Genome expansion and lineage-specific genetic innovations in the world's largest organisms (*Armillaria*)

György Sipos^{1,2*}, Arun N. Prasanna³, Mathias C. Walter⁴, Eoin O'Connor⁵, Balázs Bálint⁶, Krisztina Krizsán³, Brigitta Kiss³, Jaqueline Hess⁷, Torda Varga³, Jason Slot⁸, Robert Riley⁹, Bettina Bóka¹⁰, Daniel Rigling², Kerrie Barry⁹, Juna Lee⁹, Sirma Mihaltcheva⁹, Kurt LaButti⁹, Anna Lipzen⁹, Rose Waldron⁵, Nicola M. Moloney⁵, Christoph Sperisen², László Kredics¹⁰, Csaba Vágvolgyi¹⁰, Andrea Patrignani¹¹, David Fitzpatrick⁵, István Nagy⁶, Sean Doyle⁵, James B. Anderson¹², Igor V. Grigoriev⁹, Ulrich Güldener⁴, Martin Münsterkötter^{1,13}, László G. Nagy^{3*}

1 Functional Genomics and Bioinformatics Group, Research Center for Forestry and Wood Industry, University of Sopron, Hungary

2 Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

3 Synthetic and Systems Biology Unit, Biological Research Center, Hungarian Academy of Sciences, Szeged, Hungary

4 Department of Genome-oriented Bioinformatics, Center of Life and Food Science Weihenstephan, Technische Universität München, Freising, Germany

5 Department of Biology, Maynooth University, Maynooth, Co. Kildare, Ireland

6 Seqomics Ltd. Mórahalom, Hungary

7 Department of Botany and Biodiversity Research, University of Vienna, Vienna, Austria

8 Department of Plant Pathology, Ohio State University, Columbus, USA

9 US Department of Energy (DOE) Joint Genome Institute, Walnut Creek, CA, USA

10 Department of Microbiology, University of Szeged, Szeged, Hungary

11 Functional Genomics Center, ETH and University of Zurich, Switzerland

12 Department of Biology, University of Toronto, Canada

13 Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Neuherberg, Germany

*Author for correspondence:

László G. Nagy, e-mail: lnagy@funkenomelab.com

György Sipos, e-mail: gyoergy.sipos@wsl.ch

***Armillaria* species are both devastating forest pathogens and some of the largest terrestrial organisms on Earth. They forage for hosts and achieve immense colony sizes using rhizomorphs, root-like multicellular structures of clonal dispersal. Here, we sequenced and analyzed genomes of four *Armillaria* species and performed RNA-Seq and quantitative proteomic analysis on seven invasive and reproductive developmental stages of *A. ostoyae*. Comparison with 22 related fungi revealed a significant genome expansion in *Armillaria*, affecting several pathogenicity-related genes, lignocellulose degrading enzymes and lineage-specific genes likely involved in rhizomorph development. Rhizomorphs express an evolutionarily young transcriptome that shares features with the transcriptomes of fruiting bodies and vegetative mycelia. Several genes show concomitant upregulation in rhizomorphs and fruiting bodies and shared cis-regulatory signatures in their promoters, providing genetic and regulatory insights into complex multicellularity in fungi. Our results suggest that the evolution of the unique dispersal and pathogenicity mechanisms of *Armillaria* might have drawn upon ancestral genetic toolkits for wood-decay, morphogenesis and complex multicellularity.**

The genus *Armillaria* includes some of the most devastating forest pathogens worldwide. *Armillaria* causes root rot disease in both gymno- and angiosperms, in forests, parks, and even vineyards in more than 500 host plant species¹. Most *Armillaria* species are facultative necrotrophs, which, after colonizing and killing the root cambium, transition to a saprobic phase, decomposing dead woody tissues of the host. As saprotrophs, *Armillaria* spp. are white rot (WR) fungi, which can efficiently decompose all components of plant cell walls, including lignin, (hemi-)cellulose and pectin². They produce fleshy fruiting bodies (honey mushrooms) that appear in large clumps around infected plants and produce sexual spores. The vegetative phase of *Armillaria* is predominantly diploid rather than dikaryotic like most basidiomycetes.

Individuals of *Armillaria* can reach immense sizes and include the ‘humongous fungus’, one of the largest terrestrial organisms on Earth³, measuring up to 965 hectares and 600 tons⁴ and displaying a mutation rate ~3 orders of magnitude lower than most filamentous fungi⁵. Individuals reach this immense size via growing rhizomorphs, 1-4 mm wide dark mycelial strings that allow the fungus to bridge gaps between food sources or host plants^{1,6} (hence the name shoestring root rot). Rhizomorphs develop through the aggregation and coordinated parallel growth of hyphae, similarly to some fruiting body tissues^{7,8}. As migratory and exploratory organs, rhizomorphs can grow approximately one m/year and cross several meters underground in search for new hosts, although roles in uptake and long-range translocation of nutrients have also been proposed^{1,9,10}. Root contact by rhizomorphs is the main mode of infection by the fungus, which makes the prevention of recurrent infection in *Armillaria*-contaminated areas particularly difficult¹. Despite their huge impact on forestry, horticulture and agriculture, the genetics of the pathogenicity of *Armillaria* species is poorly understood. The only -omics data to date have highlighted a substantial repertoire of plant cell wall degrading enzymes (PCWDE) and secreted proteins, among others, in *A. mellea* and *A. solidipes*^{11,12}, while analyses of the genomes of other pathogenic basidiomycetes (e.g. *Moniliophthora*^{13,14}, *Heterobasidion*¹⁵, *Rhizoctonia*¹⁶) identified genes coding for PCWDEs, secreted and effector proteins or secondary metabolism (SM) as putative pathogenicity factors. However, the life cycle and unique dispersal strategy of *Armillaria* prefigure other evolutionary routes to pathogenicity, which, along with potential other genomic factors (e.g. transposable elements¹⁷) are not yet known.

Here, we investigate genome evolution and the origins of pathogenicity in *Armillaria* using comparative genomics, transcriptomics and proteomics. We sequenced the genomes of four *Armillaria* species to combine with that of related saprotrophic, hemibiotrophic and mycorrhizal fungi. Transcript and proteome profiling of invasive and reproductive developmental

stages shed light on the role of rhizomorphs, several putative pathogenicity factors as well as the morphogenetic mechanisms of rhizomorph and fruiting body development.

Results

We report the genomes of *A. ostoyae*, *A. cepistipes*, *A. gallica* and *A. solidipes* sequenced using a combination of PacBio and Illumina technologies. Genomes of *Armillaria* species were assembled to 103–319 scaffolds comprising 58–85 Mb and were predicted to contain 20,811–25,704 genes (Table 1). In comparison, other sequenced species of the Physalacriaceae, *Flammulina velutipes* and *Cylindrobasidium torrendii* have 12,218 (35.6Mb) and 13,940 (31.5Mb) genes, respectively, while the sister genus of *Armillaria*, *Guyanagaster necrorhiza* has 14,276 (53.6Mb)(Fig. 1). *Armillaria* species share significant synteny, comprising macro- to microsynteny (Supplementary Figure 1), whereas mesosynteny, characteristic of certain fungal groups¹⁸ was not observed. Transposable element content of *Armillaria* genomes shows a modest expansion relative to other Agaricales and even distribution along the scaffolds, suggesting that their genome expansion is not driven by transposon proliferation, as observed in other plant pathogens¹⁷ (Supplementary Figs. 2-3, Supplementary Table 1).

Phylogenomic analysis based on 835 conserved single copy genes (188,895 amino acid sites) confirmed the position of *Armillaria* in the Physalacriaceae, with *Guyanagaster* and *Cylindrobasidium* as their closest relatives (Fig. 1/a). We estimated the age of pathogenic *Armillaria* spp. at 21 million years (myr) and their divergence from *Guyanagaster* at 42 myr (Supplementary Fig. 4), coincident with decreasing temperatures and the spread of deciduous forests in the Eocene. Reconstruction of genome-wide gene duplication and loss histories in 27 Agaricales species revealed an early origin for most genes, followed by lineage-specific gene losses in most family and genus level groups, except *Armillaria*, which showed a net genome expansion: 16,687 protein-coding genes were inferred for the most recent common ancestor

(mrca) of *Armillaria* (2,012 duplications, 945 losses), as opposed to 14,720 and 14,687 for the mrca of *Armillaria* and *Guyanagaster* and that of *Armillaria*, *Guyanagaster* and *Cylindrobasidium*, respectively (Fig. 1/a,d, Supplementary Fig. 5). Further expansion to 19,272 genes was inferred for the mrca of *A. solidipes*, *A. ostoyae*, *A. cepistipes* and *A. gallica* (3,192 duplications, 607 losses), although the highly fragmented *A. mellea* assembly might cause some duplications to map to this instead of the preceding node.

Duplicated genes were enriched in functions related to chitin and cellulose binding, polysaccharide metabolism (peroxidase, lyase, hydrolase and oxidoreductase activity), peptidase activity, transmembrane transport, extracellular region and gene expression regulation ($p < 0.05$, hypergeometric test, HT, Supplementary Table 2). In line with this, 214 domains were significantly overrepresented in *Armillaria* genomes ($p < 0.05$, HT) relative to other agarics, including several peptidase, glycoside hydrolase (GH) and pectinase domains, CBM50-s, expansins, multicopper oxidases as well as several pfams related to secondary metabolism (Supplementary Table 3). We found 571 protein clusters specific to *Armillaria* and *Guyanagaster* or subclades therein (Supplementary Table 4). These included CE4 chitooligosaccharide deacetylases, CBM50-s (LysM), iron permeases (FTR1), and 19 transcription factor (TF) families, among others, although a significant proportion (70%) had no functional annotations. Taken together, these results suggest that gene family expansion was the predominant mode of genome evolution in *Armillaria* and that the observed genome expansion is largely concerned with diverse extracellular functions, including several lineage-specific innovations some of which have previously been associated with pathogenicity.

