Article type : MS - Regular Manuscript

Comparative genomics of Rhizophagus irregularis, R. cerebriforme, R. diaphanus and Gigaspora rosea highlights specific genetic features in Glomeromycotina

Emmanuelle Morin[1], Shingo Miyauchi[1], Hélène San Clemente[2], Eric CH Chen[3], Adrian Pelin[3], Ivan de la Providencia[3], Steve Ndikumana[3], Denis Beaudet[3], Mathieu Hainaut[4], Elodie Drula[4], Alan Kuo[5], Nianwu Tang[2], Sébastien Roy[6], Julie Viala[6], Bernard Henrissat[4,7], Igor V. Grigoriev[5], Nicolas Corradi[3*], Christophe Roux[2*], Francis M. Martin[1,8*]

[1] Institut national de la recherche agronomique, Université de Lorraine, Unité Mixte de Recherche Arbres/Microorganismes, Centre INRA-Grand Est-Nancy, 54280 Champenoux, France;

[2] Laboratoire de Recherche en Sciences Végétales, Université de Toulouse, UPS, CNRS 24 Chemin de Borde Rouge-Auzeville, 31320 Castanet-Tolosan, France;

[3] Department of Biology, University of Ottawa, Ottawa ON K1N 6N5, Canada;

[4] CNRS, UMR 7257, Aix-Marseille Université, 13007 Marseille, France ;

[5] US Department of Energy Joint Genome Institute (JGI), Walnut Creek, California 94598, USA;

[6] Agronutrition- rue Pierre et Marie Curie, Immeuble BIOSTEP 31670 Labège, France ;

[7] Architecture et Fonction des Macromolécules Biologiques, Aix-Marseille Université, 13007 Marseille, France;

[8] Beijing Advanced Innovation Center for Tree Breeding by Molecular Design, Beijing Forest University, 100080 Beijing, China.

ORCID :

Emmanuelle Morin	0000-0002-7268-972X
Shingo Miyauchi	0000-0002-0620-5547
Eric CH Chen	0000-0001-5367-7259
Elodie Drula	0000-0002-9168-5214
Bernard Henrissat	0000-0002-3434-8588

Igor V. Grigoriev 0000-0002-3136-8903
Nicolas Corradi 0000-0002-7932-7932
Christophe Roux 0000-0001-5688-5379
Francis M. Martin 0000-0002-4737-3715

*Joint senior authors
Authors for correspondence: Francis Martin: Tel. +33 383 39 40 80; Email: francis.martin@inra.fr; Christophe Roux: Tel. +33 5 34 32 38 04; Email: roux@lrsv.ups-tlse.fr; Nicolas Corradi: Tel: +1 613 562 58 00; Email: ncorradi@uottawa.ca

Received: 23 October 2018
Accepted: 28 December 2018

Total word count (excluding summary, references and legends): 7,102
Abstract: 196 / Introduction: 1,041 / Material and Methods: 2,091 / Results: 2,361 / Discussion: 1,286 / Acknowledgements: 127
6 color figures (Fig.1, Fig.2, Fig.4, Fig.5, Fig. 6) / 1 table
Supporting Information (6 figures; 17 tables)

Summary

• Glomeromycotina is a lineage of early diverging Fungi establishing arbuscular mycorrhizal (AM) symbiosis with land plants. Despite their major ecological role, genetic bases of their obligate mutualism are largely unknown, hindering our understanding of their evolution and biology.

• We compared the genomes of Glomerales (Rhizophagus irregularis, Rhizophagus diaphanus, Rhizophagus cerebriforme) and Diversisporales (Gigaspora rosea) species, together with those of saprotrophic Mucoromycota, to identify gene families and processes associated with these lineages and to understand the molecular underpinning of their symbiotic lifestyle.

• Genomic features in Glomeromycotina appear to be very similar with a very high content in transposons and protein-coding genes, extensive duplications of protein kinase genes, and loss of genes coding for lignocellulose degradation, thiamin biosynthesis and cytosolic fatty acid synthase. Most symbiosis-related genes in R. irregularis and G. rosea are specific to Glomeromycotina. We also confirmed that the present species have a homokaryotic genome organization.

• The high interspecific diversity of Glomeromycotina gene repertoires, affecting all known protein domains, as well as symbiosis-related orphan genes, may explain the known adaptation of Glomeromycotina to a wide range of environmental settings. Our findings contribute to an increasingly detailed portrait of genomic features defining the biology of AM fungi.

Key words: arbuscular mycorrhizal fungi, carbohydrate-active enzymes, fungal evolution, interspecific variation, protein kinases, transposable elements.

Introduction

The Glomeromycotina is a division of early diverging Fungi (Mucoromycota sensu Spatafora et al., 2016) with 315 described species (www.amf-phylogeny.com/amphylo_species.html). Members of this sub-phylum are able to establish AM symbiosis in association with 71% of land plants (Brundrett and Tedersoo, 2018). The mutualistic relationship established by AM fungi has a substantial impact on growth, development and ecological fitness of plants in natural and agricultural ecosystems (van der Heijden et al., 2015). This symbiotic association emerged over 410 million of years ago (Mya) (Strullu-Derrien et al., 2018), and is considered ancestral in land plant evolution (Spatafora et al., 2016; Martin et al., 2017; Field & Pressel, 2018). It is thought that obligate mutualistic Glomeromycotina derived from saprotrophic ancestors from the Mucoromycota lineage (Spatafora et al., 2016), although different views on the appropriate taxonomic rank of AM fungi are currently present in the research community. Here, we consider that Mucoromycota comprises Glomeromycotina, Mortierellomycotina, and Mucoromycotina and is sister to Dikarya (Spatafora et al., 2016). Despite the fact that the first AM symbionts originated >410 Mya, features of their genomes can be reconstructed through phylogenetically-informed comparisons among extant symbiotic Glomeromycotina and saprotrophic Mucoromycota. To harness this potential, genome sequences of divergent Glomeromycotina species with different life histories are needed. To date, only three species of Glomeromycotina have their genome published, namely Rhizophagus irregularis (Schenck & Sm.) Walker & Schüßler (Tisserant et al., 2013; Li et al., 2014; Chen et al., 2018; Maeda et al., 2018), R. clarus (Nicolson & Schenck) Walker & Schüßler (Kobayashi et al., 2018) and Diversispora epigaea (Daniels & Trappe) Walker & Schüßler (formerly Glomus versiforme) (Sun et al., 2018), meaning that the gene repertoires of most species of Glomeromycotina have yet to be sequenced, analyzed and compared.

