTITHG+FSSA+ R +E AA+ R + FV MVLPL L ++H+ M

Sbjct: 1573 ANLPGTITHGMFSSASVRALIENWAADSVSSRVRGYTCQFVDMVLPNTALKTSIQHVMI 1632

Query: 1696 DGNKVIKVFETFNQNGDKVIDGQAEVKQPATVYVFTGQGSQEQGMGMDLYNSSPTAKALWD 1755
 +G K+IK ET N++ V+ G+AE++QP T +VFTGQGSQEQGMGMDLY +S A+ +W+

Sbjct: 1633 NGRKLIKFETRNEDDVVLTGEAEIEQPVTTTFVFTGQGSQEQGMGMDLYKTSKAAQDVWN 1692

Query: 1756 EADGHLRSTMGFSILEIINQNPLTKTIHFGGVRGHTIRERYMSMTYESTDADGNTVTLP 1815
 AD H + T GFSIL+I+ NP+ TIHFGG +G IRE Y +M +E T DG T +

Sbjct: 1693 RADNHFKDYGFSILDIVINNPNVNLTIHFGGEGKGRIRENYSAMIFE-TIVDGKLKTEKI 1751

Query: 1816 FPEISASSQSYTFQSPKGLLFATQFAQIALVLVELSAFRDMKQRLITPDAAFAGHSLGE 1875
 F EI+ S SYTF+S KGLL ATQF Q AL L+E +AF D+K +GLI DA FAGHSLGE

Sbjct: 1752 FKEINEHSTSYTFRSEKGLLSATQFTQPALTLMEKAAFEDLKSGLIPADATFAGHSLGE 1811

Query: 1876 YXXXXXXXXXXXXXXXXXCDVVFYRGLSMQNCVQRDPVTGASNYAMMACNPLRIVGEGNPAF 1935
 Y +VVFYRG++MQ V RD + G SNY M+A NP R+ +F

Sbjct: 1812 YAALASLADVMSIESLVEVVFYRGMTMQVAVPRDEL-GRSNYGMIAINPGRVAA----SF 1866

Query: 1936 AGQALAEIVDSISKESGRLLQIVNHNVAQQYVYTAGDLVALTTLGNVLFNFIKIQKIDLKK 1995
 + +AL +V+ + K +G L++IVN+NV QQYV AGDL AL T+ NVLNF+K+QKID+ +

Sbjct: 1867 SQEALQYVVERVGRKRGWLVEIVN+VENQQYVAAGDLRALDVTNVLNFIKLQKIDIIE 1926

Query: 1996 LNEIMPIEEVREKLAIEVRELSAALKEEMKGR-IELKRGFATIPLPGIDVPFHRSYLTG 2054
 L + + +EEV L EI+ E A+ + +K R ++L+RGFA IPL GI VPFHYS YL

Sbjct: 1927 LQKSLSEEEVGHLEIIDE---ASKKSAVKPRPLKLERGFACIPLVGISVPFHSTYLMN 1983

Query: 2055 GVAPFRSYLSKRINVDAIRPQLLVGRYIPNLVAAPFELSKSYAELIFAQTDSPRLDKVLK 2114
 GV PF+S+L K I + ++ L G+YIPNL A PF+++K Y + ++ T S + +++

Sbjct: 1984 GVKPFKSFLKKNIIKENVKVARLAGKYIPNLAKPFQVTKEYFQDVYDLTGSEPIKEIID 2043

Query: 2115 NWEKDE 2120
 NWEK E

Sbjct: 2044 NWEKYE 2049

Alignment 2: Blast Alignment between *U. maydis* putative FAS and *S. cerevisiae* FAS alpha subunit.

Score = 1304 bits (3375), Expect = 0.0
 Identities = 670/1429 (46%), Positives = 921/1429 (64%), Gaps = 33/1429 (2%)

Query: 2124 PDRVQRLTATLVVELLAYQFASPVRWIETQDIFFTDYNFERLVEFGPSPTLVGMAKRTQA 2183
 P+ Q L L+ ELLAYQFASPVRWIETQD+F D+N ER+VE GPSPTL GMA+RT

Sbjct: 3 PEVEQELAHILLTELLAYQFASPVRWIETQDVFLKDFNTERVVEIGPSPTLAGMAQRTLK 62

Query: 2184 LKYSVDLAQGRKRVMYCSSKDKEALYYAFADDE----SEXXXXXXXXXXXXXXXXXXX 2239
 KY + D A R + C SKD + +YY E E

Sbjct: 63 NKYESYDAALSLHREILCYSKDAKEIYYTPDPSELAAKEEPAKEEAPAPTPAASAPAPAA 122

Query: 2240 XXXXXXXXXXXXXXXXXXXXDEPLKAVDTVRAILAQKLKKNISEVPLSKSIKDSVGGKSVL 2299
DEP+KA + ++A KLKK++ +P+SK+IKD VGGKS +
Sbjct: 123 AAPAPVAAAAPAAAAEIADEPVKASLLHLVLAHKLKSLDSIPMSKTIKDLVGGKSTV 182

Query: 2300 SNELVSDLLAEFGNXXXXXXXXXXXXXSALQAGYTGTGKTTSNLVTRTVAGKLPGGFN 2359
NE++ DL EFG Q ++G LGK +S+L++R ++ K+PGGF
Sbjct: 183 QNEILGDLGKEFGTTPKEETPLEELAETFQDTFSGALGKQSSSLLSRLISSKMPGGFT 242

Query: 2360 LSAVKSHFQKAWGLGAGRTDGALLVAVTMEPAKRLGSEAEAKTWLDSVAQTYAQLAGISL 2419
++ + + Q WGL +GR DG LLVA++ EPA RLGSEA+AK +LDS+AQ YA + G+ L
Sbjct: 243 ITVARKYLQTRWGLPSGRQDGVLLVALSNEPAARLGSEADAKAFLDSMAQKYASIVGVDL 302

Query: 2420 AQAXXXXXXXXXXXA----ISSEELDKLKAQQDEHVRRQIQVLERYLNMDGRAGGRLA 2475
+ A A I + L+++ R+Q+QVL RYL MD G R
Sbjct: 303 SSAASASGAAGAGAAAGAMIDAGALEEITKDHKVLARQQLQVLARYLKMOLDNGERKFL 362

Query: 2476 STKXXXXXXXXXXXXXXXXXHGDTYNGIMPKEKVRHFSSYWNWARQDAMTLLYDIIY 2535
K G+ + NG+ F+ KK R F S WNW+Q ++L ++II+
Sbjct: 363 KEKDTVAELQAQLDYLNAELGFEFFVNGVATSF SRKKARTFDSSWNWAKQSLLSLYFEIIH 422

Query: 2536 GRLTTVDREITARCITLMNRADPSLLRYMQYHIDQIDANQGPTYALAKEYGQQLIENCRE 2595
G L VDRE+ + I +MNR++ +L+++M+YHI D +G Y L K G+QLIENC++
Sbjct: 423 GVLKNVDREVSEAINIMNRSNDALTKFMEYHISNTDETKGENYQLVKTLGEQLIENCKQ 482

Query: 2596 ACSEDAVYKEVHAPTAPHTeidARGNIAYSEVkrVGRKLEAYVKEMAAGSRLLGAGSQV 2655
D VYK+V PT P T ID GNI YSE R VRKL YV+EMA G + SQ
Sbjct: 483 VLDVDPVYKDVAKPTGPKTAIDKNGNITYSEEPREKVRKLSQYVQEMALGGPITKE -SQP 541

Query: 2656 NLDKAQENVAKLWTLIRNEPSVSKLSKATIKSLYSEVVRSLGXXXXXXXXXXXXXXXXXL 2715
+++ V K + ++ +S ++ + LYS++++ L +
Sbjct: 542 TIEEDLTRYKAISAQADKQDISSSTRVEFEKLYSMLMKFLESSKEIDPSQTTQLAGMDV 601

Query: 2716 SSSTQRPQVSDAASTEDRT-----PFLHVKKRV-AGQWQLSRKLTSVYLD 2759
+ + + AS +++ PFLH+++K AG W+ R+L+S++LD
Sbjct: 602 EDALDKDSTKEVASLPNKSTISKTVSSTIPRETIPFLHLRKKTPAGDWKYDRQLSSLFLD 661

Query: 2760 ILGEMASQGVSFKGRNALLTGVGKASIGVEVVKGLLAGGARVVITTSYSRSRVVEYYQRI 2819
L + A GV+FK + L+TG GK SIG EV++GLL GGA+VV+TTS +S+ V +YYQ I
Sbjct: 662 GLEKAAFNGVTFKDKYVLITGAGKGSIGAEVLQGLLQGGAKVVVTTSRFSKQVTDYYQSI 721

Query: 2820 YQEVGSRGSTLTVPFNQASKQDQVDNLVKYIY----GSLDLDLDFVVPFAAISENGREI 2874
Y + G++GSTL VVPFNQ SKQDV+ L+++IY G L DLD ++PFAAI E G E+
Sbjct: 722 YAKYGAKGSTLIVVPFNQGSKQDVEALIEFIYDTEKNGGLGWDLDIIPFAAIEPQGI 781

Query: 2875 DGIDDKSELAHRLMLTNLVRLLGAIKTEKASRNFDRPTQVLLPLSPNHGTFGGDGLYGE 2934
+ ID KSE AHR+MLTN++R++G +K +K++R +TRP QV+LP+SPNHGTFGGDG+Y E
Sbjct: 782 EHIDSKSEFAHRIMLTNILRMMGCVKKQKSARGIETRPQVILPMSPNHGTFGGDGMYS 841

Query: 2935 SKLGLLETLNRWESESWGQYLCITGAIIGWCRGTGLMSSSNTVAEEIEKLGCRFTSTTEM 2994
SKL LETL NRW SESW L + GAIIGW RGTGLMS++N +AE IEK+G RTFS EM
Sbjct: 842 SKLSLETLFNRWHSSESWANQLTVCGAIIGWTRGTGLMSANNIAEGIEKMGVRTFSQKEM 901

Query: 2995 AFNLLGLLAPTVSQVEALLVDLNGMDKIVDLAEQTGKFRAAIQEASSTKRALVADN 3054
AFNLLGLL P V + Q ++ DLNGG+ + +L E T K R + E S ++A+ +
Sbjct: 902 AFNLLGLLTPEVVLCQKSPVMADLNGGLQFVPELKEFTAKLRKELVETSEVRKAVSIET 961