Putative pathogenicity-related genes in *Armillaria*

Plant pathogenic fungi possess diverse gene repertoires to invade host plants and modulate their immune system¹⁸⁻²⁰. We cataloged 20 families of putative pathogenesis-related genes to assess if *Armillaria* share expansions of these families with other plant pathogens (Supplementary Table 5). *Armillaria* species are enriched in expansins ($p=4 \times 10^{-5}$, Fisher exact test, FET) and possess many cerato-platanin genes, which contribute to unlocking cell-wall polysaccharide complexes and cause cell death in host plants, respectively and might act as a first-line cell-lysis weaponry during invasion (Fig. 1/a). Carboxylesterases and salicylate hydroxylases show moderate enrichment in both *Armillaria* and *Moniliophthora* species, along with other weakly pathogenic taxa (e.g. *Marasmius fiardii*). Salicylate hydroxylases have been implicated in developing a tolerance to salicylic acid, which is released to block the jasmonic acid defense pathway of plants upon infection. In contrast, cutinases (CE5) are missing from *Armillaria* species but present in *Moniliophthora* spp.¹⁴ that primarily infect through leaves. LysM domains (CBM50) are overrepresented in *Armillaria* compared to other Agaricales ($p=2 \times 10^{-8}$, FET), while GH75 chitosanases are exclusively found in *Armillaria* and *Moniliophthora* species (Fig1/a, Supplementary Table 5). In other plant pathogens, these are involved in modifying fungal cell wall composition and/or capturing chitin residues to mask chitin-triggered immune signals in host plants²¹⁻²⁵, suggesting similar roles in *Armillaria* too.

Homologs of SM genes reported from *Heterobasidion*¹⁵ are overrepresented in *Armillaria* species ($p=0.03 - 10^{-11}$, FET, Fig. 1/a, Supplementary Table 3,5). These include terpene cyclases, NRPS-like prenyl transferases, and halogenases, as well as trichodiene and polyprenyl synthases, which have been linked to fungal pathogenesis, virulence and competition with other microbes²⁶.

Armillaria species as wood-rotting fungi

It has been shown that plant-fungal interactions in both pathogens^{14,15,18} and mycorrhizal fungi²⁷⁻²⁹ can draw upon the PCWDE repertoire of saprotrophic ancestors. Therefore, we compared the PCWDE complement of *Armillaria* species to that of other Agaricales with diverse lifestyles. In the saprotrophic phase of their lifecycle, *Armillaria* species cause white rot on wood, which is reflected in their PCWDE repertoire. Their genomes encode lignin-, cellulose-, hemicellulose- and pectin-degrading enzymes, indicating the potential to degrade all plant cell wall (PCW) components (Fig. 1/a, Supplementary Table 5). Lignolytic families are underrepresented in *Armillaria* ($p=0.02 - 10^{-10}$, FET), whereas cellulose- and xylan-degrading families generally show similar gene counts to other WR species, with notably higher copy numbers of GH1 (β -glucosidase, $p=3 \times 10^{-6}$, FET). On the other hand, several pectinolytic families are overrepresented in *Armillaria*. Pectin degrading families include GH28, GH78 and GH88, polysaccharide lyase (PL) 1, 3, 4 and 9 as well as carbohydrate esterase 8 (CE8), of which PL3, CE8 and CBM67 rhamnose binding modules are significantly enriched in *Armillaria* spp. ($p=0.02 - 10^{-11}$, FET) compared to WR Agaricales. The pectinolytic repertoire of *Armillaria* is unusual for WR fungi³⁰ and might indicate links to dicot pathogenicity²⁰. The PCWDE repertoire of *Armillaria* species underpins their ability to act powerful WR decayers and provides a pool to draw upon as necrotrophic pathogens. It might enable them to gain access to wood and avoid competition with other microbes by damaging live trees, a strategy that is unavailable to most WR fungi.

Expression profiles of rhizomorphs

Armillaria species spread either clonally by rhizomorphs or with sexual spores produced on fruiting bodies. Rhizomorphs are unique multicellular structures of *Armillaria* species, but their functions and morphogenetic origins are debated¹⁰. We obtained gene expression and proteome profiles for actively growing rhizomorph tips (0-5 cm from the apex) of *A. ostoyae* C18

buried in sawdust-rich medium and four fruiting body stages (Supplementary Table 6). Differential expression analysis identified 1,303 and 1,610 over- and underexpressed genes in rhizomorphs relative to vegetative mycelium grown on the same medium ($FC_{VM} > 4$, $p \leq 0.05$), respectively, marking one of the largest expression change in our experiments (Fig. 2). Similarly, the highest number of unique proteins ($n=729$) was detected in rhizomorphs compared to vegetative mycelia (Fig2/c-d). Upregulated genes were enriched for GO terms related to carbohydrate, lipid and secondary metabolism, hydrophobins and pectinolysis (Supplementary Table 6). Global expression profiles suggest that rhizomorphs are transitional between vegetative mycelium and fruiting bodies (Supplementary Figs. 6-7). Expression profiles of several PCWDE families and putative pathogenicity-related genes (cerato-platanins, expansins) in rhizomorphs resembles vegetative mycelia (Supplementary Figs 8-12), whereas that of many putative morphogenesis-related genes was shared with fruiting bodies.

Rhizomorphs express an evolutionarily young transcriptome, with most upregulated genes specific to *A. ostoyae* or the *Armillaria* clade (including *Guyanagaster* Fig. 2/b). Most of these young genes lacked functional annotations but were conserved in *Armillaria* species, suggesting important functions in the genus. Genes belonging to older phylostrata had comparatively more functional annotation terms. We found 414 genes, which had expression maxima in rhizomorphs and were significantly upregulated relative to vegetative mycelium ($FC_{VM} \geq 4$, Supplementary Table 7). The most highly upregulated genes included expansins ($\log_2 FC_{VM} = 10.44$), bZip and C₂H₂ transcription factors, three caspase genes, hydrophobins, cytochrome p450s, GHs as well as several unannotated genes. We found an overexpression of H₂O₂ neutralizing 5-oxoprolinases and signs of intensified biogenesis and cargo of extracellular proteins along the secretory pathway (Supplementary Figs 13-14).

We observed significant overexpression of three cerato-platanins, three expansins, two carboxylesterases, as well as SM-related trichodiene (five genes), polyketide (six genes) synthases and a polyprenyl synthase, among others (Supplementary Table 8). Notably, all

expansins showed upregulation in fruiting bodies too, which could indicate a role in fungal cell wall remodeling instead of cellulose degradation, to which fungal expansin-related genes have mostly been linked³¹. A gene coding for a family 6 bacterial extracellular solute-binding protein showed an expression peak in rhizomorphs but low or negligible expression in other developmental stages. Bacterial homologs of this gene function as Fe³⁺ and thiamin transporters and are required for the pathogen's survival in the host. This gene is a member of an *Armillaria*-specific cluster with homologs in all *Armillaria* species and bacterial proteins as closest BLAST hits (although ML gene trees don't support HGT). Its Fe³⁺ transporting ability and nearly exclusive expression in rhizomorphs makes us speculate of a role in substrate and/or host exploitation.

A diverse suite of PCWDE encoding genes was expressed at high levels in rhizomorphs, although overall, PCWDE gene expression was higher in vegetative mycelium (Supplementary Figs. 8-11). Pectin-degrading CAZy families showed the most remarkable upregulation, including several GH28s, pectinesterases, a GH88 and PL3 gene (Supplementary Fig. 11), although the latter showed elevated expression throughout fruiting body development also. Similarly, two HTP genes and eight laccases were significantly overexpressed (Supplementary Table 8), consistent with records of laccase activity in rhizomorphs³². However, both HTPs and three laccases were also upregulated in fruiting bodies, with HTPs mostly in gill tissues, suggesting a potential role in morphogenesis instead of lignin degradation. Cellulose, hemicellulose and xylan-degrading genes showed highest expression in vegetative mycelium, with the exception of three GH3, a GH43, GH76, AA9 and lignolytic POD genes (Supplementary Fig. 11). Overall, the significant expression of PCWDE genes in rhizomorphs (Supplementary Figs. 8-11) suggest that rhizomorphs may not only explore the surrounding for new substrates, but could actively assimilate nutrients from available food sources. Taken together, the expression data show that under our experimental settings rhizomorphs expressed only a

moderate set of pathogenicity-related and PCWDE genes, but which combined could confer invasive properties to rhizomorphs.

Morphogenesis

The morphogenetic machinery underlying rhizomorph development is among the least known aspects of the biology of *Armillaria*. As multicellular structures, rhizomorphs express a variety of genes encoding cell-wall proteins, including hydrophobins, pore-forming toxins, two CBM67 and four ricin-B-lectins, an annexin and a cell-wall integrity sensor, among others, indicating several fruiting body-like functions, such as hyphal adhesion, communication or defense. We found cell-wall biosynthesis genes to be generally but moderately upregulated in rhizomorphs (and stipes) (Supplementary Table 9, Supplementary Fig. 15). Further, several GMC oxidoreductases, two mating-type pheromone receptors, CBM50-s, a CBM5_12 and chitooligosaccharide deacetylase genes were significantly overexpressed (the latter two also in stipes). Four hydrophobins reached their highest expression in rhizomorphs whereas two showed high expression in rhizomorphs and stipes, but not caps or vegetative mycelium (Supplementary Table 9, Supplementary Fig. 16). A homolog of the *Cryptococcus* red and far-red sensing Tco3 photoreceptor was expressed at high levels ($\log_2FC=4.36$). Although its function in *Cryptococcus* is unknown³³, its overexpression in rhizomorphs and in brown film forming mycelia of *Lentinula*³⁴ could suggest morphogenesis-related functions.

We detected 19 significantly upregulated TFs, ten of which had peak expression in rhizomorphs across our experimental conditions. Although based on global TF expression, rhizomorphs are most similar to vegetative mycelium (Supplementary Fig. 17), several TFs showed shared overexpression in rhizomorphs and fruiting bodies. For example, the expression of AROS_01275 peaked in rhizomorphs and stipes, whereas a zf-Mynd TF was highly upregulated in rhizomorphs and all fruiting body stages, and is thus a candidate for governing complex multicellular development in *A. ostoyae*.