The strain DAOM197198 of R. irregularis was the first Glomeromycotina genome to be sequenced (Tisserant et al., 2013, Lin et al., 2014). This genome showed that R. irregularis has substantial phylogenetic relationships with saprotrophic Mortierellomycotina and shares several genetic and metabolomic features with early diverging fungi in Mucoromycotina (Tisserant et al., 2013; Chang et al., 2015;

Spatafora et al., 2016; Uehling et al., 2017). It also provided unprecedented insights into molecular bases of the AM symbiosis, sexual reproduction and physiology in an iconic representative of Glomeromycotina. The DAOM197198 genome is homokaryotic with a low nucleotide sequence polymorphism, and one of the largest fungal genomes, with an unusually high content of transposable elements (TE) and a strikingly high number of gene duplications (Tisserant et al., 2013, Lin et al., 2014). DAOM197198 experienced the loss of several, otherwise widely conserved Mucoromycotina genes with functions related to cell wall polysaccharide degradation, and overall primary and secondary metabolism which could explain its obligate biotrophy. These features have recently been corroborated by the sequencing of five additional isolates from *R. irregularis* (Chen et al., 2018), *R. clarus* (Kobayashi et al., 2018) and *D. epigaea* (Sun et al., 2018). Most importantly, no gene encoding multidomain de novo fatty acid synthase was detected in the genome of these species as initially suggested by Wewer and co-authors (2014) based on the analysis of *R. irregularis* gene repertoire. Esterified palmitic acid is transferred from plant roots to symbiotic mycelium and this lipid export pathway, together with soluble carbohydrates, contributes a substantial amount of carbon to symbiotic hyphae of *R. irregularis* (Bravo et al., 2017; Luginbuehl et al., 2017).

Isolates of *R. irregularis* harvested from the same field harbor a very large variability in their gene repertoire affecting most known cellular and biochemical functions, as well as putative mycorrhiza-induced small secreted effector-like proteins (MiSSPs) and other differentially expressed symbiotic genes with no known function (Chen et al., 2018). High variability is also found in active transposable elements. These findings indicate a substantial divergence in the functioning capacity of *R. irregularis* isolates, and as a consequence, their genetic potential for adaptation to biotic and abiotic changes.

Although transcriptomic assemblies were recently obtained from a number of AM fungi (Salvioli et al., 2016; Tang et al., 2016; Beaudet et al., 2018), our view of the genomic features of Glomeromycotina subphylum is still highly biased by the fact that they have been obtained with species that shared a last common ancestor with other AM relatives many millions of years ago. As of today, molecular bases of genomic adaptations that facilitated evolutionary processes to the obligate symbiotic lifestyle

throughout the Glomeromycotina phylum are unknown and can only be elucidated by using additional full genome sequences from various clades of AM fungi. Acquiring genomic information from additional Glomeromycotina species is also needed to corroborate their genomic idiosyncrasies, the high intraspecific genome diversity found in R. irregularis (Chen et al., 2018) and its impact on species delimitation (Bruns et al., 2018).

In the present study, we provide a comparative analysis of four genomes of Glomeromycotina symbionts, namely R. irregularis DAOM197198, R. diaphanus (Morton & Walker) Walker & Schüßler MUCL43196, R. cerebriforme DAOM227022 in Glomales and Gigaspora rosea Nicolson & Schenck DAOM194757 in Diversisporales. Genomes of R. diaphanus, R. cerebriforme and Gigaspora rosea have been sequenced and annotated for this study as they belong to the more diversified Glomeromycotina clades (Redecker et al., 2013) and they also present contrasted developmental, ecological and symbiotic traits (Bonfante and Genre, 2008). Our aims are to assess whether the known genome features of R. irregularis are shared by other clades of AM fungi and to provide new insights into the evolutionary genome dynamics of the genome in the ancestral lineage leading to Glomeromycotina at two broad levels: gene family origin and diversification, and conservation of gene repertoire features. Our analysis focuses on inter-species genome diversity in key gene categories involved in symbiosis development and functioning and differential gene family expansion and contraction. We also confirm the occurrence of genes potentially related to mating in these supposedly ancient clones and a low genetic diversity among their co-existing nuclei. Comparison of AM fungal genomes with those of Mortierella elongata (Mortierellomycotina) and representative Mucoromycotina species indicates extensive copy number variations in genes involved in nutrient acquisition, developmental pathways, and primary and secondary metabolism. This study, together with the recent analyses of Chen et al. (2018), Kobayashi et al. (2018), Maeda et al. (2018) and Sun et al. (2018), have expanded and refined our understanding of the genomic heritage of AM symbionts.

Methods and Materials

Production of fungal materials

Spores and mycelium of R. irregularis DAOM197198 (aka DAOM181602) and G. rosea DAOM194757, produced on carrot root organ cultures, were obtained from Agronutrition (Labège, France). Carrot root organ cultures of R. diaphanus MUCL 43196 and R. cerebriforme DAOM227022 were obtained from the Glomeromycota in vitro Collection (GINCO) located at Agriculture Canada (Ottawa, Canada).

De novo genome assembly

High molecular weight genomic DNA of R. irregularis, R. diaphanus, R. cerebriforme and G. rosea was extracted from large amounts of mycelium produced on carrot root organ cultures as described in Tisserant et al. (2013) and Ropars et al. (2016). DNA was used to construct paired-end (2 x 125 bp) TruSeq Nano libraries and mate-pair libraries (with insert sizes of 3 and 8 kbp) using Nextera Mate Pair Sample Prep Kit. Libraries were sequenced using the Illumina HiSeq 2500 platform (Illumina, Inc., San Diego, CA, USA) at the GeT-PlaGe sequencing facility (Toulouse, France). Low quality sequences and sequencing adapters were trimmed from the raw Illumina reads using Trimmomatic (Bolger et al., 2014). The adapter sequences on mate-pair sequences were removed using the software Nextclip with default parameters (Leggett et al., 2014). Sequences were assembled using AllPathsLG version 43460 (Gnerre et al., 2011) as described in Chen et al. (2018). Scaffolds were queried against the NCBI's nonredundant nucleotide database by using BLASTn and sequences with >90% identity and 75% coverage to plant or bacterial sequences were considered as contaminants and removed. Sequences with a GC%>45 were also considered as bacterial contaminants and discarded (Tisserant et al., 2013). The putative MAT-loci of Paraglomus sp., Claroideoglomus claroideum, Gigaspora rosea, Scutellospora castanea and Glomus macrocarpum were identified along preliminary genome surveys of these species using reciprocal BLAST procedures (Ropars et al. 2016).