Query: 3055 SADFKVIKGSAAEALHKKVHINPRSNFTFRFPEVGSLLQNLREIAKSYSQ--MDLEKVVVI 3112
+ + KV+ G++A+A + +V I PR+N FPE+ + ++IA + + +DLE+V+V+
Sbjct: 962 ALEHKVVNGNSADAAYAQVEIQPRANIQLDFPELKPYPKQVKIAPAELEGLDLERVIVV 1021

Query: 3113 TGFAEVGPWGSRRTRWEMAFGHLTNEGVIELAWIMGLIKHHDGQLKSGATYVGWVDVKS 3172
TGFAEVGPWGS+RTRWEMAFG + EG +E+AWIMG I +H+G LK G Y GWVD K+
Sbjct: 1022 TGFAEVGPWGSARTRWEMAFGEFSLEGCVEMAWIMGFISYHNGNLK-GRPYTGWVDSKT 1080

Query: 3173 GEPVDDKDIRANYEEHIIKHAGIRLIEPELFRNYDPNRKGFQOEIENHLDLEPFVEVSAAE 3232
EPVDDKD++A YE I+++GIRLIEPELF Y+P +K QE+ + DLEPFE S E
Sbjct: 1081 KEPVDDKDVAKYETSILEHSGIRLIEPELFNGYNPEKKEMIQEVIVEEDLEPFASKET 1140

Query: 3233 ASKYLEHGDKVDSWQDANTGGWVFLKKGARIFVPKAVAFDRLVAGQLPTGWSGQRYGI 3292
A ++K +HGDKVD ++ TG + V L KGA +++PKA+ FDRLVAGQ+PTGW+ + YGI

Sbjct: 1141 AEQFKHQHGDKVDIFEIPETGEYSVKLLKGATLYIPKALRFDRVLVAGQIPTGWNAKTYGI 1200

Query: 3293 PDDIVSQVDRRTLWALVCASEALINSGVSDPYELYKHFPKSVKVGSCIGSGMGMTSLSKM 3352
 DDI+SQVD TL+ LV EA I SG++DPYE+YK+ H S+VG+C GSGMGG+++L M

Sbjct: 1201 SDDIISQVDPITLFLVLSVVEAFIASGITDPYEMYKYVHVSEVGNCSSGSGMGGVSALRGM 1260

Query: 3353 FRDRRNDTDVQNDILQETFINSVAGWVNLVMSAAGPIKTVVGACATALQSIDVAVETIL 3412
 F+DR D VQNDILQE+FIN+++ WVN+L++S++GPIKT VGACAT+++S+D+ VETIL

Sbjct: 1261 FKDRFKDEPVQNDILQESFINTMSAWVNMLLISSSGPIKTPVGACATSVESVDIGVETIL 1320

Query: 3413 SGKAELMLAGGFDDLSEEGSYEFANMKATSNNGKEELAAGREPSEHSRPFSTSRSGFMESQ 3472
 SGKA + + GG+DD EEGS+EF NMKATSN EE GR P+E SRP +T+R+GFME+Q

Sbjct: 1321 SGKARICIVGGYDDFQEEGSFEFGNMKATSNLTLEEFHGRTPAEMSRPATTTNRNGFMEAQ 1380

Query: 3473 GCGVQVLTAAANAIKLGMPIQSIVAYSNTHTDGAGRSLPAPGHGVIGSA 3521
 G G+Q++ A A+K+G+PI IVA + T TD GRS+PAPG G++ +A

Sbjct: 1381 GAGIQIIMQADLALKMGPVIYGIVAMAATATDKIGRSVPAPGKGILTAA 1429

Score = 254 bits (650), Expect = 3e-70
 Identities = 146/366 (39%), Positives = 216/366 (59%), Gaps = 10/366 (2%)

Query: 3524 MQRALARFNLTADDIGVISAHMTATKANDKNESHVYQELFTRMGRTPGHAVPVMAQKWLK 3583
 ++ ALA + LT DD+GV S H T+TKANDKNES E+ +GR+ G+ V + QK+L

Sbjct: 1522 LRGAATYGLTIDDLGVASFHGTSTKANDKNESATINEMMKHLGRSEGNPVIGVFQKFLT 1581

Query: 3584 GHAKGGAAAWALNGVMQALHTSIVCGNRNADDISPELQKFDMLLYNSTSIHRTPHHVNA 3643
 GH KG A AW +NG +Q L++ I+ GNRNAD++ L++F+ +LY S ++ V A

Sbjct: 1582 GHPKGAAGAWMMNGALQILNSGIIPGNRNADNVDKILEQFEYVLYPSKTL--KTDGVRV 1639

Query: 3644 AVSSFGFGQVGGIILVLAHAILGRLSDKAFFEEYVARRKARQHQTTRMHSA LTKEDLVR 3703
 +++SFGFGQ GG +V+H ++ G +++ + EYVA+ AR+ Y H+ + L

Sbjct: 1640 SITSFGFGQKGGQAI VVHPDYLYGAITEDRYNEYVAKVSAREKSAYKFFHNGMIYNKLFV 1699

Query: 3704 IKDNRWPAAELEDQVLMNLNARAVEIGDSYGFKAPLPMPALATVEDTIVPAKPANVAAQ 3763
 K++ P+ ELE+ V ++ AR + S + + +D+ + A A

Sbjct: 1700 SKEHAPYTDELEEDVYLDPLARVSKDKKSGSLTFNSKNIQS---KDSYNANTIIETAKM 1755

Query: 3764 QENLEKMMGHVAGVGIDCEKISTFSPDNDNFVSRNFTDVEIEYCRAQPDPRASFGRFCA 3823
 EN+ K GVG+D E I++ +ND F+ RNFT EIEYC AQP ++SF G + A

Sbjct: 1756 IENMTKEKVSNGGVGVDELITSINVENTFIERNFTPQEIEYCQAQPSVQSSFAGTWSA 1815

Query: 3824 AEAVVKSGLGVRTQGAGASLKDVEIV-PTPEGPQVKLYGEALKASNGSRFL--VSISHGD 3879
 EAV KSLGV++ G GA+LKD+EIV P V+L+G A KA+ + VSISH D

Sbjct: 1816 KEAVFKSLGVKSLGGGAALKDIEIVRVNKNAPAVELHGNAKKAAEEAGVTDVKVVSISHDD 1875

Query: 3880 DTAMAI 3885
 A+A+

Sbjct: 1876 LQAVAV 1881

Alignment 3: Blast Alignment of *U. maydis* FAS with *M. globosa* PKS (MGL_1024)

Score = 62.4 bits (150), Expect = 1e-09
 Identities = 62/277 (22%), Positives = 111/277 (40%), Gaps = 33/277 (11%)

Query: 3388 GPIKTVVGACATALQSIDVAVETILSGKAELMLAGGFD-DLSEEGSYEFANMKATSNNGKE 3446
 GP + AC++++ + +A IL G ++ + + L+ S F+ + S

Sbjct: 167 GPSVYLDTACSSITCLHLARAAILRGDCD MAVVAVNILAPHASLSFSTLGTLS---- 222

Query: 3447 ELAAGREPSEHSRPFSTSRSGFMESQGCGVQVLTAAANAIKLGMPIQSIVAYSNTHTDGA 3506
 PS F S G+ +GC V +L A+++G I S + + + +G

Sbjct: 223 -----PSGLCHTFDASADGYSRGE GCTVIILQRTNEAVRMGSHIYSEITGTAINANGK 275

Query: 3507 GRSLPAPGHGVIGSADGMQRALARFNLTAD-----DIGVISAHMTATKANDKNESHVYQ 3560
 G+S+ P DG Q+ A + D + + H T T D E++

Sbjct: 276 GKSITLP-----DGPQKKAATYAAAYRDAKREPHEAAYVECHGTGTPVGDPIEANAVG 327

Query: 3561 ELFTRMGRTPGHAVPVMAQKWLKCGHAKGGAAAWALNGVMQALHTSIVCGNRNADDISPEL 3620
 E+FT GR + + + K GH + A L L ++ + + + SP +

Sbjct: 328 EVFT-PGREKDDYLRIGSAKTNVGHLEPAAGLVGLIKNCYTLDMGLMLPHLHFNKPSPRI 386

```

Query: 3621 QKFDMLLYNSTSIHRTPHHVNA-----AAVSSFGFG 3651
      +  D L+   T      P + +      ++SSFGFG
Sbjct: 387  RWDDFLIKVQTETEPLPPNKLSKDGKFIVSLSSFGFG 423

```

***M. globosa* gene MGL_1024 likely encodes a PKS.** The *M. globosa* genome contains one gene, MGL_1024, that is similar to both type I PKS and animal FAS. Type I PKSs and animal FASs are closely related to each other and have similar domain structures (Fig. 1). We believe that MGL_1024 most likely encodes a PKS instead of a FAS. First, blast searches indicate that MGL_1024 is most similar to PKSs in the database (SI Table 6). Among animal proteins, the best match is a sea urchin PKS. Second, the C- terminal domain encoded by MGL_1024 is a thioester reductase commonly found in nonribosomal peptide synthetase and PKSs (SI Table 7) instead of the thioesterase found in animal FAS. Third, *M. globosa* are lipid-dependent for growth, and the absence of a functional fatty acid synthase provides a straightforward explanation of this lipid dependency.

References

1. Green P <http://www.phrap.org>.
2. Gordon D, Abajian C, Green P (1998) *Genome Res* 8: 195-202.
3. Gordon D, Desmarais C, Green P (2001) *Genome Res* 11: 614-625.
4. Boekhout T, Kamp M, Gueho E (1998) *Med Mycol* 36: 365-372.
5. Boekhout T, Bosboom RW (1994) *Syst. Appl. Microbiol.* 17: 147-153.
6. Mott R (1997) *Comput Appl Biosci* 13: 477-478.
7. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) *Nucleic Acids Res* 25: 3389-3402.
8. Birney E, Clamp M, Durbin R (2004) *Genome Res* 14: 988-995.
9. Guigo R (1998) *J Comput Biol* 5: 681-702.
10. Korf I, Flicek P, Duan D, Brent MR (2001) *Bioinformatics* 17 Suppl 1: S140-148.
11. Eddy SR (1998) *Bioinformatics* 14: 755-763.
12. Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, White O (2001) *Nucleic Acids Res* 29: 41-43.
13. Eddy SR <http://hmmer.janelia.org/>.
14. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, *et al.* (2003) *BMC Bioinformatics* 4: 41.
15. Tatusov RL <ftp://ftp.ncbi.nih.gov/pub/COG/old/util/xugnitor.c>.
16. Lowe TM, Eddy SR (1997) *Nucleic Acids Res* 25: 955-964.
17. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B (2000) *Bioinformatics* 16: 944-945.