Fruiting bodies

Fruiting body production is probably the largest morphogenetic transition in the life cycle of Basidiomycota. It involves the reorganization of hyphal growth patterns and the execution of a complex developmental program. Indeed, we detected the largest number of differentially expressed genes (DEGs) and the second largest change in proteome in Stage I primordia, including 1644 up- and 1513 downregulated genes ($FC_{VM} > 4$, $p \geq 0.05$, Supplementary Table 10) and 435 unique proteins. Upregulated genes were enriched for gene expression regulation, lipid metabolism, amino acid transport and ribonuclease activity (Fig. 2/a, Supplementary Table 9) and included ten TFs and two Dicer-like genes. Following stage I primordia, we tracked expression levels in two cell lineages. While the stipe lineage showed minor changes throughout development (<150 DEGs, <300 unique proteins), cap differentiation included up to 1037 DEGs and 646 unique proteins related to signal transduction, carbohydrate metabolism or the regulation of biological processes (Supplementary Table 10). Genes upregulated in gills ($n=502$, 56 unique proteins) were enriched for functions related to protein phosphorylation, carbohydrate metabolism, protein kinase activity and ribonucleotide binding.

Hydrophobin genes showed stage and tissue-specific expression patterns, with four genes significantly overexpressed in stipes, one in stage I primordia and stage II caps and another from stage I primordia through young fruiting bodies, with lower expression values in later stages. Hydrophobins were generally not expressed in cap tissues (Supplementary Fig. 16), suggesting alternative sources of hydrophobin-related functionalities in *Armillaria* caps. An analysis of cysteine-rich, hydrophobic proteins (excluding hydrophobins) revealed 33 and 13 developmentally regulated genes with expression maxima in fruiting bodies and caps, respectively (Supplementary Table 11). In addition, three cell wall galactomannoproteins show high expression in fruiting bodies (up to $\log FC_{VM} = 13.5$, Supplementary Table 12), two of which were expressed only in caps, whereas AROS_19505 was highly expressed throughout

development. These genes are homologous to the *Aspergillus* hydrophobic cell-surface protein HsbA, which has been hypothesized to recruit hydrolytic enzymes during lignocellulose degradation³⁵ and appressorium attachment to host surfaces³⁶. As lignocellulose-related processes are inactive in fruiting bodies, their expression suggests specific roles during cap development, possibly in cell-wall remodeling. Several chitin metabolism-related genes were upregulated in fruiting bodies, including GH88 chitinases, a CBM50 containing gene, a GH75 and six chitooligosaccharide deacetylases (Supplementary Table 10), which might be related to generating development-specific cell-wall architectures. Interestingly, certain HTPs, GMC oxidoreductases (AA3) and pectinolytic genes, generally linked to lignocellulose degradation, were also upregulated in fruiting bodies (Supplementary Fig. 18).

Defense-related genes in rhizomorphs and fruiting bodies

Defense against predators and competitors is a fundamental property for complex multicellular structures. We detected a significant proportion of putative defense-related genes of *A. ostoyae* upregulated in rhizomorphs, fruiting bodies or both (Supplementary Table 13). Fruiting bodies expressed three fungal hemolysins: a deuterolysin and a bacterial MTX2 pore-forming toxin gene were highly upregulated across fruiting body development, whereas a second MTX2 homolog showed upregulation in stipes and rhizomorphs. The insecticidal activities of these hemolysins³⁷ and the observed expression patterns suggest a role in defense of fruiting body structures. Two cerato-platanin genes were significantly upregulated in rhizomorphs and fruiting bodies, as well as four ricin-B-lectin genes, previously implicated in defense against nematodes and insects³⁷. A homolog of the nematocidal *Coprinopsis ccl2*³⁸ and a gene encoding a thaumatin-like protein were highly expressed from stage I primordia to young fruiting bodies mature fruiting bodies and vegetative mycelium. Taken together, the upregulated genes indicate

a phylogenetically and functionally diverse defense arsenal expressed in both fruiting bodies and rhizomorphs.

Shared morphogenetic machineries between rhizomorphs and fruiting bodies

Rhizomorphs share a complex multicellular organization with fruiting bodies. Consistent with this, we observed several genes with similar expression patterns in rhizomorphs and various fruiting body tissues (Fig. 3). These included two mating type pheromone receptor and the white collar 1-2 genes, which mediate the initiation of fruiting body development and could point to shared developmental origins of rhizomorphs and fruiting bodies or two CBM67 genes that were upregulated in rhizomorphs and all fruiting body samples relative to vegetative mycelium. There were 442 genes that showed >fourfold elevated expression over vegetative mycelium and relatively constant expression across rhizomorph and fruiting body samples (Fig. 3/a, Supplementary Table 14), suggesting they are linked to complex multicellularity in *A. ostoyae*. A systematic analysis identified 2,225 genes showing higher expression in rhizomorphs and stipes than in vegetative mycelium and corresponding cap tissues, of which 63 were at least fourfold more abundant in all four pairs of successive developmental stages (Fig. 3/b). These included most of the yeast cell wall biosynthesis pathway homologs (Supplementary Table 9, Supplementary Fig. 19), pore forming toxins, hydrophobins, TFs and SM genes, among others. For caps and rhizomorphs 1,728 and 28 such genes were found, including several GMC oxidoreductases, a GH88, a ricin-b-lectin and other SM genes (Fig. 3/c, Supplementary Table 15). This indicates that rhizomorph development extensively draws upon fruiting body genes, and makes us speculate that fruiting body development could have been the cradle for the evolution of rhizomorphs in *Armillaria* spp.

To assess whether co-regulated genes possess a common cis-regulatory signature, we searched for putative TF binding sites (TFBS) by *de novo* motif discovery³⁹ in 1kb upstream of 63 stipe/rhizomorph and 28 cap/rhizomorph co-expressed genes as well as in promoters of those that showed constant high expression in all complex multicellular structures (Fig. 3). High-scoring motifs in each gene set grouped into eight clades in a UPGMA analysis. To validate these motif predictions, we checked whether the predicted motifs match known consensus TFBS sequences of *Saccharomyces* in the JASPAR⁴⁰ database. All predicted motifs matched experimentally determined yeast TFBS sequences in JASPAR, which, although the phylogenetic distance is large between *S. cerevisiae* and *A. ostoyae*, suggests strong biological relevance to our predictions. Taken together, the detected co-expressed genes and the presence of shared motifs in their promoters could suggest common regulatory mechanisms for rhizomorphs and fruiting bodies and probably represent elements of the regulatory networks governing multicellular development as well as cap and stipe differentiation in *Armillaria*.

Discussion

Forest pathogens in the genus *Armillaria* evolved from saprotrophic ancestors in the Agaricales. They have unusually large genomes for WR saprotrophs, which evolved mostly by gene family diversification, in contrast with genome expansions in other plant pathogenic fungi, which are primarily driven by TE proliferation¹⁷. We found many lineage-specific genes (including putative pathogenicity factors and pectinolytic families), coincident overrepresentation of several pathogenicity-related genes with other Basidiomycota (above all *Moniliophthora* species), suggesting convergent origins of pathogenicity in the Agaricales. *Armillaria* species encode a full complement of PCWDE genes, which comes as no surprise given their ability to cause WR, but provides a resource to draw upon in the evolution of pathogenicity. This could give *Armillaria* spp. early access to dead wood and a strategy to bypass competition with other microbes.

Rhizomorphs are some of the most unique structures of *Armillaria* spp. that enable them to become the largest terrestrial organism on Earth^{3,5}. They express a wide array of genes involved in secondary metabolism, defense, PCW degradation and to a lesser extent pathogenesis, which indicate active nutrient uptake and adaptations to a soil-borne lifestyle where competition with other microbes and defense against predators are crucial. This underpins their role as exploratory and assimilating organs adapted to bridge gaps between food sources or potential host plants. In terms of morphogenesis, rhizomorphs resemble fruiting bodies, with many stipe- and to a smaller extent cap-upregulated genes showing concomitant upregulation and shared cis-regulatory signatures. Both rhizomorphs and fruiting bodies show key traits of complex multicellularity, such as 3-dimensional organization, cell adhesion or a highly integrated developmental program. We hypothesize that the observed similarity in gene expression indicates common developmental origins and suggest that the evolution of rhizomorphs may have extensively drawn upon the genetic toolkit of fruiting body development in the Agaricomycotina. We identified genes putatively involved in cap- and stipe morphogenesis, as well as some co-expressed in all complex multicellular stages and candidate TF binding motifs within their promoters. These represent putative building blocks of the gene regulatory circuits governing mushroom development and enabled us to zero in on the genetic bases of complex multicellularity in fungi. This study has provided comparative genomics insights into the evolution, pathogenicity and multicellular development of a group of devastating forest pathogens. It should facilitate further understanding of the biology of *Armillaria*, which, combined with new genomic resources and *in planta* interrogation of their pathogenic behavior could bring the development of efficient strategies for containing the spread and damage of *Armillaria* root-rot disease in various forest stands soon within reach.

References

- 1 Baumgartner, K., Coetzee, M. P. & Hoffmeister, D. Secrets of the subterranean pathosystem of *Armillaria*. *Mol Plant Pathol* **12**, 515-534, doi:10.1111/j.1364-3703.2010.00693.x (2011).
- 2 Floudas, D. *et al.* The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science* **336**, 1715-1719, doi:10.1126/science.1221748 (2012).
- 3 Smith, M. L., Bruhn, J. N. & Anderson, J. B. The fungus *Armillaria bulbosa* is among the largest and oldest living organisms. *Nature* **356**, 428-431 (1992).
- 4 Stanosz, G. R., Patton, R. F. & Spear, R. N. Structure of *Armillaria* rhizomorphs from Wisconsin aspen stands. *Canadian Journal of Botany* **65**, 2124-2127, doi:10.1139/b87-292 (1987).
- 5 Anderson, J. B. & Catona, S. Genomewide mutation dynamic within a long-lived individual of *Armillaria gallica*. *Mycologia* **106**, 642-648, doi:10.3852/13-367 (2014).
- 6 Pareek, M., Cole, L. & Ashford, A. E. Variations in structure of aerial and submerged rhizomorphs of *Armillaria luteobubalina* indicate that they may be organs of absorption. *Mycological research* **105**, 1377-1387, doi:<http://dx.doi.org/10.1017/S0953756201004622> (2001).
- 7 Moore, D. *Fungal Morphogenesis*. (Cambridge University Press, 1998).
- 8 Yafetto, L., Davis, D. J. & Money, N. P. Biomechanics of invasive growth by *Armillaria* rhizomorphs. *Fungal genetics and biology : FG & B* **46**, 688-694, doi:10.1016/j.fgb.2009.04.005 (2009).
- 9 Morrison, D. J. Ion uptake by rhizomorphs of *Armillaria mellea*. *Canadian Journal of Botany* **53**, 48-51, doi:10.1139/b75-006 (1975).
- 10 Pareek, M. *Structure and role of rhizomorphs of Armillaria luteobubalina*, The University of New South Wales, (2006).
- 11 Collins, C. *et al.* Genomic and proteomic dissection of the ubiquitous plant pathogen, *Armillaria mellea*: toward a new infection model system. *J Proteome Res* **12**, 2552-2570, doi:10.1021/pr301131t (2013).
- 12 Ross-Davis, A. L. *et al.* Transcriptome of an *Armillaria* root disease pathogen reveals candidate genes involved in host substrate utilization at the host–pathogen interface. *Forest Pathology* **43**, 468-477, doi:10.1111/efp.12056 (2013).