Genome annotation

Gene prediction and functional annotation (Gene Ontology (GO), Eukaryotic Orthologous Groups of Proteins (KOG), Kyoto Encyclopedia of Genes and Genomes (KEGG), proteases (MEROPS database) have been carried out using the Joint Genome Institute (JGI) Annotation Pipeline. This bioinformatic pipeline detects and masks repeats and transposable elements (TE), predicts genes, characterizes each conceptually translated protein, chooses a best gene model at each locus to provide a filtered working set, clusters the filtered sets into draft gene families and creates a JGI Genome Portal at the MycoCosm database with tools for public access and community-driven curation of the annotation (Grigoriev et al., 2014). The quality of the draft assemblies was evaluated by using conserved fungal proteins with Benchmarking Universal Single-Copy Orthologs (BUSCO version 3.0.2; Simão et al., 2015). We used default parameter values, the fungal BUSCO set (Fungi odb9 gene set; http://buscodev.ezlab.org/datasets/fungiodb9.tar.gz), and performed searches with HMMER version 3.1. Carbohydrate-active enzymes, so-called CAZymes, including glycoside hydrolases (GH), glycosyl transferases (GT), polysaccharide lyases (PL), carbohydrate esterases (CE), enzymes that act in conjunction with other CAZymes (Auxiliary activities, AA), carbohydrate-binding modules (CBM) and enzymes distantly related to plant expansins (EXPN), were identified using the CAZy database (www.cazy.org) annotation pipeline (Lombard et al., 2014). Secreted proteins were identified using a custom pipeline including SignalP v4, WolfPSort, TMHMM, TargetP, and PS-Scan algorithms as reported in Pellegrin et al. (2016).

Prediction of transposable elements (TE) was carried out as described in Payen et al. (2017). De novo repeat sequences were identified in unmasked genome assemblies, downloaded from JGI MycoCosm (Grigoriev et al., 2014), using RepeatScout 1.0.5 with default parameters (sequences ≥50 bp, ≥10 occurrences) (Price et al., 2005). Filtered sequences were annotated by searching homologous sequences against the fungal references in RepBase version 22.08 (http://www.girinst.org/server/RepBase/index.php) using tBLASTx (Altschul et al., 1990). The coverage of TE, including unknown categories, in genomes was estimated by masking the genome assemblies using RepeatMasker open 4.0.6 (http://www.repeatmasker.org). Output files generated from the procedures above were

integrated, the genome size and repeat element coverage were calculated, and the results were visualised using a set of custom R scripts named Transposon Identification Nominative Genome Overview (TINGO) (available on request).

The putative MAT-loci have been deposited in GenBank and are available under the accession numbers MH445370 to MH445379. The new genome assembly and annotation from R. irregularis DAOM197198 have been published in Chen et al. (2018), whereas genome assemblies from R. diaphanus MUCL43196 and R. cerebriforme DAOM227022 have been published in Ropars et al. (2016). The genome assembly of G. rosea DAOM194757 has been produced for this study.

RNA extraction, sequencing and expression analysis

For gene expression profiling, all biological samples were produced in triplicates. Spores of R. irregularis DAOM197198 and G. rosea DAOM194757 were germinated during seven days in liquid M medium (Bécard and Fortin, 1988) in the dark at 30°C with 2% CO_2. Transcripts from these germinating hyphae were used as reference (non-symbiotic control) for calculating the gene expression ratio. Intraradical mycelium of R. irregularis and G. rosea colonising Brachypodium distachyon genotype Bd21 were collected from pot cultures (see Kamel et al. (2017) for details).

Total RNA extraction, sequencing procedure and expression analyses were performed according to Tisserant et al. (2011) for R. irregularis, and Tang et al. (2016) and Kamel et al. (2017) for G. rosea. In brief, one to three µg of total RNA was extracted from germinating hyphae and mycorrhizal roots using the RNeasy Plant Mini RNA Extraction Kit (Qiagen, Germany) and stored at -80°C until further analysis. cDNA library construction and sequencing were performed at the GeT-PlaGe sequencing facility according to standard Illumina protocols. Bioinformatic procedures for transcript profiling were detailed in Kamel et al. (2017): trimmed paired-end reads were mapped onto predicted genes from R. irregularis (genome assembly Rhiir2_1) and G. rosea (genome assembly Gigro1) using CLC Genomics Workbench (Qiagen) with stringent settings (similarity and length read mapping criteria at 98% and 95%, respectively). Total mapped paired-end reads for each gene were calculated and total read counts were normalized as fragments per kilobase of gene model per million fragments mapped (FPKM). Detailed description of the RNA-Seq analysis (i.e.,

specifying reads and reference, defining read mapping options, calculating expression values) can be found in the CLC Workbench online manual at: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/950/index.php?manual=RNA_Seq_analysis.html). FPKM from genes expressed in intraradicular mycelium were compared to those of germinating hyphae as a reference. Fold-change values were calculated by proportion-based test statistics (Baggerly et al., 2003) with a False Discovery Rate (FDR) correction for multiple testing (Benjamini et al., 1995). For the present study, we used very stringent parameters and retained only genes showing an expression >5-fold higher in intraradical mycelium compared to germinating hyphae (FDR $\leq$ 0.05). Among the 26,183 high-confidence genes predicted in R. irregularis, 17,876 were expressed in hyphae from germinating spores and 12,890 in roots of B. distachyon. Among the 31,291 high-confidence genes predicted in G. rosea, 13,987 genes were expressed in hyphae from germinating spores and 11,896 genes were expressed in roots of B. distachyon. The high number of genes expressed in hyphae from germinating spores indicates that this non-symbiotic mycelium was transcriptionally very active and thus, can be used as an appropriate control in transcriptome comparisons. Here, fungal genes showing a higher expression in symbiotic roots compared to germinating hyphae, referred to as symbiosis-related genes. They consist in all genes involved in fungal development and metabolism in plant tissues. These so-called symbiosis-related genes are candidate genes for further functional analyses of symbiotic functions.