18. Slater GS, Birney E (2005) *BMC Bioinformatics* 6: 31.
19. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) *Nucleic Acids Res* 32: D277-280.
20. Haas BJ, Delcher AL, Wortman JR, Salzberg SL (2004) *Bioinformatics* 20: 3643-3646.
21. Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, *et al.* (2005) *Nucleic Acids Res* 33: D192-196.
22. Tuckwell D, Lavens SE, Birch M (2006) *Mycol Res* 110: 1140-1151.
23. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) *J Mol Biol* 340: 783-795.
24. Rozen S, Skaletsky H (2000) *Methods Mol Biol* 132: 365-386.
25. Snel B, Lehmann G, Bork P, Huynen MA (2000) *Nucleic Acids Res* 28: 3442-3444.
26. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) *Nucleic Acids Res* 25: 4876-4882.
27. Castresana J (2000) *Mol Biol Evol* 17: 540-552.
28. Guindon S, Gascuel O (2003) *Syst Biol* 52: 696-704.
29. Jones DT, Taylor WR, Thornton JM (1992) *Comput Appl Biosci* 8: 275-282.
30. Deangelis YM, Saunders CW, Johnstone KR, Reeder NL, Coleman CG, Kaczvinsky JR, Jr., Gale C, Walter R, Mekel M, Lacey MP, *et al.* (2007) *J Invest Dermatol* 127: 2138-2146.

31. Gemmer CM, DeAngelis YM, Theelen B, Boekhout T, Dawson Jr TL, Jr. (2002) *J Clin Microbiol* 40: 3350-3357.
32. Koyama T, Kanbe T, Ishiguro A, Kikuchi A, Tomita Y (2001) *J Dermatol Sci* 26: 201-208.
33. Keough T, Lacey MP, Fieno AM, Grant RA, Sun Y, Bauer MD, Begley KB (2000) *Electrophoresis* 21: 2252-2265.
34. Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, Prensier G, Barbe V, Peyretilade E, Brottier P, Wincker P, *et al.* (2001) *Nature* 414: 450-453.
35. Slamovits CH, Fast NM, Law JS, Keeling PJ (2004) *Curr Biol* 14: 891-896.
36. Smulian AG (2001) *Fungal Genet Biol* 34: 145-154.
37. JGI in <http://www.jgi.doe.gov/>.
38. Castoe TA, Stephens T, Noonan BP, Calestani C (2007) *Gene* 392: 47-58.
39. Smit AFA, Hubley R, Green P *RepeatMasker Open-3.1*.
40. Fitzpatrick DA, Logue ME, Stajich JE, Butler G (2006) *BMC Evol Biol* 6: 99.
41. Kuramae EE, Robert V, Snel B, Weiss M, Boekhout T (2006) *FEMS Yeast Res* 6: 1213-1220.
42. James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, Cox CJ, Celio G, Gueidan C, Fraker E, Miadlikowska J, *et al.* (2006) *Nature* 443: 818-822.
43. Kämper J, Kahmann R, Bolker M, Ma LJ, Brefort T, Saville BJ, Banuett F, Kronstad JW, Gold SE, Muller O, *et al.* (2006) *Nature* 444: 97-101.
44. Freitag M, Williams RL, Kothe GO, Selker EU (2002) *Proc Natl Acad Sci USA* 99: 8802-8807.

45. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S, *et al.* (2003) *Nature* 422: 859-868.
46. Lippman Z, Martienssen R (2004) *Nature* 431: 364-370.
47. Rehmeier C, Li W, Kusaba M, Kim YS, Brown D, Staben C, Dean R, Farman M (2006) *Nucleic Acids Res* 34: 4685-4701.

SI Table 5. *M. restricta* and *M. globosa* have similar putative PKS genes

<i>M. restricta</i> contig	Identity (%)	MGL PKS start (AA)	MGL PKS end (AA)	MRE start (bp)	MRE end (bp)	e-value
MRE_CONTIG_1609	69.86	2	74	1459	1677	2e-22
MRE_CONTIG_2021	83.82	161	772	1871	36	0.0
MRE_CONTIG_2602	76.28	816	1788	2	2923	0.0
MRE_CONTIG_1843	79.79	1917	2297	2	1144	2e-179
MRE_CONTIG_1324	76.16	2396	2737	1136	105	8e-153

Alignment summary of tblastn search of *M. globosa* (MGL) PKS (MGL 1024) against *M. restricta* (MRE) genome. From five contigs of *M. restricta* DNA, there is sequence similar to 87% (2,381/2,737) of the length of the putative *M. globosa* PKS gene.

SI Table 6. Best hits of *M. globosa* putative PKS gene to NCBI nr database

Rank	Accession No.	Description	Classification	E value	PKS annotation? Y=yes
1	NP_925775.1	similar to polyketide synthase [<i>Gloeobacter violaceus</i> PCC 7421]	Bacteria	3.00E-130	Y
2	AAK57186.1	MxaB2 [<i>Stigmatella aurantiaca</i>]	Bacteria	4.00E-130	Y
3	ABA29782.1	MmxB [<i>Cystobacter fuscus</i>]	Bacteria	4.00E-130	Y
4	ZP_00108795.1	COG3321: Polyketide synthase modules and related proteins [<i>Nostoc punctiforme</i> PCC 73102]	Bacteria	1.00E-129	Y
5	CAD19091.1	StiG protein [<i>Stigmatella aurantiaca</i>]	Bacteria	9.00E-127	Y
6	YP_890975.1	mycocerosic acid synthase [<i>Mycobacterium smegmatis</i> str. MC2 155]	Bacteria	3.00E-126	
7	NP_924900.1	similar to polyketide synthase [<i>Gloeobacter violaceus</i> PCC 7421]	Bacteria	1.00E-125	Y
8	YP_632694.1	polyketide synthase type I [<i>Myxococcus xanthus</i> DK 1622]	Bacteria	3.00E-125	Y
9	NP_485689.1	hypothetical protein all1649 [<i>Nostoc</i> sp. PCC 7120]	Bacteria	2.00E-121	Y
10	YP_325234.1	Beta-ketoacyl synthase [<i>Anabaena variabilis</i> ATCC 29413]	Bacteria	4.00E-121	
11	ZP_01424647.1	beta-ketoacyl synthase [<i>Herpetosiphon aurantiacus</i> ATCC 23779]	Bacteria	1.00E-120	
12	NP_867299.1	mycocerosate synthase [<i>Rhodopirellula baltica</i> SH 1]	Bacteria	3.00E-120	
13	YP_529192.1	Polyketide synthase modules and related proteins-like [<i>Saccharophagus degradans</i> 2-40]	Bacteria	1.00E-119	Y
14	NP_349947.1	Polyketide synthase pksE (short-chain alcohol dehydrogenase,acyl-carrier-protein S-malonyltransferase,3-oxoacyl-(acyl-carrier-protein) synthase I domains) [<i>Clostridium acetobutylicum</i> ATCC 824]	Bacteria	4.00E-119	Y
15	YP_324603.1	Beta-ketoacyl synthase [<i>Anabaena variabilis</i> ATCC 29413]	Bacteria	4.00E-119	
16	ZP_01628022.1	polyketide synthase [<i>Nodularia spumigena</i> CCY9414]	Bacteria	6.00E-118	Y
17	NP_486720.1	polyketide synthase [<i>Nostoc</i> sp. PCC 7120]	Bacteria	4.00E-116	Y
18	ABK32263.1	AmbH [<i>Polyangium cellulosum</i>]	Bacteria	4.00E-115	Y
19	ABM21570.1	CrpB [<i>Nostoc</i> sp. ATCC 53789]	Bacteria	4.00E-115	Y
20	AAT70104.1	CurI [<i>Lyngbya majuscula</i>]	Bacteria	6.00E-115	Y
21	DAA05847.1	TPA_inf: polyketide synthase [<i>Lytechinus variegatus</i>]	Metazoa	1.00E-114	Y
22	ZP_00107886.1	COG3321: Polyketide synthase modules and related proteins [<i>Nostoc punctiforme</i> PCC 73102]	Bacteria	2.00E-114	Y
23	AAT70107.1	CurL [<i>Lyngbya majuscula</i>]	Bacteria	6.00E-114	Y
24	ZP_01088963.1	polyketide synthase [<i>Blastopirellula marina</i> DSM 3645]	Bacteria	8.00E-114	Y
25	ZP_00110105.1	COG3321: Polyketide synthase modules and related proteins [<i>Nostoc punctiforme</i> PCC 73102]	Bacteria	1.00E-113	Y
26	YP_632255.1	polyketide synthase type I [<i>Myxococcus xanthus</i> DK 1622]	Bacteria	3.00E-113	Y
27	ZP_01462080.1	aminotransferase, class III family [<i>Stigmatella aurantiaca</i> DW4/3-1]	Bacteria	5.00E-113	
28	AAK57187.1	MxaC [<i>Stigmatella aurantiaca</i>]	Bacteria	5.00E-113	Y