- 411 13 Meinhardt, L. W. *et al.* Genome and secretome analysis of the hemibiotrophic fungal
412 pathogen, *Moniliophthora roreri*, which causes frosty pod rot disease of cacao:
413 mechanisms of the biotrophic and necrotrophic phases. *BMC genomics* **15**, 164,
414 doi:10.1186/1471-2164-15-164 (2014).
- 415 14 Mondego, J. M. *et al.* A genome survey of *Moniliophthora perniciosa* gives new insights
416 into Witches' Broom Disease of cacao. *BMC genomics* **9**, 548, doi:10.1186/1471-2164-9-
417 548 (2008).
- 418 15 Olson, A. *et al.* Insight into trade-off between wood decay and parasitism from the
419 genome of a fungal forest pathogen. *New Phytol* **194**, 1001-1013, doi:10.1111/j.1469-
420 8137.2012.04128.x (2012).
- 421 16 Hane, J. K., Anderson, J. P., Williams, A. H., Sperschneider, J. & Singh, K. B. Genome
422 sequencing and comparative genomics of the broad host-range pathogen *Rhizoctonia*
423 *solani* AG8. *PLoS Genet* **10**, e1004281, doi:10.1371/journal.pgen.1004281 (2014).
- 424 17 Raffaele, S. & Kamoun, S. Genome evolution in filamentous plant pathogens: why
425 bigger can be better. *Nat Rev Micro* **10**, 417-430 (2012).
- 426 18 Ohm, R. A. *et al.* Diverse lifestyles and strategies of plant pathogenesis encoded in the
427 genomes of eighteen Dothideomycetes fungi. *PLoS Pathog* **8**, e1003037,
428 doi:10.1371/journal.ppat.1003037 (2012).
- 429 19 Giraldo, M. C. & Valent, B. Filamentous plant pathogen effectors in action. *Nat Rev*
430 *Micro* **11**, 800-814, doi:10.1038/nrmicro3119 (2013).
- 431 20 Kubicek, C. P., Starr, T. L. & Glass, N. L. Plant cell wall-degrading enzymes and their
432 secretion in plant-pathogenic fungi. *Annu Rev Phytopathol* **52**, 427-451,
433 doi:10.1146/annurev-phyto-102313-045831 (2014).
- 434 21 de Jonge, R. *et al.* Conserved fungal LysM effector Ecp6 prevents chitin-triggered
435 immunity in plants. *Science* **329**, 953-955, doi:10.1126/science.1190859 (2010).
- 436 22 Liu, W. *et al.* Recent progress in understanding PAMP- and effector-triggered immunity
437 against the rice blast fungus *Magnaporthe oryzae*. *Molecular plant* **6**, 605-620,
438 doi:10.1093/mp/sst015 (2013).
- 439 23 Akcapinar, G. B., Kappel, L., Sezerman, O. U. & Seidl-Seiboth, V. Molecular diversity of
440 LysM carbohydrate-binding motifs in fungi. *Curr Genet* **61**, 103-113,
441 doi:10.1007/s00294-014-0471-9 (2015).

- 442 24 Lee, W. S., Rudd, J. J., Hammond-Kosack, K. E. & Kanyuka, K. Mycosphaerella
443 graminicola LysM effector-mediated stealth pathogenesis subverts recognition through
444 both CERK1 and CEBiP homologues in wheat. *Mol Plant Microbe Interact* **27**, 236-243,
445 doi:10.1094/mpmi-07-13-0201-r (2014).
- 446 25 Liu, H., Zhang, B., Li, C. & Bao, X. Knock down of chitosanase expression in
447 phytopathogenic fungus Fusarium solani and its effect on pathogenicity. *Current*
448 *Genetics* **56**, 275-281, doi:10.1007/s00294-010-0299-x (2010).
- 449 26 Fox, E. M. & Howlett, B. J. Secondary metabolism: regulation and role in fungal biology.
450 *Current opinion in microbiology* **11**, 481-487, doi:10.1016/j.mib.2008.10.007 (2008).
- 451 27 Kohler, A. *et al.* Convergent losses of decay mechanisms and rapid turnover of
452 symbiosis genes in mycorrhizal mutualists. *Nat Genet* **47**, 410-415, doi:10.1038/ng.3223
453 (2015).
- 454 28 Martin, F., Kohler, A., Murat, C., Veneault-Fourrey, C. & Hibbett, D. S. Unearthing the
455 roots of ectomycorrhizal symbioses. *Nat Rev Micro* **14**, 760-773,
456 doi:10.1038/nrmicro.2016.149 (2016).
- 457 29 Peter, M. *et al.* Ectomycorrhizal ecology is imprinted in the genome of the dominant
458 symbiotic fungus Cenococcum geophilum. *Nature communications* **7**, 12662,
459 doi:10.1038/ncomms12662 (2016).
- 460 30 Rytioja, J. *et al.* Plant-Polysaccharide-Degrading Enzymes from Basidiomycetes.
461 *Microbiology and Molecular Biology Reviews : MMBR* **78**, 614-649,
462 doi:10.1128/MMBR.00035-14 (2014).
- 463 31 Saloheimo, M. *et al.* Swollenin, a Trichoderma reesei protein with sequence similarity to
464 the plant expansins, exhibits disruption activity on cellulosic materials. *European journal*
465 *of biochemistry* **269**, 4202-4211 (2002).
- 466 32 Worrall, J. J., Chet, I. & Huttermann, A. Association of rhizomorph formation with laccase
467 activity in armillaria spp. *Journal of General Microbiology* **132**, 2527-2534 (1986).
- 468 33 Lee, K. T. *et al.* Systematic functional analysis of kinases in the fungal pathogen
469 Cryptococcus neoformans. *Nature communications* **7**, 12766,
470 doi:10.1038/ncomms12766 (2016).
- 471 34 Tang, L.-h. *et al.* Transcriptome analysis of candidate genes and signaling pathways
472 associated with light-induced brown film formation in Lentinula edodes. *Applied*
473 *microbiology and biotechnology* **97**, 4977-4989, doi:10.1007/s00253-013-4832-y (2013).

- 474 35 Ohtaki, S. *et al.* Novel hydrophobic surface binding protein, HsbA, produced by
475 *Aspergillus oryzae*. *Appl Environ Microbiol* **72**, 2407-2413, doi:10.1128/aem.72.4.2407-
476 2413.2006 (2006).
- 477 36 Soanes, D. M., Chakrabarti, A., Paszkiewicz, K. H., Dawe, A. L. & Talbot, N. J. Genome-
478 wide transcriptional profiling of appressorium development by the rice blast fungus
479 *Magnaporthe oryzae*. *PLoS Pathog* **8**, e1002514, doi:10.1371/journal.ppat.1002514
480 (2012).
- 481 37 Plaza, D. F., Lin, C.-W., van der Velden, N. S. J., Aebi, M. & Künzler, M. Comparative
482 transcriptomics of the model mushroom *Coprinopsis cinerea* reveals tissue-specific
483 armories and a conserved circuitry for sexual development. *BMC genomics* **15**, 492,
484 doi:10.1186/1471-2164-15-492 (2014).
- 485 38 Kunzler, M. Hitting the sweet spot-glycans as targets of fungal defense effector proteins.
486 *Molecules* **20**, 8144-8167, doi:10.3390/molecules20058144 (2015).
- 487 39 Pavesi, G., Mereghetti, P., Mauri, G. & Pesole, G. Weeder Web: discovery of
488 transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic*
489 *acids research* **32**, W199-203, doi:10.1093/nar/gkh465 (2004).
- 490 40 Mathelier, A. *et al.* JASPAR 2016: a major expansion and update of the open-access
491 database of transcription factor binding profiles. *Nucleic Acids Res* **44**, D110-115,
492 doi:10.1093/nar/gkv1176 (2016).
- 493 41 Prospero, S., Holdenrieder, O. & Rigling, D. Comparison of the virulence of *Armillaria*
494 *cepistipes* and *Armillaria ostoyae* on four Norway spruce provenances. *Forest Pathology*
495 **34**, 1-14 (2004).
- 496 42 Tsykun, T., Rigling, D. & Prospero, S. A new multilocus approach for a reliable DNA-
497 based identification of *Armillaria* species. *Mycologia* **105**, 1059-1076, doi:10.3852/12-
498 209 (2013).
- 499 43 Ford, K. L., Baumgartner, K., Henricot, B., Bailey, A. M. & Foster, G. D. A reliable in vitro
500 fruiting system for *Armillaria mellea* for evaluation of *Agrobacterium tumefaciens*
501 transformation vectors. *Fungal biology* **119**, 859-869, doi:10.1016/j.funbio.2015.06.007
502 (2015).
- 503 44 Kues, U. & Navarro-Gonzalez, M. How do Agaricomycetes shape their fruiting bodies?
504 1. Morphological aspects of development. *Fungal Biol Rev* **29**, 63-97,
505 doi:10.1016/j.fbr.2015.05.001 (2015).
- 506 45 Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read
507 SMRT sequencing data. *Nature methods* **10**, 563-569, doi:10.1038/nmeth.2474 (2013).

508 46 English, A. C. *et al.* Mind the gap: upgrading genomes with Pacific Biosciences RS long-
509 read sequencing technology. *PloS one* **7**, e47768, doi:10.1371/journal.pone.0047768
510 (2012).