Detailed information on the protocols and data are available at National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) portal (accession numbers: GSE67906 to GSE67911 for G. rosea and GSE67913 to GSE67926 for R. irregularis).

Protein orthology

To assess the orthology between gene sets from the eight species of Mucoromycota sensu Spatafora et al. (2016), we downloaded gene models of Mortierella elongata AG-77 v2.0 (Uehling et al., 2017), Mucor circinelloides CBS277 v2.0 (Corrochano et al., 2016), Phycomyces blakesleeanus NRRL1555 v2.0 (Corrochano et al., 2016) and Rhizopus microcarpus ATCC52814 v1.0 (Lastovetsky et al., 2016) from JGI

MycoCosm database. We clustered the predicted proteins of these taxa, together with the present Glomeromycotina predicted proteins, with FastOrtho using 50% identity and 50% coverage (Wattam et al., 2013). We selected the latter parameters because compared fungal species are highly divergent. We discussed the protein families (orthogroups) in expansion in each species relative to the other species only when the differences were statistically supported (Wattam et al., 2013). Based on this clustering, we determined (1) the set of predicted proteins shared by the eight species (i.e., core genes), (2) sets of predicted proteins encoded in at least two genomes (i.e., dispensable genes), and (3) sets of predicted proteins unique to a genome (i.e., species-specific genes, which are also referred to as taxonomically restricted genes). To define sets of species-specific genes in a broader context, we also searched for orthologous sequences (50% identity and 50% coverage) in a wider set of genomes, ca. all fungal genomes publicly available at MycoCosm. For each gene sets, we also identified duplicated genes. Note that genes families were also automatically clustered by the JGI prediction pipeline and the clusters can be visualized, ranked and compared at the 'CLUSTERS' page of the JGI Glomeromycotina genome portals, e.g. https://genome.jgi.doe.gov/clm/run/Rhiir2_1-FM2-Glomeromycota-only.2100;Pe_ufO?organism=Rhiir2_1. Multigene families were analysed for statistically significant evolutionary changes in protein family size using the CAFE and FastOrtho programs (Han et al., 2013; Wattam et al., 2013) with default parameters. The genomes of R. clarus (Kobayashi et al., 2018) and D. epigaea (Sun et al., 2018) were not included in these analyses, as they were publicly released after this manuscript submission.

Phylogenomic analysis

A phylogenomic tree was constructed using the eight above mentioned Mucoromycota genomes and four outgroup genomes. We identified 784 gene clusters with only one protein-coding gene per species by clustering protein sequences using FastOrtho (Wattam et al., 2013) with the following parameters: 50% identity and 50% coverage. Each cluster was then aligned with MAFFT 7.221 (Katoh & Standley, 2002), and ambiguous regions (containing gaps and poorly aligned) were eliminated and single-gene alignments were concatenated with GBLOCKS 0.91b (Castresana, 2000). A

maximum likelihood inference for our phylogenomic dataset was achieved with RAxML 7.7.2 (Stamatakis, 2014) using the standard algorithm, the PROTGAMMAWAG model of sequence evolution and 1000 bootstrap replicates.

dN/dS calculation

For this analysis, we only used gene nucleotide sequences defined as 1-1-1-1-1 orthologues by FastOrtho. Average dN and dS values for predicted transcripts were calculated using the BioPerl's DNAStatistics package (Stajich et al., 2002). The package uses a simple count method for dN and dS calculations, which is sufficient for our purposes of finding divergent orthogroups in the clusters as defined above. Alignments were confirmed by visual inspection. To ensure a conserved analysis, poorly aligned loci were discarded from the average dN and dS analysis and the final results were plotted using the R program. A list of the genes showing evidence of rapid sequence evolution in the Glomeromycotina genomes and their putative function (KOG definition) can be found in the Supporting Information.

Single nucleotide polymorphism (SNP)

For SNP calling and allelic frequencies plots, each set of paired-end and mate-pair data sets used for Glomeromycotina assemblies was mapped independently against the respective corresponding reference genome assemblies downloaded from the JGI portals using the Burrows–Wheeler Alignment (BWA) tool, with the BWA-MEM algorithm (Li & Durbin, 2009). The mapping tool is specifically designed for sequences ranging from 70 bp to 1 Mbp and is recommended for high-quality queries. SAMtools (Li et al., 2009) was then used to convert SAM files into sorted BAM files and to merge the different data sets of the same species together, to obtain a single sorted BAM file for each isolate. SNPs were called using FreeBayes v0.9.18-3-gb72a21b (Garrison & Marth, 2012), with the following parameters: -K (that is, output all alleles that pass input filters), excluding alignments with mapping quality less than 20 (-m 20) and taking into account only SNPs with at least two alternate reads (-C 2). SNPs were filtered to avoid the analysis of false positives (that is, SNPs originating from misalignment and/or paralogy) using vcffilter from the vcf-lib library according to (1) the read depth (maximum read depth: DP < 1.25× genome mean coverage; minimum

read depth: DP > 0.75× genome mean coverage), (2) the type of SNPs (only considering SNPs, not indels: TYPE = snp), (3) considering only one alternative allele (NUMALT = 1) and (4) the reference allele observation (RO > 1).

Data availability

Full genome, predicted gene and transcript sequences of R. irregularis DAOM197198, R. diaphanus MUCL 43196, R. cerebriforme DAOM 227022 and G. rosea DAOM194757 can be accessed at:
https://genome.jgi.doe.gov/Rhiir2_1/Rhiir2_1.home.html;
https://genome.jgi.doe.gov/Rhidi1/Rhidi1.home.html;
https://genome.jgi.doe.gov/Rhice1_1/Rhice1_1.home.html;
https://genome.jgi.doe.gov/Gigro1/Gigro1.home.html.

Genomic resources are also available at GenBank under the following accession numbers: R. irregularis DAOM197198 version 2, AUPC02000000/PRJNA208392; R. diaphanus, QKKE01000000/PRJNA430014; R. cerebriforme, QKYT01000000/PRJNA430010; and G. rosea, QKWP01000000/PRJNA430513.