29	ZP_01466498.1	cyclopropane-fatty-acyl-phospholipid synthase family [<i>Stigmatella aurantiaca</i> DW4/3-1]	Bacteria	6.00E-113	
30	AAF19814.1	MtaF [<i>Stigmatella aurantiaca</i>]	Bacteria	6.00E-113	Y
31	ZP_00514739.1	Beta-ketoacyl synthase:Acyl transferase region:Zinc-containing alcohol dehydrogenase superfamily:Phosphopantetheine-binding [<i>Crocospaera watsonii</i> WH 8501]	Bacteria	1.00E-112	
32	ZP_01728761.1	Short-chain dehydrogenase/reductase SDR [<i>Cyanothece</i> sp. CCY0110]	Bacteria	2.00E-112	
33	ZP_01423839.1	beta-ketoacyl synthase [<i>Herpetosiphon aurantiacus</i> ATCC 23779]	Bacteria	2.00E-112	
34	ZP_01425084.1	beta-ketoacyl synthase [<i>Herpetosiphon aurantiacus</i> ATCC 23779]	Bacteria	2.00E-112	
35	XP_001258284.1	polyketide synthase, putative [<i>Neosartorya fischeri</i> NRRL 181]	Fungi	3.00E-112	Y
36	XP_793564.2	PREDICTED: hypothetical protein [<i>Strongylocentrotus purpuratus</i>]	Metazoa	3.00E-112	
37	AAS98783.1	JamL [<i>Lyngbya majuscula</i>]	Bacteria	4.00E-112	Y
38	CAD19087.1	StiC protein [<i>Stigmatella aurantiaca</i>]	Bacteria	4.00E-112	Y
39	YP_951124.1	beta-ketoacyl synthase [<i>Mycobacterium vanbaalenii</i> PYR-1]	Bacteria	1.00E-111	
40	YP_325237.1	Beta-ketoacyl synthase [<i>Anabaena variabilis</i> ATCC 29413]	Bacteria	2.00E-111	
41	YP_632696.1	polyketide synthase [<i>Myxococcus xanthus</i> DK 1622]	Bacteria	3.00E-111	Y
42	AAF00958.1	McyE [<i>Microcystis aeruginosa</i> PCC 7806]	Bacteria	4.00E-111	Y
43	ZP_01629641.1	Amino acid adenylation [<i>Nodularia spumigena</i> CCY9414]	Bacteria	5.00E-111	
44	YP_001104811.1	polyketide synthase type I [<i>Saccharopolyspora erythraea</i> NRRL 2338]	Bacteria	7.00E-111	Y
45	YP_120537.1	putative polyketide synthase [<i>Nocardia farcinica</i> IFM 10152]	Bacteria	8.00E-111	Y
46	XP_001209753.1	hypothetical protein ATEG_07067 [<i>Aspergillus terreus</i> NIH2624]	Fungi	1.00E-110	Y
47	YP_632467.1	polyketide synthase [<i>Myxococcus xanthus</i> DK 1622]	Bacteria	1.00E-110	Y
48	XP_661216.1	hypothetical protein AN3612.2 [<i>Aspergillus nidulans</i> FGSC A4]	Fungi	3.00E-110	
49	BAB69235.1	modular polyketide synthase [<i>Streptomyces avermitilis</i>]	Bacteria	3.00E-110	Y
50	BAB12211.1	polyketide synthase and peptide synthetase [<i>Microcystis aeruginosa</i>]	Bacteria	3.00E-110	Y
51	YP_632697.1	polyketide synthase type I [<i>Myxococcus xanthus</i> DK 1622]	Bacteria	4.00E-110	Y
52	NP_828360.1	modular polyketide synthase [<i>Streptomyces avermitilis</i> MA-4680]	Bacteria	4.00E-110	Y
53	ZP_00107887.1	COG3321: Polyketide synthase modules and related proteins [<i>Nostoc punctiforme</i> PCC 73102]	Bacteria	5.00E-110	Y
54	CAD19089.1	StiE protein [<i>Stigmatella aurantiaca</i>]	Bacteria	6.00E-110	Y
55	YP_001068538.1	beta-ketoacyl synthase [<i>Mycobacterium</i> sp. JLS]	Bacteria	7.00E-110	
56	AAF19810.1	MtaB [<i>Stigmatella aurantiaca</i>]	Bacteria	7.00E-110	Y
57	ZP_01467359.1	hypothetical protein STIAU_0273 [<i>Stigmatella aurantiaca</i> DW4/3-1]	Bacteria	8.00E-110	
58	YP_637421.1	beta-ketoacyl synthase [<i>Mycobacterium</i> sp. MCS]	Bacteria	1.00E-109	

59	AAO64407.1	NdaF [<i>Nodularia spumigena</i>]	Bacteria	1.00E-109	Y
60	AAF26921.1	polyketide synthase [<i>Polyangium cellulosum</i>]	Bacteria	1.00E-109	Y
61	NP_485688.1	hypothetical protein all1648 [<i>Nostoc</i> sp. PCC 7120]	Bacteria	2.00E-109	Y
62	ZP_00901438.1	putative type I polyketide synthase [<i>Pseudomonas putida</i> F1]	Bacteria	3.00E-109	Y
63	YP_324486.1	Short-chain dehydrogenase/reductase SDR [<i>Anabaena variabilis</i> ATCC 29413]	Bacteria	4.00E-109	
64	XP_001275051.1	polyketide synthase, putative [<i>Aspergillus clavatus</i> NRRL 1]	Fungi	5.00E-109	Y
65	CAD29794.1	peptide synthetase [<i>Planktothrix agardhii</i> NIVA-CYA 126/8]	Bacteria	8.00E-109	
66	AAK57190.1	MxaF [<i>Stigmatella aurantiaca</i>]	Bacteria	1.00E-108	Y
67	AAK57188.1	MxD [<i>Stigmatella aurantiaca</i>]	Bacteria	1.00E-108	Y
68	AAF62883.1	epoD [<i>Polyangium cellulosum</i>]	Bacteria	1.00E-108	Y
69	AAF19813.1	MtaE [<i>Stigmatella aurantiaca</i>]	Bacteria	2.00E-108	Y
70	YP_324485.1	Short-chain dehydrogenase/reductase SDR [<i>Anabaena variabilis</i> ATCC 29413]	Bacteria	2.00E-108	
71	ABB92693.1	EpoD [synthetic construct]	other sequences	3.00E-108	Y
72	ABA29781.1	MmxC [<i>Cystobacter fuscus</i>]	Bacteria	4.00E-108	
73	AAR90269.1	polyketide synthase [<i>Cochliobolus heterostrophus</i>]	Fungi	4.00E-108	Y
74	YP_953934.1	beta-ketoacyl synthase [<i>Mycobacterium vanbaalenii</i> PYR-1]	Bacteria	5.00E-108	
75	CAJ46689.1	polyketide synthase [<i>Chondromyces crocatus</i>]	Bacteria	5.00E-108	Y
76	CAD89776.1	MelE protein [<i>Melittangium lichenicola</i>]	Bacteria	6.00E-108	Y
77	AAT70108.1	CurM [<i>Lyngbya majuscula</i>]	Bacteria	1.00E-107	Y
78	ZP_01463114.1	oxidoreductase, short chain dehydrogenase/reductase family [<i>Stigmatella aurantiaca</i> DW4/3-1]	Bacteria	1.00E-107	
79	XP_001262991.1	polyketide synthase, putative [<i>Neosartorya fischeri</i> NRRL 181]	Fungi	2.00E-107	Y
80	ZP_01466838.1	oxidoreductase, short chain dehydrogenase/reductase family [<i>Stigmatella aurantiaca</i> DW4/3-1]	Bacteria	3.00E-107	
81	YP_723339.1	beta-ketoacyl synthase [<i>Trichodesmium erythraeum</i> IMS101]	Bacteria	3.00E-107	
82	ZP_01547720.1	fatty acid synthase transmembrane protein [<i>Stappia aggregata</i> IAM 12614]	Bacteria	3.00E-107	
83	ZP_00110106.1	COG3321: Polyketide synthase modules and related proteins [<i>Nostoc punctiforme</i> PCC 73102]	Bacteria	4.00E-107	Y
84	AAAY89049.1	polyketide synthase [<i>Polyangium cellulosum</i>]	Bacteria	4.00E-107	Y
85	XP_754155.1	polyketide synthase type I, putative [<i>Aspergillus fumigatus</i> Af293]	Fungi	5.00E-107	Y
86	YP_388101.1	Amino acid adenylation [<i>Desulfovibrio desulfuricans</i> G20]	Bacteria	5.00E-107	
87	AAAY00025.1	SA1_PKSA [uncultured bacterial symbiont of <i>Discodermia dissoluta</i>]	Bacteria	8.00E-107	Y
88	YP_434162.1	Polyketide synthase modules and related protein [<i>Hahella chejuensis</i> KCTC 2396]	Bacteria	9.00E-107	Y
89	CAD19092.1	StiH protein [<i>Stigmatella aurantiaca</i>]	Bacteria	9.00E-107	Y
90	NP_841435.1	putative type I polyketide synthase WcbR [<i>Nitrosomonas europaea</i> ATCC 19718]	Bacteria	9.00E-107	Y
91	YP_884821.1	type I modular polyketide synthase [<i>Mycobacterium</i>]	Bacteria	1.00E-106	Y

		<i>smegmatis</i> str. MC2 155]			
92	BAE66270.1	unnamed protein product [<i>Aspergillus oryzae</i>]	Fungi	1.00E-106	
93	EAT79359.1	hypothetical protein SNOG_13032 [<i>Phaeosphaeria nodorum</i> SN15]	Fungi	2.00E-106	
94	CAJ46690.1	polyketide synthase [<i>Chondromyces crocatus</i>]	Bacteria	3.00E-106	Y
95	AAK57189.1	MxaE [<i>Stigmatella aurantiaca</i>]	Bacteria	3.00E-106	Y
96	YP_639998.1	beta-ketoacyl synthase [<i>Mycobacterium</i> sp. MCS]	Bacteria	3.00E-106	
97	XP_001273762.1	polyketide synthase, putative [<i>Aspergillus clavatus</i> NRRL 1]	Fungi	4.00E-106	Y
98	AAR90267.1	polyketide synthase [<i>Cochliobolus heterostrophus</i>]	Fungi	6.00E-106	Y
99	AAF00959.1	McyD [<i>Microcystis aeruginosa</i> PCC 7806]	Bacteria	8.00E-106	Y
100	AAO62584.1	polyketide synthase type 1 [<i>Anabaena</i> sp. 90]	Bacteria	8.00E-106	Y

The *M. globosa* putative polyketide synthase (PKS) gene (MGL_1024) was searched against the NCBI nr database, and the top 100 hits are summarized below. In total, 86 of the hits are from bacteria, 11 from fungi, and two from animals. The remaining hit was synthetic. Most (68) of the hits contain either polyketide synthase or PKS in the annotation description. Every hit contains COG3321 (PKS modules and related proteins). Only one hit is annotated as a fatty acid synthase (FAS), that of fatty acid synthase transmembrane protein (hit #82; ZP_01547720.1) from the bacterium *Stappia aggregata*. We have found no experimental evidence to support this protein as a fatty acid synthase (FAS).