511 47 Lam, K. K., LaButti, K., Khalak, A. & Tse, D. FinisherSC: a repeat-aware tool for
512 upgrading de novo assembly using long reads. *Bioinformatics* **31**, 3207-3209,
513 doi:10.1093/bioinformatics/btv280 (2015).

514 48 Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant
515 detection and genome assembly improvement. *PloS one* **9**, e112963,
516 doi:10.1371/journal.pone.0112963 (2014).

517 49 Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively
518 parallel sequence data. *Proceedings of the National Academy of Sciences of the United*
519 *States of America* **108**, 1513-1518, doi:10.1073/pnas.1017351108 (2011).

520 50 Martin, J. *et al.* Rnnotator: an automated de novo transcriptome assembly pipeline from
521 stranded RNA-Seq reads. *BMC genomics* **11**, 663, doi:10.1186/1471-2164-11-663
522 (2010).

523 51 Grigoriev, I. V. *et al.* MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic*
524 *acids research* **42**, D699-704, doi:10.1093/nar/gkt1183 (2014).

525 52 Darling, A. E., Carey, L. & Feng, W.-C. in *ClusterWorld Conference & Expo and the 4th*
526 *International Conference on Linux Cluster: The HPC Revolution 2003* (San Jose,
527 California, 2003).

528 53 Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale
529 detection of protein families. *Nucleic acids research* **30**, 1575-1584 (2002).

530 54 van Stijn, D. *Graph Clustering by Flow Simulation*, University of Utrecht, (2000).

531 55 Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Molecular*
532 *biology and evolution* **25**, 1307-1320, doi:10.1093/molbev/msn067 (2008).

533 56 Smith, S. A. & Dunn, C. W. Phyutility: a phyloinformatics tool for trees, alignments and
534 molecular data. *Bioinformatics* **24**, 715-716, doi:10.1093/bioinformatics/btm619 (2008).

535 57 Nagy, L. G. *et al.* Genetic bases of fungal wood decay predicted by phylogenomic
536 analyses of gene-phenotype coevolution. *Mol Biol Evol* **accepted with revisions** (2016).

537 58 Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with
538 thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690, doi:bt1446
539 [pii]10.1093/bioinformatics/btl446 (2006).

540 59 Sanderson, M. J. Estimating absolute rates of molecular evolution and divergence times:
541 a penalized likelihood approach. *Molecular biology and evolution* **19**, 101-109 (2002).

542 60 Hibbett, D., Grimaldi, D. & Donoghue, M. Fossil mushrooms from Miocene and
543 Cretaceous ambers and the evolution of Homobasidiomycetes. *Am J Bot* **84**, 981 (1997).

544 61 Matheny, P. B. *et al.* Major clades of Agaricales: a multilocus phylogenetic overview.
545 *Mycologia* **98**, 982-995 (2006).

546 62 Poinar, G. O., Jr. & Buckley, R. Evidence of mycoparasitism and hypermycoparasitism in
547 Early Cretaceous amber. *Mycological research* **111**, 503-506,
548 doi:10.1016/j.mycres.2007.02.004 (2007).

549 63 Nagy, L. G. *et al.* Latent homology and convergent regulatory evolution underlies the
550 repeated emergence of yeasts. *Nature communications* **5**, 4471,
551 doi:10.1038/ncomms5471 (2014).

552 64 Wu, Y. C., Rasmussen, M. D., Bansal, M. S. & Kellis, M. TreeFix: statistically informed
553 gene tree error correction using species trees. *Syst Biol* **62**, 110-120,
554 doi:10.1093/sysbio/sys076 (2013).

555 65 Mitchell, A. *et al.* The InterPro protein families database: the classification resource after
556 15 years. *Nucleic acids research* **43**, D213-221, doi:10.1093/nar/gku1243 (2015).

557 66 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
558 improvements in performance and usability. *Molecular biology and evolution* **30**, 772-
559 780, doi:10.1093/molbev/mst010 (2013).

560 67 Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating
561 signal peptides from transmembrane regions. *Nature methods* **8**, 785-786,
562 doi:10.1038/nmeth.1701 (2011).

563 68 Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element
564 diversification in de novo annotation approaches. *PloS one* **6**, e16526,
565 doi:10.1371/journal.pone.0016526 (2011).

566 69 Hoede, C. *et al.* PASTEC: an automatic transposable element classification tool. *PloS*
567 *one* **9**, e91929, doi:10.1371/journal.pone.0091929 (2014).

568 70 Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for
569 de novo detection of LTR retrotransposons. *BMC bioinformatics* **9**, 18,
570 doi:10.1186/1471-2105-9-18 (2008).

571 71 Edgar, R. C. & Myers, E. W. PILER: identification and classification of genomic repeats.
572 *Bioinformatics* **21 Suppl 1**, i152-158, doi:10.1093/bioinformatics/bti1003 (2005).

573 72 Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in
574 sequenced genomes. *Genome research* **12**, 1269-1276, doi:10.1101/gr.88502 (2002).

575 73 Quesneville, H. *et al.* Combined evidence annotation of transposable elements in
576 genome sequences. *PLoS computational biology* **1**, 166-175,
577 doi:10.1371/journal.pcbi.0010022 (2005).

578 74 Kohany, O., Gentles, A. J., Hankus, L. & Jurka, J. Annotation, submission and screening
579 of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC bioinformatics*
580 **7**, 474, doi:10.1186/1471-2105-7-474 (2006).

581 75 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for
582 differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-
583 140, doi:10.1093/bioinformatics/btp616 (2010).

584 76 Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression
585 analysis of RNA-seq data. *Genome biology* **11**, R25, doi:10.1186/gb-2010-11-3-r25
586 (2010).

587 77 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing
588 and microarray studies. *Nucleic acids research* **43**, e47, doi:10.1093/nar/gkv007 (2015).

589 78 Dolan, S. K. *et al.* Regulation of nonribosomal peptide synthesis: bis-thiomethylation
590 attenuates gliotoxin biosynthesis in *Aspergillus fumigatus*. *Chem Biol* **21**, 999-1012,
591 doi:10.1016/j.chembiol.2014.07.006 (2014).

592 79 Moloney, N. M. *et al.* The iron-responsive microsomal proteome of *Aspergillus*
593 *fumigatus*. *J Proteomics* **136**, 99-111, doi:10.1016/j.jprot.2015.12.025 (2016).

594 80 O'Keeffe, G. *et al.* RNA-seq reveals the pan-transcriptomic impact of attenuating the
595 gliotoxin self-protection mechanism in *Aspergillus fumigatus*. *BMC genomics* **15**, 894,
596 doi:10.1186/1471-2164-15-894 (2014).

597 81 Owens, R. A. *et al.* Interplay between Gliotoxin Resistance, Secretion, and the
598 Methyl/Methionine Cycle in *Aspergillus fumigatus*. *Eukaryot Cell* **14**, 941-957,
599 doi:10.1128/ec.00055-15 (2015).

600 82 Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized
601 p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*
602 **26**, 1367-1372, doi:10.1038/nbt.1511 (2008).

- 603 83 Pavesi, G., Mauri, G. & Pesole, G. An algorithm for finding signals of unknown length in
604 DNA sequences. *Bioinformatics* **17 Suppl 1**, S207-214 (2001).
- 605 84 Mahony, S. & Benos, P. V. STAMP: a web tool for exploring DNA-binding motif
606 similarities. *Nucleic acids research* **35**, W253-258, doi:10.1093/nar/gkm272 (2007).
- 607 85 Gupta, S., Stamatoyannopolous, J. A., Bailey, T. & Noble, W. S. Quantifying similarity
608 between motifs. *Genome biology* **8**, R24 (2007).

609

610 Acknowledgements

611 This work was supported by the 'Momentum Program' of the Hungarian Academy of Sciences
612 under LP2014/12, and an OTKA-ERC_HU #118722 from the NRD Office to L.G.N. Fungal
613 research at Maynooth University is part-funded by a Science Foundation Ireland Investigator
614 award to SD (12/IP/1695). LC-MS facilities were funded by a competitive award from Science
615 Foundation Ireland (12/RI/2346 (3)). G.S. acknowledges support from a WSL grant for a guest
616 researcher in Switzerland and funding from GINOP-2.3.2-15-2016-00052. The work conducted
617 by the U.S. Department of Energy Joint Genome Institute (JGI), a DOE Office of Science User
618 Facility, was supported by the Office of Science of the U.S. Department of Energy under
619 Contract No. DE-AC02-05CH11231. IN was supported by a Janos Bolyai Research Fellowship
620 of the Hungarian Academy of Sciences. We also thank Renate Heinzelmänn for verifying the
621 haploid status of the *A. ostoyae* and *A. cepistipes* isolates used for genome sequencing and
622 Catherine Aime and Francis Martin for providing access to unpublished genome data produced
623 by JGI. Two anonymous reviewers are thanked for their constructive comments on earlier
624 versions of the paper.

625

Author contributions

G.S. and L.G.N. conceived the study, G.S., K.B., C.S., J.L., U.G., S.M., K.L., A.L., R.W., M.C.W., A.P., I.V.G. and M.M. performed genome sequencing, assembly and annotation, N.M.M., E.O. and S.D. performed the proteomic study, D.F. analyzed mating genes, J.H. performed analyses of TE content, A.P., B.K., I.N., Be.B., L.K., C.V. and B.B. obtained and analyzed RNA-Seq data for *A. ostoyae*, A.P., K.K., J.S., R.R. and L.G.N. performed comparative genomic and phylogenomic analyses, T.V. performed the molecular clock dating. L.G.N, A.P., J.A., G.S. wrote the manuscript. D.R. contributed tested strains for sequencing. All authors read and commented on the manuscript.

Competing Financial Interests

The authors declare no competing financial interests.