Results

General genome features and phylogeny

The nuclear genomes of R. irregularis, R. diaphanus and R. cerebriforme in Glomerales and Gigaspora rosea in Diversiporales were sequenced and assembled. They ranged from 126 to 598 Mbp with an estimated content of 21,549 to 31,291 protein-coding genes (Table 1, Supplementary Table S1 in Supporting Information). Glomeromycotina genomes are significantly larger than saprotrophic Mucoromycota genomes (Table 1). No evidence of whole genome duplication events (i.e., no segmental duplications) was found (see Synteny tools on the JGI portals) and this larger size is mainly driven by TE proliferation. TE content ranges from 20% (R. diaphanus) to 63% (G. rosea) of total assemblies (Table 1, Fig. 1). However, the exact repetitive fraction of G. rosea and Rhizophagus genomes is likely larger; their highly repetitive nature (Fig. 1) has contributed to the assembly fragmentation, hindering the annotation of an unknown TE proportion. The distribution of TE categories notably varies between G. rosea and Rhizophagus spp. (Fig. 1), with the former harbouring a larger genome coverage of

Gypsy LTR, Tad1, hAT and Mariner/Tc1. The number of Penelope retroelement copies in G. rosea is >3,900, whereas only 206 copies are found in R. cerebriforme and none in R. irregularis and R. diaphanus, indicating that invasions by different types of TE took place independently in different AM fungi. There are hints of older TE propagation events in the four genomes with a long tail of low similarity TE copies (data not shown). As a result of massive TE proliferations, Glomeromycotina genomes show a very high level of structural rearrangements and a macrosynteny was only observed between R. irregularis and R. diaphanus (Supplementary Fig. S1), consistent with their close phylogenetic proximity.

Over 97% of a benchmark set of conserved fungal BUSCO genes, a proxy to genome completeness (Simão et al., 2015), were found in Glomeromycotina assemblies (Supplementary Table S2) and up to 94% of RNA-Seq reads from fungal libraries mapped to the gene repertoire (see Info page on JGI genome portals), indicating that assembled genomes capture most of the coding gene space.

To have a robust phylogenetic framework for our comparative analyses, we investigated phylogenetic relationships between the sequenced Glomeromycotina and other Mucoromycota. A phylogeny based on a concatenation of 784 single copy, orthologous protein sequences (Fig. 2) strongly supports the erection of Mucoromycota to unite Glomeromycotina, Mortierellomycotina and Mucoromycotina (Spatafora et al., 2016; Uehing et al., 2017).

Glomeromycotina-specific gene families: gains and losses

We compared the gene repertoires encoded by sampled Mucoromycota taxa and identified sub-phylum- and species-specific gene families that might contribute to genome trait diversification. We separately clustered predicted protein sequences of either the four species of Glomeromycotina or the eight species of Mucoromycota to infer orthologous gene groups (orthogroups) (Wattam et al., 2013). We then identified (i) sets of core genes shared by all Mucoromycota or all Glomeromycotina species; (ii) sets of dispensable genes shared by at least two species of Mucoromycota or Glomeromycotina; (iii) sets of species-specific genes only found in a single genome (Fig. 3, Supplementary Table S3). For each category, we also identified single copy and duplicated genes. As expected for species that diverged >450 million years ago

(Uehling et al., 2017), clustering the predicted protein sequences of the eight Mucoromycota led to a very restricted core set of genes (Fig. 3, Supplementary Table S3). In the other hand, we identified 5,463 to 5,703 conserved (core) genes, 24 to 27% of them being duplicated genes, in Glomeromycotina species (Supplementary Table S3). Each AM species is characterized by a large set of species-specific genes, which are also referred to as taxonomically restricted genes. Within this context, the very high proportion of species-specific-genes in G. rosea (64%) with a higher frequency of multi-allelic copy numbers (Fig. 3, Supplementary Table S3) is intriguing and partly reflects the large taxonomic divergence between this taxon and those sequenced so far. These sets of Glomeromycotina species-specific genes are noticeably distinct as they have a shorter gene size, fewer exons, and a lower proportion of expressed sequences than the conserved genes (Supplementary Table S4), suggesting they might be evolutionarily young genes.

The expansion and contraction of gene families (i.e., orthogroups) in the different lineages of Mucoromycota were determined by using the gene family modeling pipeline CAFE (Han et al., 2013) (Supplementary Fig. S2) and FastOrtho (Supplementary Tables S5 to S8). Across the phylogeny, the number of orthologous gene families gained on Glomeromycotina and Mucoromycotina lineages, relative to their most recent common ancestor (MRCA), are in the same range, from 63 to 259 (<10% of the orthologous protein sets). Gene family loss was also rampant during the diversification throughout the Mucoromycota lineages and is larger in G. rosea.

It is noteworthy that several Glomeromycotina gene families are strikingly expanded (i.e., they contain a larger set of duplicated genes) or not shared with Mucoromycotina or M. elongata (duplicated species-specific genes in Fig. 3, Supplementary Tables S5 to S8). They include large gene families encoding protein domains related to signaling kinases, such as tyrosine kinase specific for activated GTP (p21cdc42Hs) and ubiquitination-associated BTB/POZ domain-containing proteins (Supplementary Table S9). Tyrosine kinases are often associated to Sel1 repeats which can serve as adaptor proteins for the assembly of macromolecular complexes under cellular stress (Mittl and Schneider-Brachert, 2007).

Hierarchical clustering of the presence and abundance of the different Pfam protein domains found in the genomes of Mucoromycota species (this study) and R. irregularis

isolates (Chen et al., 2018) (Fig. 4A) identified genome-wide patterns of functional domain content among these fungi. Glomeromycotina clustered together, whereas Mucoromycotina species clustered with M. elongata. Among Glomerales, R. diaphanus was closely related to the five sequenced R. irregularis isolates, whereas R. cerebriforme displays a substantial divergence in its Pfam domain distribution. Although clustering with Glomerales, G. rosea displayed a Pfam domain distribution pattern very different from these species, pointing to large differences in metabolic, developmental and signalling pathways between AM fungi. Pfam categories showing a substantial differential abundance contain genes encoding transcriptional factors, e.g. Myb proteins and DNA polymerase, but also key factors involved in cell structure, such as adaptins and kinesins. In Rhizophagus spp., the distribution of Pfam domains corroborated the higher occurrence of proteins predicted to have a role in signaling pathways and protein-protein interactions (see above, Tables S5, S6, S7 and S9). The G. rosea gene set is enriched in AMP-binding and tetratricopeptide repeat region (TPR)-domain containing proteins, H^+-ATPases, NUDIX hydrolases, aspartyl proteases, cytochromes P450, and methyltransferases (Fig. 4A).