The best matches are at the N-terminal region of the PKS, containing the KS and MAT domains. The ER and KR domains show the next best matches. The DH and ACP domains show little homology with other proteins. There is only one hit from the top 100 (hit# 35; XP_001258284.1 from the fungus *Neosartorya fischeri*) that shows similarity at the C-terminal thioester-reductase domain.

It is interesting that there are also two animal hits in the top 100 matches to *M. globosa* putative PKS. Both hits are from sea urchin (*Lytechinus variegatus* [green sea urchin, tpg|DAA05847.1|, hit#21] and *Strongylocentrotus purpuratus* [purple sea urchin,

ref[XP_793564.2, hit#36])). The hit from *S. purpuratus* was shown experimentally to be a PKS (38).

In contrast, when using each of the two *U. maydis* putative FAS genes to search against the NCBI nr database, the top 100 hits are all annotated as fungal FAS or hypothetical proteins (not shown).

SI Table 7. Top blast hits from a search using the C-terminal domain of *M.*

globosa putative PKS

Rank	Accession#	Description	Species	Classification	E-value
1	NP_489397.1	polyketide synthase [<i>Nostoc</i> sp. PCC 7120]	<i>Nostoc</i> sp. PCC 7120	Bacteria	5.00E-36
2	AAK57184.1	MxaA [<i>Stigmatella aurantiaca</i>]	<i>Stigmatella aurantiaca</i>	Bacteria	2.00E-35
3	YP_323105.1	Thioester reductase [<i>Anabaena variabilis</i> ATCC 29413]	<i>Anabaena variabilis</i> ATCC 29413	Bacteria	2.00E-35
4	AAA03658.1	putative polyketide synthase	<i>Anabaena</i> sp.	Bacteria	3.00E-35
5	YP_632693.1	non-ribosomal peptide synthase MxaA [<i>Myxococcus xanthus</i> DK 1622]	<i>Myxococcus xanthus</i> DK 1622	Bacteria	4.00E-34
6	ZP_01628049.1	polyketide synthase [<i>Nodularia spumigena</i> CCY9414]	<i>Nodularia spumigena</i> CCY9414	Bacteria	5.00E-34
7	Q70LM4	Linear gramicidin synthetase subunit D [Includes: ATP-dependent tryptophan adenylase (TrpA) (Tryptophan activase); ATP-dependent D-leucine adenylase (D-LeuA) (D-Leucine activase); Leucine racemase [ATP-hydrolyzing] ; ATP-dependent tryptophan adenylase (TrpA) (Tryptophan activase); ATP-dependent glycine adenylase (GlyA) (Glycine activase); Linear gramicidin--PCP reductase]	<i>Brevibacillus parabrevis</i>	Bacteria	7.00E-32
8	ZP_00107100.1	COG3320: Putative dehydrogenase domain of multifunctional non-ribosomal peptide synthetases and related enzymes [<i>Nostoc punctiforme</i> PCC 73102]	<i>Nostoc punctiforme</i> PCC 73102	Bacteria	2.00E-31
9	AAR21587.1	Alpha-aminoadipate reductase [<i>Cryptococcus neoformans</i> var. <i>neoformans</i>]	<i>Cryptococcus neoformans</i> var. <i>neoformans</i>	Fungi	4.00E-31
10	XP_965396.1	probable alpha-aminoadipate reductase large subunit [MIPS] [<i>Neurospora crassa</i> OR74A]	<i>Neurospora crassa</i> OR74A	Fungi	4.00E-31
11	XP_571745.1	aminoadipate-semialdehyde dehydrogenase [<i>Cryptococcus neoformans</i> var. <i>neoformans</i> JEC21]	<i>Cryptococcus neoformans</i> var. <i>neoformans</i> JEC21	Fungi	5.00E-31
12	ZP_01630202.1	polyketide synthase [<i>Nodularia spumigena</i> CCY9414]	<i>Nodularia spumigena</i> CCY9414	Bacteria	6.00E-31
13	NP_486681.1	polyketide synthase [<i>Nostoc</i> sp. PCC 7120]	<i>Nostoc</i> sp. PCC 7120	Bacteria	7.00E-31
14	ZP_01632060.1	Non-ribosomal peptide synthase [<i>Nodularia spumigena</i> CCY9414]	<i>Nodularia spumigena</i> CCY9414	Bacteria	1.00E-30
15	XP_503627.1	hypothetical protein [<i>Yarrowia lipolytica</i>]	<i>Yarrowia lipolytica</i> CLIB122	Fungi	2.00E-30
16	XP_001220511.1	conserved hypothetical protein [<i>Chaetomium globosum</i> CBS 148.51]	<i>Chaetomium globosum</i> CBS 148.51	Fungi	2.00E-30
17	ZP_00514134.1	Thioester reductase [<i>Crocospaera watsonii</i> WH 8501]	<i>Crocospaera watsonii</i> WH 8501	Bacteria	5.00E-30
18	ZP_00516263.1	Non-ribosomal peptide synthase:Amino acid adenylation:Thioester reductase [<i>Crocospaera watsonii</i> WH 8501]	<i>Crocospaera watsonii</i> WH 8501	Bacteria	5.00E-30
19	AAO23334.1	NcpB [<i>Nostoc</i> sp. ATCC 53789]	<i>Nostoc</i> sp. ATCC 53789	Bacteria	7.00E-30
20	ZP_01632190.1	probable non-ribosomal peptide synthetase [<i>Nodularia spumigena</i> CCY9414]	<i>Nodularia spumigena</i> CCY9414	Bacteria	1.00E-29
21	XP_721197.1	alpha-aminoadipate reductase [<i>Candida albicans</i> SC5314]	<i>Candida albicans</i> SC5314	Fungi	2.00E-27
22	YP_729393.1	polyketide synthase, putative [<i>Synechococcus</i> sp. CC9311]	<i>Synechococcus</i> sp. CC9311	Bacteria	2.00E-27
23	EAQ70925.1	hypothetical protein MGCH7_ch7g332 [<i>Magnaporthe grisea</i> 70-15]	<i>Magnaporthe grisea</i> 70-15	Fungi	3.00E-27
24	XP_721469.1	alpha-aminoadipate reductase [<i>Candida albicans</i> SC5314]	<i>Candida albicans</i> SC5314	Fungi	3.00E-27
25	XP_001266925.1	alpha-aminoadipate reductase large subunit, putative [<i>Neosartorya fischeri</i> NRRL 1831]	<i>Neosartorya fischeri</i> NRRL 1831	Fungi	4.00E-27

		putative [<i>Neosartorya fischeri</i> NRRL 181]	NRRL 181		
26	XP_366535.1	hypothetical protein MG02611.4 [<i>Magnaporthe grisea</i> 70-15]	<i>Magnaporthe grisea</i> 70-15	Fungi	4.00E-27
27	ABO15888.1	polyketide synthase [<i>Streptomyces</i> sp. UC 11065]	<i>Streptomyces</i> sp. UC 11065	Bacteria	5.00E-27
28	XP_001214153.1	predicted protein [<i>Aspergillus terreus</i> NIH2624]	<i>Aspergillus terreus</i> NIH2624	Fungi	7.00E-27
29	ZP_01647305.1	amino acid adenylation domain [<i>Salinispora arenicola</i> CNS205]	<i>Salinispora arenicola</i> CNS205	Bacteria	2.00E-26
30	AAC44129.1	saframycin Mx1 synthetase A	<i>Myxococcus xanthus</i>	Bacteria	2.00E-26
31	AAY42631.1	HetM [<i>Nodularia spumigena</i>]	<i>Nodularia spumigena</i>	Bacteria	4.00E-26
32	XP_001271987.1	alpha-aminoadipate reductase large subunit, putative [<i>Aspergillus clavatus</i> NRRL 1]	<i>Aspergillus clavatus</i> NRRL 1	Fungi	4.00E-26
33	CAC22111.1	aminoadipate reductase enzyme [<i>Acremonium chrysogenum</i>]	<i>Acremonium chrysogenum</i>	Fungi	4.00E-26
34	XP_751705.1	alpha-aminoadipate reductase large subunit, putative [<i>Aspergillus fumigatus</i> Af293]	<i>Aspergillus fumigatus</i> Af293	Fungi	5.00E-26
35	XP_360747.1	hypothetical protein MG03290.4 [<i>Magnaporthe grisea</i> 70-15]	<i>Magnaporthe grisea</i> 70-15	Fungi	5.00E-26
36	NP_594314.1	hypothetical protein SPAP7G5.04c [<i>Schizosaccharomyces pombe</i> 972h-]	<i>Schizosaccharomyces pombe</i> 972h-	Fungi	6.00E-26
37	XP_451939.1	unnamed protein product [<i>Kluyveromyces lactis</i>]	<i>Kluyveromyces lactis</i> NRRL Y-1140	Fungi	6.00E-26
38	XP_663214.1	hypothetical protein AN5610.2 [<i>Aspergillus nidulans</i> FGSC A4]	<i>Aspergillus nidulans</i> FGSC A4	Fungi	8.00E-26
39	ZP_01430275.1	Amino acid adenylation [<i>Salinispora tropica</i> CNB-440]	<i>Salinispora tropica</i> CNB-440	Bacteria	1.00E-25
40	CAD43185.1	alpha-aminoadipate reductase [<i>Kluyveromyces lactis</i>]	<i>Kluyveromyces lactis</i>	Fungi	1.00E-25
41	AAX11423.1	nonribosomal peptide synthetase 10 [<i>Gibberella moniliformis</i>]	<i>Gibberella moniliformis</i>	Fungi	1.00E-25
42	CAB97252.1	alpha-aminoadipate reductase [<i>Pichia farinosa</i>]	<i>Pichia farinosa</i>	Fungi	2.00E-25
43	ABO15860.1	polyketide synthase [<i>Streptomyces vitaminophilus</i>]	<i>Streptomyces vitaminophilus</i>	Bacteria	2.00E-25
44	ZP_00109915.1	COG0318: Acyl-CoA synthetases (AMP-forming)/AMP-acid ligases II [<i>Nostoc punctiforme</i> PCC 73102]	<i>Nostoc punctiforme</i> PCC 73102	Bacteria	3.00E-25
45	XP_386683.1	hypothetical protein FG06507.1 [<i>Gibberella zeae</i> PH-1]	<i>Gibberella zeae</i> PH-1	Fungi	4.00E-25
46	ZP_00828169.1	COG1020: Non-ribosomal peptide synthetase modules and related proteins [<i>Yersinia frederiksenii</i> ATCC 33641]	<i>Yersinia frederiksenii</i> ATCC 33641	Bacteria	4.00E-25
47	XP_001213135.1	L-aminoadipate-semialdehyde dehydrogenase large subunit [<i>Aspergillus terreus</i> NIH2624]	<i>Aspergillus terreus</i> NIH2624	Fungi	6.00E-25
48	BAE58192.1	unnamed protein product [<i>Aspergillus oryzae</i>]	<i>Aspergillus oryzae</i>	Fungi	6.00E-25
49	XP_386217.1	hypothetical protein FG06041.1 [<i>Gibberella zeae</i> PH-1]	<i>Gibberella zeae</i> PH-1	Fungi	7.00E-25
50	CAK38804.1	unnamed protein product [<i>Aspergillus niger</i>]	<i>Aspergillus niger</i>	Fungi	8.00E-25