Figure Legends and Tables

Fig. 1. Genome evolution and phylogenomics in *Armillaria*. **(a)** Reconstructed gene duplication/loss histories along a time-calibrated phylogeny of 27 fungal species shows the expansion of protein coding gene repertoires in *Armillaria* spp. The heatmap shows gene copy numbers for PCWDE and pathogenicity-related gene families in the 27 species examined. **(b)** Assembly sizes (in Mb) for 27 species broken down by TEs (orange) and non-TEs (gray). Data for *A. mellea* are incomplete due to the highly fragmented available assembly (29,300 scaffolds). *Tricholoma matsutake* has been truncated for clarity. See Supplementary Fig. 2 for details **(c)** Numbers of predicted genes by phylogenetic conservation. **(d)** Numbers of duplicated (+) and lost (-) genes inferred for the Physalacriaceae, showing the net genome expansion in

Armillaria spp. and the structure of their mating loci showing high levels of synteny. The locus comprises of four homeodomain (HD) transcription factors arranged into *MAT-Aα* and *MAT-Aβ* respectively. *β-fg* and *MIP* genes anchor the locus. Genes containing carbohydrate esterase (CE), proline racemase (PR), deoxyribose-phosphate aldolase (DPA), metacarpases (MC) and dehydrogenase/reductase (DH) domains (pfam) are located within the locus. For clarity, *A. mellea* is not displayed as its assembly is highly contiguous.

Fig. 2. Expression analysis of invasive and reproductive developmental stages of *A. ostoyae* C18. **(a)** Gene expression and proteomic map of ten developmental stages and tissue types. Green and red arrows denote the numbers of significantly up- and downregulated genes ($p < 0.05$, $FC > 4$), respectively. Bar diagrams show the number of detected proteins common (grey area) to successive developmental stages, and the ones unique to the first (white) or the next (black) stage. The development of rhizomorphs (RH, examples shown for growth in the dark and light, left to right) and that of stage I primordia (P1) from vegetative mycelium (VM) denote the largest expression change, whereas stage II primordia (P2), young fruiting bodies (YFB) and fruiting bodies (FB) show smaller numbers of DEGs (FDR-corrected $P < 0.05$, $FC > 4$). Each sample was sequenced in 3 biological replicates **(b)** Phylostratigraphic analysis of 1,303 gene upregulated in rhizomorphs. Inset graph shows the age distribution of all genes in the genome of *A. ostoyae*. **(c)** Number of detected proteins overlapping in the compared developmental stages. Note the similarity between vegetative mycelium and rhizomorphs and between all complex multicellular stages (267 proteins). **(d)** Protein abundance profiles in ten developmental stages and tissue types. The y-axis shows fold change values.

Fig. 3. Predicted motifs among promoter regions of co-expressed genes in *Armillaria* rhizomorphs and fruiting bodies. Putative transcription factor binding sites and expression

profiles are shown for three co-expressed gene sets identified in rhizomorph and fruiting body transcriptomes. **(a)** 442 genes that are at least fourfold upregulated relative to vegetative mycelium and show relatively constant expression in all complex multicellular structures (stage I primordia through fruiting bodies). **(b)** and **(c)** 63 and 28 genes with concomitant upregulation in rhizomorphs and stipe tissues and rhizomorphs and caps, respectively, showing over fourfold expression difference in rhizomorphs relative to mycelium and all four pairs of successive developmental stages. Highlighted are putative TF genes detected in the co-expressed gene sets.

Table 1. Summary statistics of the new genomes reported in this paper.

Species	Assembly size (# scaffolds)	N50	# genes	TE content	BUSCO	CEGMA
<i>A. cepistipes</i>	75.5 (182)	8 (3.29 Mb)	23461	34.8%	95.1%	97.9%
<i>A. gallica</i>	85.34 (319)	22 (1.04 Mb)	25704	32.6%	98.6%	99.8%
<i>A. ostoyae</i>	60.1 (106)	9 (2.28 Mb)	22705	27.3%	95.7%	99.2%
<i>A. solidipes</i>	58.01 (229)	21 (716 kb)	20811	20.9%	98.4%	99.8%

Methods

Strains and fungal material used for genome sequencing

The diploid *A. ostoyae* C18 is a field isolate from Switzerland⁴¹ and the haploid C18/9 is derived from C18 as a single spore isolate. The *A. cepistipes* B5 haploid⁴² was originally isolated from a fruiting body on *Fagus sylvatica* in Italy (H. Marxmüller) and then included in the set of haploid tester strains by J.J. Guillaumin. The *A. ostoyae* C18/9 and *A. cepistipes* B5 haploid isolates were cultured on Roth and Shaw plates (for 1 L RS: 40 g malt extract, 20g dextrose, 5g bacto peptone, 19g agar) covered with cellophane sheets, and incubated at 25°C for three weeks.

Prior to genomic DNA extraction, fungal mycelia were detached and harvested from the cellophane sheets, and frozen in liquid nitrogen. The *A. gallica* 21-2 and *A. solidipes* 28-4 strains were maintained on 2% malt extract, 2% agar medium. To produce mycelium for DNA extraction, strains were grown in liquid CYM medium (d-glucose 20g, yeast extract 2 g, peptone 2 g, MgSO₄·7H₂O 0.5 g, KH₂PO₄ 0.46 g, K₂HPO₄ 1g, 1l water) in Petri dishes with 8-10 small pieces of inoculum floated on the air-water interface. Genomic DNA was extracted using the PowerMax MOBIO DNA isolation kit according to the manufacturer's instructions. Note that Burdsall and Volk (2008) have proposed that the *A. solidipes* and *A. ostopae* are one species and that the earlier name *A. solidipes* from a North American collection should always be used in place of *A. ostopae*. For consistency with recent historical usage, however, we choose to consider *A. ostopae* in Europe as a separate, but closely related species, to *A. solidipes* in North America

We developed an *in vitro* fruiting system for *A. ostopae* C18 based on the protocol published for *A. mellea*⁴³. A modified RST medium, termed RSTO was prepared in 720 ml jars and inoculated with *A. ostopae* C18, incubated for 28 days at 24°C in the dark, then placed in a growth chamber at 15°C with a 10/14 hour light/dark cycle (light intensity: 11 µE m⁻²s⁻¹). Vegetative mycelia, rhizomorphs, stage I and II primordia, young and mature fruiting bodies were harvested as shown on Fig. 1. Fruiting body stages were defined to conform to the general notation⁴⁴ of mushroom developmental stages as closely as possible. Stage I primordia were defined as up to 1-2 mm tall fruiting body initials without a clear differentiation of a pileal section. Stage II primordia had a developed button-shaped cap, up to 7-10 mm tall. Young fruiting bodies were 30-50 mm tall with a developed hymenial cavity but prior to cap expansion. From stage II primordia, caps and stipes were separated and RNA was extracted separately. Mature fruiting bodies were separated to stipe, gills and cap. RNA was extracted by using the RNEasy Midi kit (QIAGEN) following the manufacturer's instructions. RNA quality was checked by gel

electrophoresis and the Agilent 2200 TapeStation before library preparation. Biological triplicates were analyzed for all sample types.

For RNA-Seq, *A. cepistipes* was grown on MEA (for 1L MEA: 20 g malt extract, 0.5 g yeast extract, 15 g agar), RST (15 g *Picea abies* sawdust, 30 g rice, ca. 100 ml water mixed together in 720 ml jars, sterilized and overlaid with a ca. 2cm layer of homogenized tomato followed by another round of sterilization), RSTO (RST medium with additional 100 g minced orange), Orange media (3 roughly chopped oranges in 720 ml jars) and RNA was extracted using the RNEasy Midi kit (QIAGEN).

Genome sequencing and assembly

The haploid genomes of *A. ostoyae* and *A. cepistipes* were sequenced employing the PacBio (Pacific Biosystem) RS II platform (Functional Genomics Center, ETH and University of Zurich, Switzerland; <http://www.fgc.zh.ch>). For preparing the sequencing libraries, 10µg of gDNA aliquots were mechanically sheared to an average size of 10kb using the Covaris gTube (KBiosciences p/n 520079, Brighton, UK) in an Eppendorf microcentrifuge, and the fragment size distributions were assessed applying the Bioanalyzer 2100 12K DNA-Chip assay (Agilent p/n 5067-1508). Five µg of the sheared gDNA aliquots were DNA damage repaired, end-repaired and the final SMRTbell templates were created by blunt end ligation and exonuclease treatments. The libraries were quality inspected on an Agilent Bioanalyzer 12K DNA Chip and quantified on a Qubit.1 Fluorimeter (Life Technologies). The SMRTbell was set up by using the DNA Template Prep Kit 2.0 (3 kb - <10 kb) (Pacific Biosciences p/n 001-540-835). A ready to sequence SMRTbell-Polymerase Complex was arranged by applying the P4 DNA/Polymerase binding kit 2.0 (Pacific Biosciences p/n 100-236-500) according to the manufacturer instructions. Libraries were sequenced on 15 SMRT cell v3.0 (Pacific Biosciences p/n100-171-800), taking 1 movie of 120 minutes each per SMRT cell. The MagBead loading (PacBio p/n 100-133-600) technique

served to improve the enrichment for the longer fragments. Final sequencing reports were generated for every cell, via the SMRT portal, to assess the adapter dimer contamination, the sample loading efficiency, the obtained average read length and the number of filtered sub-reads.

The HGAP3⁴⁵ workflow of the SMRT Analysis suite v2.3 was used to create an initial assembly. After removal of redundant contigs, a scaffolding using PBJelly2⁴⁶ and FinisherSC⁴⁷ followed by polishing via applying the RS Resequencing protocol (SMRT Analysis suite) was performed in four iterations. The final scaffold set was checked for miss-assemblies using the RS BridgeMapper protocol (SMRT Analysis suite) and corrected if necessary. The mitochondrial scaffolds were first identified using a BLASTn search and then circularized by merging and truncating the overlapping ends. In order to correct PacBio reads, Illumina sequencing was carried out by shotgun sequencing of a 350-450 bp library with paired-end 100 bp reads to 180 fold coverage using HiSeq 2000 by the Functional Genomic Center at the ETH, Zurich. Finally, Pilon⁴⁸ was applied to further polish the scaffolds with the genomic Illumina reads.