The functional genomic comparison made through KEGG pathway profile correlations (Fig. 4B) also showed that sequenced Glomeromycotina present a higher metabolic similarity between taxa compared to M. elongata/Mucoromycota species, corroborating and extending M. elongata genome analysis (Uehling et al., 2017). Lack of PCWDE (see below), degradation of sucrose and glycogen (i.e. invertase, glucoinvertase, glucoamylase), biosynthesis of polyketides, nonribosomal peptides, thiamin and biosynthesis of fatty acids (i.e. palmitic acid through type I fatty acid synthase) are among the most noticeable metabolic idiosyncrasies of Glomeromycotina (see KEGG comparative tool on JGI portals).

A substantial proportion (44 to 47%) of predicted Glomeromycotina genes have no sequence similarity with documented proteins in MycoCosm (Fig. 3), Pfam (Supplementary Table S1) or Eukaryotic Orthologous Groups of Proteins (KOG) databases (data not shown).

Nucleotide sequences of Glomeromycotina orthologous genes (1-1-1-1-1) were aligned to identify any evidence of accelerated sequence evolution, assuming that increased sequence divergence results from positive selection, possibly caused by

environmental pressures (Supplementary Fig. S3). The analysis revealed that most of the 100 orthologous genes showing the highest sequence divergence (i.e. red dots in Supplementary Fig. S3) encode for proteins with unknown function (Supplementary Table S10). Orthologues with putative function are involved in a large variety of biologically unrelated functions and pathways and, for example, include HMG-box transcription factors, RNA polymerases, as well as protein required for meiotic chromosome segregation (KOG2513), or mitochondrial Fe/S cluster exporters.

A restricted set of genes involved in lignin and polysaccharide degradation

The sequenced Glomeromycotina species share a limited repertoire of genes coding for secreted plant cell wall degrading enzymes (PCWDE) (Fig. 5, Supplementary Fig. S4, Supplementary Table S11). No gene encoding lignin peroxidases (AA2), cellobiohydrolases (GH6, GH7), polysaccharide lyases (PL1, PL3, PL4, PL9), lytic polysaccharide monooxygenases acting on cellulose or cellulose-binding-, carbohydrate-binding module 1 (CBM1) are encoded by sequenced Glomeromycotina genomes. In Rhizophagus spp., only a single endo-ß-1,4-endomannanase (GH5_27) is possibly acting on hemicellulose in plant cell walls. The secreted polysaccharidases annotated in Rhizophagus species are mostly acting on fungal polysaccharides (chitooligosaccharide oxidase AA7, chitin deacetylase CE4, chitinase GH18 and α-N-acetylgalactosaminidase GH27) or bacterial peptidoglycans (lysozyme GH25) (Fig. 5, Supplementary Fig. S4, Supplementary Table S11). The only carbohydrate-binding modules are chitin-binding modules (CBM18, CBM19). Remarkably, the distribution of several CAZyme families strikingly differ in G. rosea compared to Rhizophagus species, i.e., higher copy number of laccase (AA1) possibly acting on polyphenolic compounds, cellobiose dehydrogenase AA3, chitooligosaccharide oxidase AA7, chitinases GH18, α-N-acetylgalactosaminidase GH27, mannosyl-oligosaccharide α-1,2-mannosidase GH92, galactoside α-1,3/1,4-L-fucosyltransferase GT10, lipopolysaccharide β-1,4-galactosyltransferase GT25 and carbohydrate-binding modules binding to chitin (CBM14, CBM18) in G. rosea (Fig. 5, Supplementary Fig. S4, Supplementary Table S11).

Sexual reproduction

The Glomeromycotina genomes were also investigated for the presence of genomic signatures of sexual reproduction (Riley et al. 2013), particularly meiosis-specific genes (MSG), and for evidence of a homokaryotic/dikaryotic genetic organization; the latter being defined by the co-existence of one or two divergent putative mating-type MAT loci, as recently found in some R. irregularis isolates (Ropars et al., 2016). Our analyses are consistent with recent data based on analyses of R. irregularis assemblies (Ropars et al., 2016, Chen et al., 2018), as we found that all genomes encode for a complete set of MSG (Supplementary Table S12A). Furthermore, all AM fungi in this study show intra-isolate genetic variation (0.23 to 0.36 SNP per kb) (Supplementary Table S12B) that are consistent with a homokaryotic genome organization with no evidence of dikaryosis. The distribution of DNA reads mapping on all bi-allelic SNP regions were assessed and we observed allele frequencies in agreement with haploid genome patterns (Supplementary Fig. S5).

Also consistent with the homokaryotic nature of these species, each one carried a single copy of a genomic region showing similarities with a MAT locus composed of two bi-directionally transcribed genes with homeodomain regions, with coiled-coil domains and nuclear localization signals. We took advantage of the newly available genome data to determine whether the locus is structurally conserved across the AM fungal phylogeny, an indication that conservation in gene order is functionally important for the locus. Our analyses showed that the putative AM fungal MAT-locus is conserved in structure across most Glomeromycotina species investigated to date with the exception of G. rosea, including the basal genera Claroideoglomus and Paraglomus (Supplementary Fig. S6). It also shows substantial sequence divergence among the species investigated, as expected for bone-fide MAT-loci. The absence of structural conservation of HD-1-like and HD-2 genes in Gigasporaceae stands out, particulary given that the locus is conserved in other members of the Diversisporaceae; namely D. epigaea.

Secretome and candidate effectors in Glomeromycotina

G. rosea and M. elongata have the largest repertoire of secreted proteins, whereas other species have similar sets of secreted proteins, such as CAZymes, proteases and lipases

(Supplementary Table S13). R. irregularis presents a larger repertoire of small secreted proteins (SSP) compared to other Glomales. Among 436 orthogroups coding for SSPs, 250 are specific to Glomeromycotina species, while 45 are specific to G. rosea and 138 only represented in Rhizophagus species (Supplementary Table S14), confirming that AM fungi have substantial species-specific repertoire of SSPs.