The C terminus of *M. globosa* putative PKS matches TIGRFAM profile TIGR01746 (thioester reductase domain) that is constructed from the terminal domain of the fungal α -aminoadipate reductase involved in lysine biosynthesis and MxcG which functions as a reductase in myxochelin biosynthesis. If part of a larger protein, the thioester reductase domain is always found in the C terminus. The domain is commonly found in PKS's and

non-ribosomal peptide synthases. The function of the domain is to release the product by reducing the thioester to an aldehyde. In contrast, the thioesterase domain found in FAS releases the fatty acid by catalyzing the hydrolysis of the thioester linking the fatty acid to phosphopantetheine. The *M. globosa* putative PKS thioester reductase domain also matches the COG3320 that is annotated as “putative dehydrogenase domain of multifunctional non-ribosomal peptide synthetases and related enzymes.”

We searched the NCBI nr database using the C-terminal 353 residues of the *M. globosa* putative PKS. The top matches are annotated as polyketide synthase, non-ribosomal polyketide synthase, thioester reductase, and α -aminoadipate reductase. Shown below are the top 50 hits. None of the top hits is annotated as a fatty acid synthase.

SI Table 8. Fatty acid synthases found in sequenced fungal genomes

Genomes	Beta subunit protein accession #	Alpha subunit protein accession #
<i>Aspergillus fumigatus</i> Af293	XP_748739.1	XP_748738.1
<i>Aspergillus nidulans</i> FGSC A4	XP_682677.1	XP_682676.1
<i>Aspergillus niger</i> CBS 513.88	CAK45758.1	CAK36857.1
<i>Aspergillus terreus</i> NIH2624	XP_001213436.1	XP_001213437.1
<i>Aspergillus oryzae</i> RIB40	BAE62632	BAE62631
<i>Candida albicans</i> SC5314	XP_716877.1	XP_723162.1 XP_723161.1*
<i>Candida glabrata</i> CBS 138	XP_445436.1	XP_445956.1
<i>Coprinopsis cinerea</i> okayama 7#130	EAU85370.1 [†]	
<i>Cryptococcus neoformans</i> var. <i>neoformans</i> B-3501A	EAL20517.1	EAL20516.1
<i>Cryptococcus neoformans</i> var. <i>neoformans</i> JEC21	XP_571100.1	XP_571099.1
<i>Debaryomyces hansenii</i> CBS767	XP_457311.1	XP_457388.1
<i>Eremothecium gossypii</i>	NP_984945.1	NP_985412.1
<i>Gibberella zeae</i> PH-1	XP_385498.1	XP_385497.1
<i>Kluyveromyces lactis</i> NRRL Y-1140	XP_451653.1	XP_452914.1
<i>Neurospora crassa</i>	XP_962465.1	XP_962466.1
<i>Saccharomyces cerevisiae</i>	NP_012739.1	NP_015093.1
<i>Schizosaccharomyces pombe</i> 972h-	NP_594370.1	NP_593823.1
<i>Ustilago maydis</i> 521	XP_759118.1 [†]	
<i>Yarrowia lipolytica</i> CLIB122	XP_500912.1	XP_501096.1
	Beta subunit contig accession #	Alpha subunit contig accession #
<i>Ajellomyces capsulatus</i> NAm1	AAJI01002272.1	AAJI01002272.1
<i>Ascosphaera apis</i> USDA-ARSEF 7405	AARE01000101.1	AARE01001140.1
<i>Aspergillus clavatus</i> NRRL 1	NZ_AAKD03000012.1	NZ_AAKD03000012.1
<i>Aspergillus flavus</i> NRRL3357	AAIH01000282.1	AAIH01000282.1
<i>Aspergillus terreus</i> ATCC 20542	AABT01000081.1 [‡]	
<i>Botryotinia fuckeliana</i> B05.10	AAID01003361.1	AAID01003361.1
<i>Candida albicans</i> WO-1	AAFO01000026.1	AAFO01000057.1
<i>Candida tropicalis</i> MYA-3404	AAFN01000095.1	AAFN01000139.1
<i>Chaetomium globosum</i> CBS 148.51	NZ_AAFU01000788.1	NZ_AAFU01000788.1
<i>Clavisporea lusitaniae</i> ATCC 42720	AAFT01000036.1	AAFT01000002.1
<i>Coccidioides immitis</i> H538.4	AASO01002422.1	AASO01002426.1
<i>Coccidioides immitis</i> RMSCC 2394	NZ_AATX01000239.1	NZ_AATX01000239.1
<i>Coccidioides immitis</i> RS	NZ_AAEC02000038.1	NZ_AAEC02000038.1
<i>Cryptococcus neoformans</i> R265	AAFP01000549.1	AAFP01000549.1
<i>Cryptococcus neoformans</i> var. <i>grubii</i> H99	AACO02000043.1	AACO02000043.1
<i>Gibberella moniliformis</i> 7600	AAIM02000056.1	AAIM02000056.1
<i>Kluyveromyces waltii</i> NCYC 2644	AADM01000290.1	AADM01000010.1
<i>Lodderomyces elongisporus</i> NRRL YB-4239	AAPO01000050.1	AAPO01000074.1

<i>Magnaporthe grisea</i> 70-15	AACU02000456.1	AACU02000456.1
<i>Neosartorya fischeri</i> NRRL 181	NZ_AAKE03000046.1	NZ_AAKE03000046.1
<i>Phaeosphaeria nodorum</i> SN15	AAGI01000045.1	AAGI01000045.1
<i>Phanerochaete chrysosporium</i> RP-78	AADS01000039.1	AADS01000039.1
<i>Pichia guilliermondii</i> ATCC 6260	AAFM01000041.1	AAFM01000041.1
<i>Rhizopus oryzae</i> RA 99-880	AACW02000023.1	AACW02000023.1
<i>Puccinia graminis</i> f. sp. <i>tritici</i> CRL 75-36-700-3	AAWC01001139.1	AAWC01001275.1
<i>Saccharomyces bayanus</i> 623-6C	AACG02000173.1	AACG02000023.1
<i>Saccharomyces bayanus</i> MCYC 623	AACA01000289.1	AACA01000045.1
<i>Saccharomyces castellii</i> NRRL Y-12630	AACF01000400.1	AACF01000003.1
<i>Saccharomyces cerevisiae</i> RM11-1a	AAEG01000033.1	AAEG01000080.1
<i>Saccharomyces cerevisiae</i> YJM789	AAFW01000148.1	AAFW01000073.1
<i>Saccharomyces kluyveri</i> NRRL Y-12651	AACE02000141.1	AACE02000179.1
<i>Saccharomyces kudriavzevii</i> IFO 1802	AACI02001627.1	AACI02000597.1
<i>Saccharomyces mikatae</i> IFO 1815	AABZ01000096.1	AABZ01000004.1
<i>Saccharomyces paradoxus</i> NRRL Y-17217	AABY01000006.1	AABY01000030.1
<i>Schizosaccharomyces japonicus</i> yFS275	AATM01000137.1	AATM01000109.1
<i>Sclerotinia sclerotiorum</i> 1980	AAGT01000601.1	AAGT01000601.1
<i>Trichoderma reesei</i> QM9414	AAIL01000854.1	AAIL01000854.1
<i>Uncinocarpus reesii</i> 1704	AAIW01000176.1	AAIW01000176.1

The sequences of fatty acid synthase (FAS) subunits of *S. cerevisiae* (Accession XP_962465.1 and NP_015093.1) were searched against 55 fungal genomes available for blast search at NCBI. Blastp was used to search against 18 of the genomes with protein sequence available, and tblastn was used to search against the genome sequence of the remaining 37 genomes. The results were manually inspected, and putative FAS accession numbers or the chromosome/contig accession numbers were listed below. Multiple putative FASs were found in some genomes, and only the best matched accession number was reported. Putative FAS's were found in all free-living fungi but not in the Microsporidium *Encephalitozoon cuniculi* GB-M1 (34), which has only

recently been classified as a fungus. From our limited sequence of the *M. restricta* genome, there was no sign of a traditional fungal FAS. Based on the predicted size of its genome (9 Mb) and the size of the *U. maydis* FAS gene (~11 kb), there is $< 10^{-9}$ chance that we would have not sequenced this gene among the 17,914 sequencing reactions. *M. restricta* does encode a gene similar to the *M. globosa* putative PKS (SI Table 6).

*XP_723162.1 and XP_723161.1 are the accession numbers for the N-terminal and C-terminal beta subunits, respectively.

†The two subunits are fused into a single polypeptide.

‡Only the sequence of beta subunit was found in the genome. Of the estimated 35 Mb genome, 138,520 bases (0.4%) were sequenced. Therefore, the absence of an alpha subunit gene is likely due to low sequence coverage. Note that genes for both subunits were found in a different *A. terreus* strain, listed in the upper part of this table.