Draft gene models for *A. ostoyae* and *A. cepistipes* were generated by three de-novo prediction programs: 1) Fgenesh (Salamov and Solovyev 2000) with different matrices (trained on *Aspergillus nidulans*, *Neurospora crassa* and a mixed matrix based on different species); 2) GeneMark-ES (Ter-Hovhannisyan et al. 2008) and 3) Augustus (Stanke et al. 2006) with RNA-seq based transcripts as training sets. Annotation was aided by exonerate (Slater and Birney 2005) hits of protein sequences from *Armillaria* spec. to uncover gene annotation gaps and to validate de-novo predictions. Transcripts were assembled on the RNA-seq data sets using Trinity (Grabherr et al. 2011). The different gene structures and evidences (exonerate mapping, RNA-seq reads and transcripts) were visualized in GBrowse (Donlin 2009) allowing manual validation of coding sequences with a focus on Chitin, Cellulose, Pectin, Lignin, SM key genes and other genes of interest. The best fitting model per locus was selected manually and gene structures were adjusted by splitting or fusion of gene models and redefining exon-intron

boundaries if necessary. tRNAs were predicted using tRNAscan-SE (Lowe and Eddy 1997). The predicted protein sets were searched for highly conserved single (low) copy genes to assess the completeness of the genomic sequences and gene predictions. Orthologous genes to all 246 single copy genes were searched for all proteomes by blastp comparisons (eVal: 10⁻³) against the single-copy families from all 21 species available from the FunyBASE (Aguileta et al. 2008). Additionally, the proteomes were searched for the 248 core-genes commonly present in higher eukaryotes (CEGs) by Blastp comparisons (eVal: 10⁻³) (Parra et al. 2009). All genomes were analyzed using the PEDANT system (Walter et al. 2009).

The genomes and transcriptomes of *A. gallica* and *A. solidipes* were sequenced using the Illumina platform at JGI. Genomic DNA was sequenced as pairs of Illumina standard and Nextera long mate-pair (LMP) libraries. For the standard libraries, 500 ng of DNA was sheared to 270 bp using the Covaris E220 (Covaris) and size selected using SPRI beads (Beckman Coulter). The fragments were treated with end-repair, A-tailing, and ligation of Illumina compatible adapters (IDT, Inc) using the KAPA-Illumina library creation kit (KAPA biosystems). For LMP, 1µg of DNA was used to generate the library using the Nextera LMP kit (Illumina). DNA was fragmented and ligated with biotinylated linkers using the Tagmentation enzyme. The fragments were circularized via ligation followed by random shearing using the Covaris LE220 (Covaris). The mate pair fragments were purified using Streptavidin beads (Invitrogen) and treated with end repair, A-tailing, and ligation of Illumina adaptors. The final product was enriched with ten cycles of PCR.

For transcriptomics, stranded cDNA libraries were generated using the Illumina TruSeq Stranded RNA LT kit. mRNA was purified from 1µg of total RNA using magnetic beads containing poly-T oligos, fragmented and reverse transcribed using random hexamers and SSII (Invitrogen), followed by second strand synthesis. The fragmented cDNA was treated with end-pair, A-tailing, adapter ligation, and ten cycles of PCR.

All libraries were quantified using KAPA Biosystem's next-generation sequencing library qPCR kit and run on a Roche LightCycler 480 real-time PCR instrument. Except for *A. gallica* LMP, the quantified libraries were prepared for sequencing on the Illumina HiSeq sequencing platform utilizing a TruSeq paired-end cluster kit, v3(*A. gallica*) or v4(*A. ostoyae*), and Illumina's cBot instrument to generate clustered flowcells for sequencing. Sequencing of the flowcells was performed on the Illumina HiSeq2000 or HiSeq2500 sequencer using a TruSeq SBS sequencing kit, v3 or v4, respectively, following a 2x150 indexed run recipe. Sequencing of the *A. gallica* LMP library was performed on the Illumina MiSeq sequencer using a MiSeq Reagent kit, v2, following a 2x150 indexed run recipe.

Genomic reads from each pair of libraries were QC filtered for artifact/process contamination and assembled together with AllPathsLG⁴⁹. Illumina reads of stranded RNA-Seq data were used as input for *de novo* assembly of RNA contigs, assembled into consensus sequences using Rnnotator (v. 3.4)⁵⁰. Both genomes were annotated using the JGI Annotation Pipeline and made available via the JGI fungal portal MycoCosm⁵¹.

CEGMA and BUSCO were used to assess the completeness of the assemblies. We used BUSCO Version 3.0.1 with the lineage specific profile library basidiomycota_odb9 (species:selected 41:25 with 1.335 BUSCO groups) downloaded from <http://busco.ezlab.org>. We performed the whole analysis in Gene set (proteins) assessment mode. Results there combined in a folder and plotted with the script generate-plot.py

Phylogenomic analysis

We used whole-genomes of five *Armillaria* spp. and 22 Agaricales encompassing white and brown rot wood decayers, litter decomposers and ECM fungi and species from the Russulales and Boletales included as outgroups (Supplementary Table 5). An all-versus-all protein BLAST was performed using mpiBLAST-1.6.0⁵² with default parameters to find homologs based on

similarity. Homologs were clustered into protein families using the Markov Cluster Algorithm^{53,54} with an inflation parameter of 2.0. For each cluster, a multiple sequence alignment was inferred using PRANK v.150803 (Loytynoja A et al, 2005), run with default settings. Subsequently, spuriously aligned regions were excluded with trimAl v1.4.r15 (Capella-Gutierrez et al, 2009), with a gap threshold of 0.2. Next, we inferred Maximum likelihood (ML) gene trees for each cluster of a minimum of four proteins using FastTree v2.19 (Price et al, 2010). FastTree was run with CAT/GAMMA20 model for rate heterogeneity and an improved LG model⁵⁵ for substitution probabilities. Gene trees were mid-point re-rooted using Phyutility v 2.2.6⁵⁶.

Next, we screened the gene trees with a custom Perl script⁵⁷ for the presence of deep paralogs and inparalogs. Gene trees with deep paralogs were eliminated, while in trees with inparalogs, the paralog closest to the root was retained in the corresponding alignments. Finally, 835 clusters passed these criteria and had >50 amino acid sites and (>=) 60% taxon occupancy were concatenated into a supermatrix with 188895 sites. We used this supermatrix to infer species tree with RAxML PTHREADS version 7.2.8⁵⁸ using a partitioned WAG+G model, where each data partition represented a single input gene family. To assess branch support of the tree we performed a bootstrap analysis in RAxML with 100 replicates under the same model.

We used penalized likelihood (PL) implemented in the program r8s version 1.70⁵⁹ for molecular clock dating based on the optimal ML tree and two fossils and one secondary calibration. *Archaeomarasmius legettii* (94-90 Myr)⁶⁰ was used to define minimum age constraint (92 Myr) for the origin of marasmioid clade (*sensu* Matheny 2006⁶¹). *Palaeoagaricites antiquus* (100-110 Myr) is the oldest fossil can be placed within Agaricales⁶², hence we constrained the most recent common ancestors (mrca) of this order to 108 Myr minimum age. To define the minimum age constraint of the origin of Boletales to 84 Myr, we referred to the analyses of Kohler *et al.*²⁷. Maximum age constraints were defined with a wide safety window to allow for the calibrated clades to be inferred at least twice as old as their minimum ages. The mrca of the Boletales, that of the Agaricales and that of the marasmioid clade were constrained

between 84-250 Myr, 108-250 Myr and 92-180 Myr, respectively, as minimum and maximum ages. A cross-validation (CV) analysis was used to identify an optimal smoothing parameter between 10^{-20} and 10^9 . All analyses were run in four replicates (random number seeds). We applied additive penalty function and run optimization for 25 times initialized from independent starting points. Furthermore after reaching an initial solution of an optimization step, the solution was perturbed and the truncated Newton (TN) optimization was rerun 20 times. We found that the optimal smoothing parameter (λ) can be between 10^{-7} and 10^{-3} , thus we estimated divergence time with five lambda parameter ranges between 10^{-7} and 10^{-3} . By checking the gradients and the estimated ages we found a high similarity between the results of the analyses obtained with different smoothing parameters (Supplementary Fig. 2). By inspecting the CV analyses and the gradient values a smoothing parameter of 10^{-6} was chosen as the most adequate and was used to estimate divergence times.

COMPARE analysis

We analyzed the evolutionary history of gene families of *Armillaria* species and closely related Agaricales using the COMPARE pipeline⁶³. To reconstruct gene duplications and losses, a genome-wide collection of 13821 gene trees (see previous section) was first reconciled with the species tree using Treefix v1.1.10⁶⁴. Treefix was run with RAxML site-wise likelihood model, Maximum Parsimony Reconciliation model (MPR) and an --alpha threshold of 0.001 to find any alternate gene tree topologies that minimize duplication/loss cost but are statistically equivalent to ML gene tree. Orthologs were identified and recoded into presence/absence matrix, as described in Nagy et al.⁶³. We then inferred duplications and losses for each orthogroup along the species tree using Dollo parsimony. Gene trees with less than four proteins were excluded. Subsequently, we also reconstructed the genome size for a given node by summing over gains and losses to the genome size of most recent common ancestor (MRCA). GO enrichment

analysis based on the Fisher exact test with Benjamin-Hochberg correction was performed using Pfam domains mapped to protein clusters and creating GO annotations with pfam2go version 2016/10/01⁶⁵ followed by enrichment analysis at $p < 0.05$.

Copy number analyses

We analyzed the copy number distribution of selected gene families in *Armillaria*, *Moniliophthora*, and *Heterobasidion annosum* and, non-pathogenic members of Agaricales. For this, predicted proteomes of all the taxa under study were annotated with Pfam domains and IPR signatures. We scanned the protein sequences for pfams using pfamscan.pl v1.5 against the Pfam 30.0 database and InterPro signatures using InterProscan v5. Next, we built search terms for 42 PCWDE families and 25 putative pathogenicity related genes based on literature evidence ('Search_Terms' in Supplementary Table 5). In few cases, Pfam signatures were either absent or did not yield consistent results. CBM67 copy numbers were obtained based on BLAST-hits by counting homologs obtained through blasting proteins annotated by JGI as CBM67. Genomes were searched iteratively for significant hits to avoid the impact of phylogenetic distance on the detection of homologs. We annotated cellobiose dehydrogenases (CDH, AA3_1), alcohol oxidases (Aox, AA3_3) and GMC oxidoreductases (GMT, AA3_2) using a tree-based approach. To this end, proteins of *Armillaria gallica* (Armga1) were combined with homologs in other genomes, followed by multiple sequence alignment construction using MAFFT v7.273⁶⁶ with -auto option and the estimation of a gene tree with FastTree v2.19 (as above). Classification was done based on the occurrence as monophyletic groups in the phylogenetic tree. Small secreted proteins (SSPs) were defined as proteins with less than 300 amino acids and sequence-based evidence for secretion as inferred by signalP v4.1⁶⁷. Both signalP-noTM and signalP-TM models were used for the detection of signal peptides.