Conservation of symbiosis-related transcriptional signature within Glomeromycotina

The expression of R. irregularis genes was measured by RNA-Seq profiling in B. distachyon mycorrhizae; 426 R. irregularis transcripts (3.3% of the expressed genes) are expressed at a higher level in symbiotic roots compared to transcriptionally-active germinating hyphae (Supplementary Table S15A). These transcripts are potentially involved in the development and physiology of the symbiotic interaction. We assessed the evolutionary conservation of these symbiosis-related transcripts among Mucoromycota (Fig. 6A). We found that only 16 % of R. irregularis symbiosis-upregulated genes are shared by all species of Mucoromycota (cluster VIII). Most of them are coding for core metabolic functions. In addition, most transcripts from cluster IV have orthologous sequences in Glomeromycotina and one or several species of saprotrophic Mucoromycota. On the other hand, only seven % of R. irregularis symbiosis-induced genes are species-specific (cluster VII), i.e. not even shared with its closest taxa, R. diaphanus. Most of these genes code for proteins with unknown KOG functions and mycorrhiza-induced small secreted proteins (MiSSPs). Cluster II (8%) grouped R. irregularis symbiosis-induced genes, mainly coding for unknown proteins and MiSSPs, having a strong similarity with R. diaphanus, its closest relative. Transcripts of clusters III and V (36%) are shared by the Glomeromycotina species, while those of cluster VI (7%) are only encoded by G. rosea and R. cerebriforme. As expected, sequence conservation reflects the phylogenetic distance between taxa, e.g., 80 % of R. irregularis symbiosis-induced genes are found in R. diaphanus with a high sequence similarity (> 80 %). A substantial proportion of these Glomeromycotina-conserved, symbiosis-related genes have no known function. However, among genes conserved in G. rosea and Rhizophagus species (clusters III, IV, V and VIII), several are involved in primary metabolism, e.g. nitrogen and carbon assimilation, membrane transport, signaling pathways (Supplementary Table S15A). Genes putatively involved

in detoxification mechanisms are also widely represented, e.g. cytochrome P450, UDP-glucuronosyl transferase, glutathione-S-transferase and pleiotropic drug resistance proteins (PDR1-15).

Analysis of the differential gene expression during the G. rosea/B. distachyon interaction identified 989 G. rosea genes (8.3% of the expressed genes) having a higher expression in symbiotic tissues compared to germinating hyphæ (Supplementary Table S15B). We investigated the evolutionary conservation of these transcripts enriched in symbiotic tissues among Mucoromycota (Fig. 6B). Intriguingly, a larger proportion (48%, clusters V, VI and VII) of symbiosis-related G. rosea genes are conserved in the eight Mucoromycota species compared to R. irregularis. Most of them are involved in cellular and signaling processes, and metabolism. Fourteen % of symbiosis-upregulated genes (cluster IV) are specific to G. rosea, coding for proteins of unknown KOG function and MiSSPs (Fig. 6B, Supplementary Table S15B). Similarly, G. rosea symbiosis-upregulated genes shared with Rhizophagus species (cluster II) are coding for proteins of unknown KOG function and MiSSPs.

Discussion

In the present study, we investigated the evolutionary dynamics of key genomic traits in the subphylum Glomeromycotina of Mucoromycota (Spatafora et al., 2016). Our enhanced AM fungal taxon sampling, including three newly annotated genomes (R. diaphanus, R. cerebriforme, G. rosea) and an improved R. irregularis DAOM197198 assembly and annotation (Chen et al., 2018), allows us to perform both within- and across-lineage comparisons, thus covering the different time scales at which the evolution of genome features occurred. In addition, this comparative genomic study provides further insights on the gene repertoires of AM fungi. Overall, our findings show that extant Glomeromycotina genomes have been shaped by both retention of ancestral states present in saprotrophic Mucoromycota and secondary innovations, for the multiple genomic traits investigated in the present study, namely genome size, protein domain diversity and gene content.

Genomic features (e.g. genome size, gene number, TE content) are highly similar within Glomerales. In contrast, G. rosea genome is much larger (>600 Mb) with a larger coding space and higher TE content. Previously, our knowledge on AM fungal

genomics was limited to the genus Rhizophagus, mainly the model fungus R. irregularis (Tisserant et al., 2013; Li et al., 2014; Ropars et al., 2016; Chen et al., 2018; Maeda et al., 2018). Although the transcriptome of G. rosea (Tang et al., 2016) and G. margarita (Salvioli et al., 2016) have been sequenced, the genome of these representatives of the Diversiporales was not sequenced. Therefore, the present study improves our knowledge on genomics and evolutionary biology of AM fungi by including genome information on G. rosea.

Our findings reveal a remarkable convergence in genome evolution in Glomerales and Diversiporales with massive accumulation of TE, extensive gene duplications in species-specific families and signaling pathways, but also losses of genes related to saprotrophism in Mucoromycota. We identified large sets of Glomeromycotina-specific genes by comparing Mucoromycota genomes, though most of them are coding for proteins with unknow function, such as MiSSPs. Gene families in expansion that originated in lineages leading to extant AM fungal species and genes specific to the Glomeromycotina subphylum are thought to operate in pathways or developmental processes, e.g. symbiotic interactions, that distinguish AM fungi from other Mucoromycota. Confirmation of this contention will require further large scale functional analyses.

We also showed the consistent lack of enzymes involved in plant cell wall degradation, thiamin biosynthesis, and cytosolic fatty acid synthesis in the four Glomeromycotina genomes, the consistent presence of genes involved in sexual reproduction in the four genomes, genus-specific sets of small secreted proteins that may play a role in symbiont recognition and accomodation. A low proportion (16%) of genes upregulated in symbiotic tissues are conserved in Mucoromycotina genomes. Those conserved genes mainly encode for cellular and signaling processes, and pathways of the primary metabolism, and likely derived from those encoded by the saprotrophic MRCA. Several of these general genomic features have recently been confirmed in G. clarus (Kobayashi et al., 2018) and D. epigaea (Sun et al., 2018).