SI Table 9. Secreted and cell wall-associated proteins from *M. globosa*

Protein		GeneID	Predicted to be secreted ⁺	Secreted into the media [‡]			Cell wall-associated proteins [‡]								Homologs identified from <i>M. restricta</i> cultures	
				Rich		Defined	Rich				Defined					
							EDTA		SDS		EDTA	SDS				
				family	Gene annotation ⁺	1	2	3	4	5	6	7	8	9	10	11
Aspartyl Proteases	secretory aspartyl protease	MGL_0534	X													
	secretory aspartyl protease	MGL_1932	X													
	secretory aspartyl protease	MGL_3328	X													
	secretory aspartyl protease	MGL_3331	X													
	secretory aspartyl protease	MGL_4053	X													
GMC Oxidoreductases	GMC oxidoreductase	MGL_0683	X													
	GMC oxidoreductase	MGL_0750	X													
	GMC oxidoreductase	MGL_0751	X													
	GMC oxidoreductase	MGL_3881	X													
Phospho- lipases	secretory phospholipase C	MGL_3076	X													
	secretory phospholipase C	MGL_3077	X													
	secretory phospholipase C	MGL_3326	X													
Lipases	LIP1, secretory lipase (family 3)	MGL_0797	X													
	secretory lipase (family 3)	MGL_0799	X													
	secretory lipase (family LIP)	MGL_1311	X													
	secretory lipase (family 3)	MGL_1331	X													
	secretory lipase (family LIP)	MGL_3507	X													
	secretory lipase (family LIP)	MGL_4051	X													
	secretory lipase (family LIP)	MGL_4054	X													
	secretory lipase (family 3)	MGL_4197	X													
Cell Wall Modifying Enzymes	chitin deacetylase	MGL_0611	X													
	1,3-beta-glucanase	MGL_1399	X													
	similar to <i>Cryp. neoformans</i> glycoprotein	MGL_1981	X													
	1,3-beta-glucosidase	MGL_3222	X													
	1,3-beta-glucosidase	MGL_3944	X													
	1,3-beta-glucosidase	MGL_3945	X													
Allergens	similar to allergen Mal s1 / Mal f1	MGL_1303	X													
	similar to allergen Mal s8	MGL_1304	X													
	similar to allergen Mal s9	MGL_2179	X													
	similar to allergen Mal s6 / cyclophilin A	MGL_3612	X													
	similar to allergen Mal f2	MGL_4042	X													
	Other	acid sphingomyelinase	MGL_1574	X												
ferro-O2-oxidoreductase precursor		MGL_4262	X													
Unknown Function	Unknown <i>M. globosa</i> protein	MGL_0157	X													
	Unknown <i>M. globosa</i> protein	MGL_0379	X													
	Unknown <i>M. globosa</i> protein	MGL_0433	X													
	Unknown <i>M. globosa</i> protein	MGL_0535	X													
	Unknown <i>M. globosa</i> protein	MGL_0536	X													
	Unknown <i>M. globosa</i> protein	MGL_0696	X													
	Unknown <i>M. globosa</i> protein	MGL_1187	X													
	Unknown <i>M. globosa</i> protein	MGL_1316	X													
	Unknown <i>M. globosa</i> protein	MGL_1576	X													
	Unknown <i>M. globosa</i> protein	MGL_1684	X													
	Unknown <i>M. globosa</i> protein	MGL_1732	X													
	Unknown <i>M. globosa</i> protein	MGL_2131	X													
	Unknown <i>M. globosa</i> protein	MGL_2131	X													
	Unknown <i>M. globosa</i> protein	MGL_2660	X													
	Unknown <i>M. globosa</i> protein	MGL_2668	X													
	Unknown <i>M. globosa</i> protein	MGL_2673	X													
	Unknown <i>M. globosa</i> protein	MGL_2673	X													
	Unknown <i>M. globosa</i> protein	MGL_2942	X													
	Unknown <i>M. globosa</i> protein	MGL_3153	X													
	Unknown <i>M. globosa</i> protein	MGL_3210	X													
	Unknown <i>M. globosa</i> protein	MGL_3384	X													
	Unknown <i>M. globosa</i> protein	MGL_3426	X													
	Unknown <i>M. globosa</i> protein	MGL_3792	X													
	Unknown <i>M. globosa</i> protein	MGL_4104	X													
	Unknown <i>M. globosa</i> protein	MGL_4191	X													
	Unknown <i>M. globosa</i> protein	MGL_4195	X													
	Unknown <i>M. globosa</i> protein	MGL_4242	X													
	Unknown <i>M. globosa</i> protein	MGL_4259	X													

For more details on this table, please refer to *Proteomics Methods* in Section I: *SI Text*.

* Annotation for LIP1 (MGL_0797) is based on experimental evidence (30). All other annotations are based on sequence similarity to proteins of known function.

† SignalP 3.0 was used to predict amino-terminal secretion signals, as described in *Methods*.

‡ Details for Experiments 1-11:

Experiment #	1	2	3	4	5	6	7	8	9	10	11
Media type	Rich	Rich	Defined	Rich	Rich	Rich	Rich	Rich	Rich	Defined	Defined
Extraction type	Pierce PagePrep	Pierce PagePrep	Pierce PagePrep	EDTA	EDTA	EDTA	EDTA	EDTA	SDS	EDTA	SDS
Gel type	1D	1D	1D	2D	1D	2D	1D	1D	1D	1D	1D
MALDI	MALDI	MALDI	MALDI	MALDI	MALDI TOF/TOF	MALDI TOF/TOF	MALDI	MALDI	MALDI	MALDI	MALDI
Electrospray	QTOF	QTOF	QTOF	QTOF	QTOF Nano LC-MS/MS	QTOF Nano LC-MS/MS	QTOF	QTOF	QTOF	QTOF	QTOF
Edman									EDMAN		EDMAN

SI Table 10. Secreted and cell wall-associated proteins from *M. restricta*

Protein family	Gene annotation *	Predicted to be secreted [†]	MS Database that was matched [‡]		<i>M. restricta</i> Contig [§]	Corresponding <i>Malassezia</i> GeneID [§]
			<i>M. globosa</i> and NCBI nr	<i>M. restricta</i>		
Aspartyl Proteases	secretory aspartyl protease	x		x	MRE_CONTIG_0010	MGL_2125
GMC Oxidoreductases	GMC oxidoreductase	x	x			MGL_0683
	GMC oxidoreductase	x	x	x	MRE_CONTIG_0716	MGL_0750
	GMC oxidoreductase	x	x			MGL_0751
	Mal s 12 allergen precursor [<i>Malassezia sympodialis</i>]	x	x			gij58416210
Phospholipases	secretory phospholipase C	x		x	MRE_CONTIG_2150	MGL_2589
	secretory phospholipase C	x	x	x	MRE_CONTIG_2432	MGL_3075
	secretory phospholipase C	x	x	x	MRE_CONTIG_3947	MGL_3326
Lipases	LIP1, secretory lipase (family 3) [<i>M. globosa</i>]	x	x	x	MRE_CONTIG_3144	MGL_0797
Other	histidine acid phosphatase	x		x	MRE_CONTIG_1599	MGL_3567
Unknown Function	hypothetical protein			x	MRE_CONTIG_3687	MGL_0915
	hypothetical protein			x	MRE_CONTIG_1963	
	hypothetical protein			x	MRE_CONTIG_1913	MGL_1562

M. restricta cells grown in mDixon media (30) were pelleted by centrifugation at 4000 x g and washed three times with TE buffer [50 mM Tris-HCl (pH 8.0), 5 mM EDTA containing protease inhibitors (Roche)]. The cell pellet was resuspended in an equal volume of 5 M urea, 2 M thiourea, 4% CHAPS, 1% tributylphosphine and incubated for 1 hr at room temperature with gentle agitation. The cells were pelleted as above and discarded, and the proteins in the supernatant were prepared for 1D SDS PAGE. 1D SDS PAGE was performed using a 10-20% gel (Invitrogen), and the gel was stained with Coomassie blue as described (30,33). Details for the identification of proteins by MALDI also have been described (30,33).

* The annotation is based on sequence similarity to proteins of known function.

[†] SignalP 3.0 was used to predict N-terminal secretion signals, as described in *Methods*. The prediction was performed on *M. globosa* and *M. sympodialis* proteins but not on *M. restricta* proteins because so many of the corresponding *M. restricta* genes had incomplete sequence information.

[‡] Some of the *M. restricta* peptides that were identified by mass spectrometry have exact matches (indicated by an "x") in the *M. globosa* or NCBI nr databases but not in the *M. restricta* database, or vice versa. This most likely reflects the incomplete sequence coverage of the *M. restricta* genome as well as sequence differences between the homologous *M. globosa* and *M. restricta* proteins.

[§] Contig and GeneID assignments were derived from the mass spectrometry data and homology between the corresponding *M. globosa* and *M. restricta* proteins.