Analysis of transposable elements

De novo element discovery and annotation - We used the REPET package version 2.5^{68,69} to identify, classify and annotate transposable elements (TEs) within the genomes studied. Since the *A. mellea* genome contained large amounts of “chaff” contigs, we thresholded this assembly to only include contigs larger than 1000 bps prior to further analysis.

The REPET *de novo* pipeline was run using genome self-alignment as well as a search for structural features using LTRHarvest⁷⁰. Consensus sequences were clustered using Piler v1.0⁷¹, Recon 1.08⁷² and Grouper⁷³ and classified using PASTEC⁶⁹. The resulting consensus sequences were filtered for low frequency elements, short sequence repeats and sequences that were classified as putative host genes and then clustered into families using Markov clustering⁵³. For annotation, the consensus sequences collected from each species were combined into a pan-species TE library. TE identification was then carried out using the REPET anno pipeline, implementing BLASTER⁷³, RepeatMasker 4.06 and CENSOR 4.2.29⁷⁴. Configuration files for the pipelines containing detailed parameter settings as well as a sample command line script can be found at https://github.com/JackyHess/Armillaria_TE_annotations

TE organization along the genome in core *Armillaria* species - In order to gain an understanding of how TEs are organized among the core *Armillaria* genomes we investigated TE and gene content density in 50 kb genome windows. Each genome was segmented into 50 kb partitions using bedtools (<http://bedtools.readthedocs.io>) and for each partition, TE coverage and genic coverage were estimated. Prior to estimating genic coverage, we filtered annotated CDS for which 20% of the sequence overlapped with TE annotations to remove putative TE-derived genes. This reduced the number of protein-coding genes considered by up to 7000 (Supplementary Table 1). All analysis scripts can be found at https://github.com/JackyHess/Armillaria_TE_annotations.

In-depth transcriptome analysis of *A. ostoyae*

Whole transcriptome sequencing was performed using TrueSeq RNA Library Preparation Kit v2 (Illumina) according to the manufacturer's instructions. Briefly, RNA quality and quantity measurements were performed using RNA ScreenTape and Reagents on TapeStation (all from Agilent) and Qubit (ThermoFisher); only high quality (RIN >8.0) total RNA samples were processed. Next, RNA was DNaseI (ThermoFisher) treated and the mRNA was purified and fragmented. First strand cDNA synthesis was performed using SuperScript II (ThermoFisher) followed by second strand cDNA synthesis, end repair, 3'-end adenylation, adapter ligation and PCR amplification. All purification steps were performed using AmPureXP Beads (Beckman Coulter). Final libraries were quality checked using D1000 ScreenTape and Reagents on TapeStation (all from Agilent). Concentration of each library was determined using the QPCR Quantification Kit for Illumina (Agilent). Sequencing was performed on Illumina NextSeq instrument using the NextSeq Series 2x150bp high-output kit (Illumina) generating >20 million clusters for each sample.

Bioinformatic Analysis Draft genome sequence together with genome annotation file was used as a reference for *A. ostoyae* RNA-Seq analysis. Paired-end Illumina NextSeq reads were quality trimmed using CLC Genomics Workbench tool version 9.5.2 (CLC Bio/Qiagen) removing ambiguous nucleotides as well as any low quality read end using an error probability cutoff value of 0.05 (corresponding to Phred score 13). Trimmed reads were mapped using the RNA-Seq Analysis 2.1 package in CLC allowing intergenic read mapping and requiring at least 80% sequence identity over at least 80% of the read lengths; strand specificity was omitted. Reads with less than 30 equally scoring mapping positions were randomly mapped; reads with more than 30 potential mapping positions were considered as uninformative repeat reads and were excluded from the analysis (Supplementary Table 6).

“Total gene read” RNA-Seq count data was imported from CLC into R version 3.0.2. Genes were filtered based on their expression levels keeping only those features that were detected by at least five mapped reads in at least 25% of the samples included in the study. Subsequently, “calcNormFactors” from package “edgeR” version 3.4.2⁷⁵ was used to perform data normalization based on the “trimmed mean of M-values” (TMM) method⁷⁶. Log transformation was carried out by the “voom” function of the “limma” package version 3.18.13⁷⁷. Linear modeling, empirical Bayes moderation as well as the calculation of differentially expressed genes were carried out using “limma”. Genes showing at least four-fold gene expression change with an FDR value below 0.05 were considered as significant. Multi-dimensional scaling (“plotMDS” function in edgeR) was also applied to visually summarize gene expression profiles revealing similarities between samples. In addition, unsupervised cluster analysis with Euclidean distance calculation and complete-linkage clustering was carried out on the normalized data using “heatmap.2” function from R package “gplots”.

Proteomics

Whole cell lysate preparation and trypsin digestion. Two hundred mg ($\pm 5\%$) from every developmental stage of *A. ostoyae* was ground in 600 μ L of ice-cold lysis buffer (100 mM Tris-HCl, 50 mM NaCl, 20 mM EDTA, 10 % (v/v) glycerol, 1 mM PMSF and 1 μ g/mL Pepstatin A pH 7.5) using a IKA T10 basic homogenizer (IKA-Werk GmbH & Co. KG, Germany) and PT 1200 E Ergonomic Homogenizer (POLYTRON PT 1200 E; Kinematica AG, Switzerland), respectively. Samples were sonicated (Bandelin Sonoplus HD2200, Bandelin Elec., Germany) three times on ice (MS73 probe, Cycle 6, 10s, Power 20%), lysates were incubated overnight on ice at 4°C and then clarified by centrifugation at 9,700 g for 10 min. Protein was quantified using the Bradford assay and samples were normalized for protein concentration, where possible, per replicate. Samples were brought to 15% (w/v) trichloroacetic acid (TCA) for precipitation, with acetone washes. Samples were resuspended in 6M urea, 2M thiourea and 100mM Tris-HCl pH 8.0 (60-

120 µL), protein concentrations analyzed, as above, and adjusted to 1 M urea final using ammonium bicarbonate prior to trypsin digestion. Protein samples (10 µg each) were sequentially reduced and alkylated with 5 µM DTT and 15µM IAA, respectively¹¹, brought to 0.01% (v/v) ProteaseMAX and trypsin added (1.6µL; 1 µg/µL) to a final protein amount (10µg) (37°C, overnight). Samples were acidified by addition of 1µL trifluoroacetic acid (TFA). Peptide solutions were then dried in a centrifugal evaporator, resuspended in 0.5% (w/v) TFA and desalted using C₁₈ Ziptips (Millipore[®] Ziptips C18)⁷⁸⁻⁸¹.

LC-MS, data processing and interpretation. Peptide mixtures were analyzed using a Thermo Fisher Q-ExactiveTM mass spectrometer coupled to a Dionex RSLCnano for LC-MS/MS analysis. LC gradients operated from 3-40%B over 40 min, with data collection using a Top15 method for MS/MS scans⁸². LC-MS spectra chromatograms were analysed manually using rawMeat software. Raw files corresponding to the aforementioned spectra were then processed against an *A. ostoyae* predicted protein database using MaxQuant (version 1.3.0.5) and further filtered and visualized in Perseus (Version1.4.1.3)⁸². MaxQuant parameters were as described in Owens *et al.*⁸¹. Peptide intensity values were normalized to log2 values in Perseus. Only samples represented in all replicates of a sample group were taken (n = 1 to 4). Proteins common to all stages were z –score normalised, averaged and hierarchical clustering was performed to generate heat maps and profile plots. Venn diagrams were prepared for all proteins (unique and abundant proteins inclusive) in the entire data set (<http://bioinformatics.psb.ugent.be>).

Proteomic analysis revealed the presence of 2,549 proteins, whereby 39.7% (1012/2,549) proteins were common to all *A. ostoyae* developmental stages and 25.2% (643/2,549) were unique to individual developmental stages. Proteins were unique to rhizomorph stage (RH; n = 286), vegetative mycelia (VM; n = 163), young fruiting body (YFB; n = 97), primordial stages 1 and 2 (PS1,2; n = 73) and mature fruiting body (MFB; n = 24). Quantitative changes in protein abundance were also evident between different development

stages, albeit to different extents. For instance, comparing RH to VM stages, individual proteins underwent increased ($n = 184$) and decreased ($n = 125$) abundance (total detected $n = 2190$), whereas comparative analysis of YFB against MFB revealed only eight proteins with increased abundance (total detected $n = 1203$). Comparative analysis of all other developmental stages yielded intermediate changes in the abundance of specific proteins.

Motif Discovery

We used 1 kb sequences of co-expressed genes (which most likely contain their promoter), located upstream of the start codon, including 5' UTR, to predict putative TFBS. We performed *de novo* motif discovery in our co-expressed gene sets using Weeder 2.0^{39,83}. We created frequency files from promoter regions of all genes of *A. ostoyae* using the Weeder2.0 frequencymaker. Frequencies were counted on both strands for motif lengths 6, 8 and 10, allowing 1, 2 and 3 mismatches, respectively. Using the inferred frequency counts as reference, both strands of the sequences were scanned using Weeder, allowing the detection of a maximum of 100 motifs. Next, we grouped the motifs based on Pearson correlation co-efficient, aligned using the ungapped Smith-Waterman algorithm and clustered using UPGMA, all executed via the STAMP webtool⁸⁴. Finally, we constructed familial binding profiles for each motif group and searched for matching DNA motifs in the JASPAR Core (Fungi) 2016 database using the TOMTOM webserver⁸⁵.

Data availability

Genome assemblies and annotation were also deposited at DDBJ/EMBL/GenBank under the following accessions *A. ostoyae*: FUEG01000001-FUEG01000106, *A. cepistipes*: FTRY01000001-FTRY01000182, *A. gallica*: NKEW000000000 and *A. solidipes*: NKHM000000000. Gene Expression Omnibus (GEO) archive of the sequenced *A. ostoyae* libraries was deposited in NCBI's GEO Archive at <http://www.ncbi.nlm.nih.gov/geo> under accession GSE100213.

a

b

c

d

a

bioRxiv preprint doi: <https://doi.org/10.1101/166231>; this version posted July 20, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

b

c

d