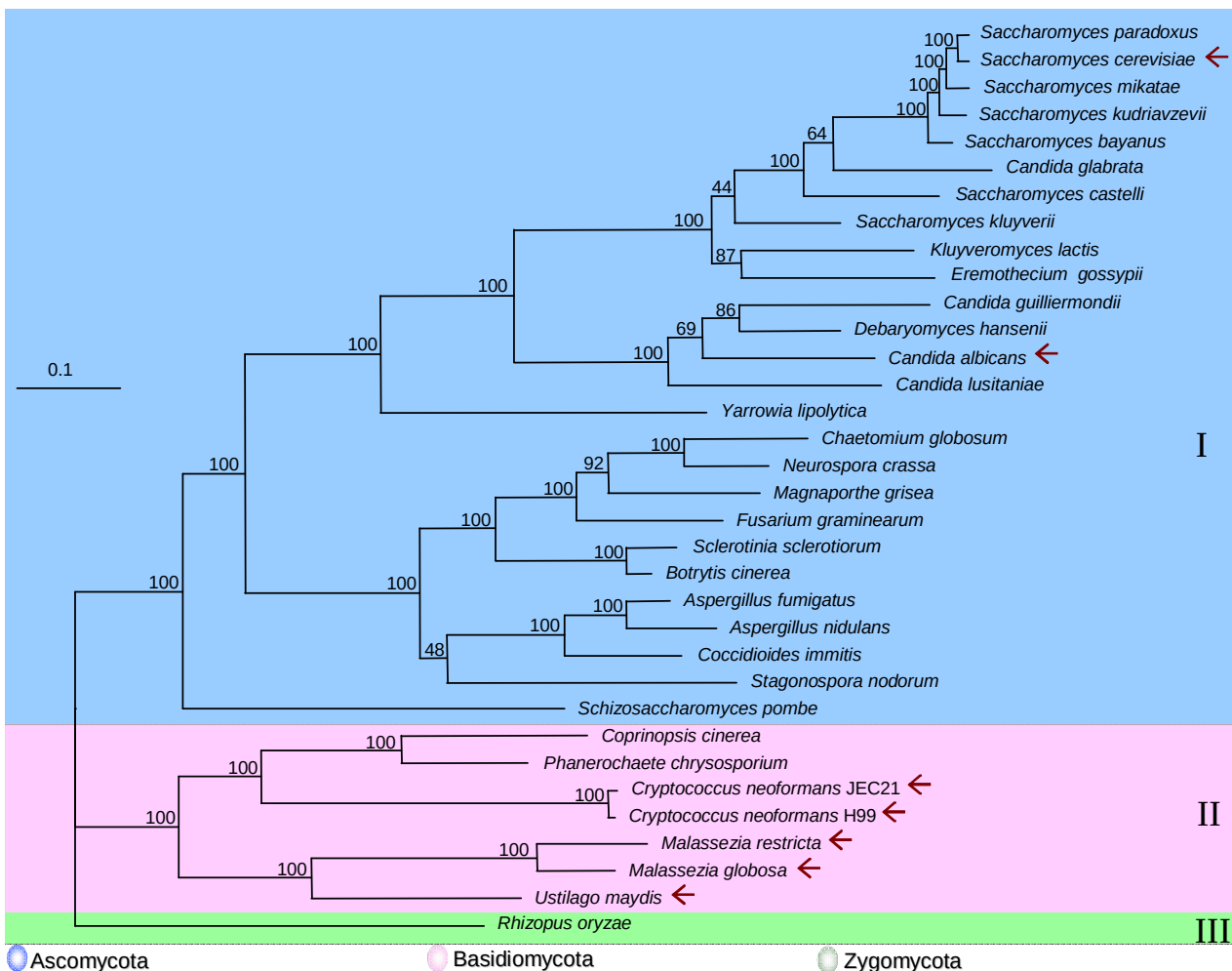


Fig. 4. Fungal phylogenetic tree based on 28 orthologous proteins present in 34 complete genomes. See SI Table 11 for a list of these proteins. The clades I, II and III correspond to the Ascomycota, Basidiomycota and Zygomycota phyla, respectively. The topology is congruent with the results based on different phylogenetic methods such as supertree (40), and concatenation of 531 proteins (41), 153 proteins (40) and six genes (42). The arrows indicate the genomes compared in Table 1.

SI Table 11. KOG number, functional category, and predicted function used for phylogenetic analysis of 34 complete fungal genomes

KOG number	Functional category	Predicted function
KOG0304	A	mRNA deadenylase subunit
KOG0816	A	Protein involved in mRNA turnover
KOG3013	A	Exosomal 3'-5' exoribonuclease complex, subunit Rrp4
KOG1534	AK	Putative transcription factor FET5
KOG1598	K	Transcription initiation factor TFIIIB, Brf1 subunit
KOG3104	K	Mod5 protein sorting/negative effector of RNA Pol III synthesis
KOG2626	KB	Histone H3 (Lys4) methyltransferase complex, subunit CPS60/ASH2/BRE2
KOG0991	L	Replication factor C, subunit RFC2
KOG2267	L	Eukaryotic-type DNA primase, large subunit
KOG2671	L	Putative RNA methylase
KOG2851	L	Eukaryotic-type DNA primase, catalytic (small) subunit
KOG0436	J	Methionyl-tRNA synthetase
KOG1750	J	Mitochondrial/chloroplast ribosomal protein S12
KOG3185	J	Translation initiation factor 6 (eIF-6)
KOG2874	JD	rRNA processing protein
KOG1173	DO	Anaphase-promoting complex (APC), Cdc16 subunit
KOG1746	DO	Defender against cell death protein/oligosaccharyltransferase, epsilon subunit
KOG3800	O	Predicted E3 ubiquitin ligase containing RING finger, subunit of transcription/repair factor TFIIH and CDK-activating kinase assembly factor
KOG0809	U	SNARE protein TLG2/Syntaxin 16
KOG3103	U	Rab GTPase interacting factor, Golgi membrane protein
KOG2788	C	Glycosyltransferase
KOG0325	CH	Lipoyltransferase
KOG3327	F	Thymidylate kinase/adenylate kinase
KOG2794	H	Delta-aminolevulinic acid dehydratase
KOG3789	P	Nitrogen permease regulator NLRG/NPR2
KOG0645	R	WD40 repeat protein
KOG2728	R	Uncharacterized conserved protein with similarity to phosphopantothenoylcysteine synthetase/decarboxylase
KOG0282	S	mRNA splicing factor

The categories are grouped in the four main categories: (i) Information Storage and Processing ([J] Translation, ribosomal structure and biogenesis, [A] RNA processing and modification, [K] Transcription, [L] Replication, recombination and repair, [B] Chromatin structure and dynamics); (ii) Cellular Processing and Signaling ([D] Cell cycle control, cell division, chromosome partitioning, [U] Intracellular trafficking, secretion,

and vesicular transport, [O] Posttranslational modification, protein turnover, chaperones); (iii) Metabolism ([C] Energy production and conversion, [F] Nucleotide transport and metabolism, [H] Coenzyme transport and metabolism, [P] Inorganic ion transport and metabolism, and (iv) Poorly Characterized ([R] General function prediction only, [S] Function unknown) (14).

SI Table 12. Genes reported in Fig. 2, found in clusters within the *M. globosa* genome

Cluster ID	Contig ID	Gene ID	Description
C1	mgl_sf_13.1	MGL_4048	hypothetical protein
		MGL_4049	hypothetical protein
		MGL_4050	hypothetical protein
		MGL_4051	similar to secretory lipase (family LIP)
		MGL_4052	similar to secretory lipase (family LIP)
		MGL_4053	similar to secretory aspartyl protease
		MGL_4054	similar to secretory lipase (family LIP)
		MGL_4055	hypothetical protein
C2	mgl_sf_8.1	MGL_3326	similar to secretory phospholipase c
		MGL_3327	similar to secretory aspartyl protease
		MGL_3328	similar to secretory aspartyl protease
		MGL_3329	similar to secretory aspartyl protease
		MGL_3330	similar to secretory aspartyl protease
		MGL_3331	similar to secretory aspartyl protease
		MGL_3332	hypothetical protein
		MGL_3333	hypothetical protein
C3	mgl_sf_8.1	MGL_3334	similar to secretory aspartyl protease
		MGL_3072	hypothetical protein
		MGL_3073	hypothetical protein
		MGL_3074	hypothetical protein
		MGL_3075	similar to secretory phospholipase C
		MGL_3076	similar to secretory phospholipase C
		MGL_3077	similar to secretory phospholipase C
C4	mgl_sf_1.2	MGL_3078	hypothetical protein
		MGL_0797	LIP1, secretory lipase (family 3)
		MGL_0798	similar to secretory lipase (family 3)
		MGL_0799	similar to secretory lipase (family 3)
C5	mgl_sf_3.1	MGL_0800	similar to secretory lipase (family 3)
		MGL_1572	hypothetical protein
		MGL_1573	similar to acid sphingomyelinase
		MGL_1574	similar to acid sphingomyelinase
		MGL_1575	similar to acid sphingomyelinase
		MGL_1576	hypothetical protein
C6	mgl_sf_10.1	MGL_1577	hypothetical protein
		MGL_3562	hypothetical protein
		MGL_3563	hypothetical protein
		MGL_3564	hypothetical protein
		MGL_3565	similar to secretory aspartyl protease
		MGL_3566	hypothetical protein
		MGL_3567	Similar to histidine acid phosphatase
		MGL_3568	similar to acid sphingomyelinase

The genes are listed in the order represented in Fig. 3, with the top gene in this table referring to the leftmost gene in Fig. 1C.

SI Table 13. Similarities of *C. albicans* aspartyl proteases to *M. globosa* aspartyl proteases

Query ID	Hit ID	Pct ID
orf19.1891	MGL_3195	54.42
orf19.2082	MGL_3329	22.04
orf19.242	MGL_3195	27.71
orf19.3708	MGL_3195	28.13
orf19.3839	MGL_3195	23.03
orf19.5542	MGL_3195	25.53
orf19.5585	MGL_3195	26.25
orf19.5714	MGL_3195	28.61
orf19.5716	MGL_3195	26.14
orf19.6001	MGL_3195	27.43
orf19.6481	MGL_3331	24.59
orf19.6928	MGL_3195	23.41
orf19.756	MGL_3195	23.16
orf19.852	MGL_3565	22.90
orf19.853	MGL_3327	28.14
MGL_0534	orf19.5714	23.68
MGL_0599	orf19.1891	29.55
MGL_1932	orf19.6001	25.75
MGL_2125	orf19.3708	24.78
MGL_3195	orf19.1891	54.42
MGL_3327	orf19.3708	26.57
MGL_3328	orf19.3708	25.00
MGL_3329	orf19.5714	23.55
MGL_3330	orf19.242	20.72
MGL_3331	orf19.6001	25.00
MGL_3565	orf19.5542	25.65
MGL_4053	orf19.3708	24.25
MGL_4243	orf19.6001	20.00
MGL_4267	orf19.3708	26.18

Fourteen *M. globosa* secretory aspartyl proteases and 15 *C. albicans* aspartyl proteases belong to the PFAM Asp family. These proteases were compared to the proteases from the other species. The best matches from the other genome and the percentage of amino acid identity are shown. ID designations starting with MGL represent *M. globosa* genes, and ID designations starting with orf19 represent *C. albicans* genes. One of the *M. globosa* secretory aspartyl proteases belongs to the PFAM Peptidase_A22B and is not shown here.

SI Table 14. Repeat elements in the *M. globosa* genome

	Number of elements	Length occupied (bp)	Percentage of genome
Retroelements	119	13,066	0.15
LINEs	40	4,332	0.05
R1/LOA/Jockey	22	2,443	0.03
LTR elements	77	8,679	0.10
Ty1/Copia	23	1,615	0.02
Gypsy/DIRS1	50	6,835	0.08
DNA transposons	53	3,604	0.04
hobo-Activator	14	873	0.01
En-Spm	21	1,555	0.02
MuDR-IS905	12	461	0.01
Total interspersed repeats		18,215	0.20
Small RNA	30	8,287	0.09
Satellites	13	2,745	0.03
Simple repeats	657	28,620	0.32
Low complexity	202	12,477	0.14

Repeat elements were identified using RepeatMasker (39). Only 0.78% of the genome contains repeat elements. None of the *C. albicans*, *U. maydis*, *C. neoformans*, and *S. cerevisiae*-specific repeat sequences were identified in the *M. globosa* genome. The identified repeat elements with more than 10 copies are listed.