The very large sets of species-specific genes found in each clade of Glomeromycotina suggest that de novo gene construction followed by extensive gene duplications, and/or fast sequence evolution of pre-existing genes is a hallmark of the sampled AM fungal genomes. As these species-specific genes have no ortholog in other

sequenced taxa, they evolved independently in each AM species, i.e. they are not derived from ancestral saprotrophic Mucoromycota. This is particularly true for G. rosea which displays the largest gene repertoire of sequenced AM fungi so far. Several of these expanding gene families are coding for symbiosis-upregulated orphan genes, that are possibly playing a role in symbiosis. For example, we found dozens of MiSSPs in each taxa of AM fungi that may code for candidate effector proteins. It remains to investigate whether they play a role in host specificity and in symbiosis development as suggested by Kamel et al. (2017) and Zeng et al. (2018). Maeda et al. (2018) showed that TE contribute to gene duplication in several gene families in R. irregularis. Investigating the role of TE in the massive gene duplications observed in G. rosea will require a genome assembly of higher quality. The present study confirms and extends our initial findings (Tisserant et al., 2013) that protein kinase genes, such as those coding for protein tyrosine kinases, are among the largest gene families identified in AM fungal taxa. It is tempting to speculate that this large number of protein sensors play a role in symbiotic interactions, such as host specificity, and in planta accomodation of AM fungi.

Selected species in the Mucoromycotina sub-phylum includes fast growing, early colonizers of carbon-rich substrates such as Mucor and Rhizopus. They possess a substantial number of lignocellulose-degrading enzymes, although their set of PCWDE, especially those targeting crystalline cellulose (e.g., GH6, GH7, AA9 and CBM1) (Fig. 5, Supplementary Fig. S4), is lower than wood decayers in Dikarya (Kohler et al., 2015; Uehling et al., 2017). As obligate biotrophs, AM fungi do not need a repertoire of polysaccharide degrading enzymes because they derive most (if not all) of their carbon from their hosts, but the complete lack of genes acting on plant cell wall polysaccharides in Rhizophagus species is intriguing and gives rise to the question of how hyphae colonize the apoplastic space of host roots and how they degrade host cell walls to colonize host cells. The few remaining enzymes, i.e., multicopper oxidases (AA1), endoglucanase GH5_7, xyloglucanase GH5_12 and xyloglucosyltransferase GH16, are prime candidates for further functional analysis of fungal colonisation in planta. In addition to the loss of their saprotrophic enzymatic arsenal, AM fungi are lacking genes needed for thiamine biosynthesis, secondary metabolites and cytoplasmic fatty acid synthesis. This reduction in the biosynthetic ability is also observed in several

obligate biotrophic pathogens (Spanu, 2012). The evolutionary mechanism behind this convergent gene loss is not known, but it is supporting the assumption that their function has become obsolete due to the obligate biotrophic lifestyle.

The analysis of the present AM fungal genomes, together with the recently published R. clarus (Kobayashi et al., 2018) and D. epigaea (Sun et al., 2018) genomes, confirmed that AM fungal genomes are haploid and their genomic polymorphism is very low (0.14 to 0.35 SNP per kb). It also confirmed the presence of the gene machinery usually related to sex (e.g., MSG and putative MAT-loci) in these putative asexual clonal lineages. These genes are likely involved in the recently observed inter-nuclear recombination taking place in the dikaryotic life-stage of the R. irregularis isolates A4 and A5 (Chen et al., 2018b).

Obviously, we cannot sequence the genome of the unknown MRCA of Glomeromycotina and Mucoromycotina to identify the gene set involved in the transition from saprotrophism to symbiosis and obligate mutualism. Sequencing genomes of a much larger and diverse set of Mucoromycota associated to early land plants and of fine root endophytes (Field & Pressel, 2018) may facilitate the reconstruction of the genome of these ancient species which gave rise to the symbiotic lineage(s). Although Endogonales and Glomeromycotina are not sister groups and represent independent origins of mycorrhizal lifestyle within Mucoromycota, it is worth mentioning that the most prominent genome features of ectomycorrhizal Endogonaceae is their high TE content and a reduced number of PCWDE (Chang et al., 2018).

In conclusion, the present genome comparison refines our understanding of what makes Glomeromycotina unique. Their genomic features have arisen repeatedly in several independent lineages, likely as a result of convergence of evolutionary traits, suggesting that such adaptations can be favoured by selection. It is not yet known whether the identified genomic features are shared by the 315 AM fungal species. Despite the global dominance of Glomerales over the other AM fungal families, it is crucial to further corroborate our findings with improved sampling of other taxa from the more ancient, non-Glomerales families, such as Archeosporales and Paraglomerales.

Acknowledgements

We would like to thank Annegret Kohler, Claude Murat, Alexandre Poliakov, Christine Strullu-Derrien and Joseph Spatafora for discussions and advices. FMM's work is funded by the Laboratory of Excellence ARBRE (ANR-11-LABX-0002-01), Region Lorraine, European Regional Development Fund and Beijing Advanced Innovation Center for Tree Breeding by Molecular Design, Beijing Forest University. NC's work is supported by the Discovery program from Natural Sciences and Engineering Research Council of Canada (NSERC-Discovery), an Early Researcher Award from Ontario Ministry of Research and Innovation (ER13-09-190) and ZygoLife project funded by the U.S. National Science Foundation (DEB 1441677). CR is funded by the Laboratory of Excellence TULIP (ANR-10-LABX-41). Genome annotations were performed using the annotation pipeline of the U.S. DOE JGI (Contract DE-AC02-05CH11231) (1000 Fungal Genomes project and Mycorrhizal Genomics Initiative).

Author contributions

F.M.M., C.R. and N.C. planned and designed the research, wrote the manuscript, and helped with data analysis. E.M., H.S.C., E.C.H.C., S.M., A.P., I.D.L.P, M.H., E.D., and B.H. performed bioinformatic analyses; E.M. and N.T. performed the transcriptome analyses. A.K. and I.V.G. supervised the JGI gene prediction pipeline. S.R. and J.V. produced the biological material. I.D.L.P., S.N., D.B. produced DNA material and DNA sequences. CR and FMM were joint senior authors on this work.

References

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

Baggerly KA., Deng L, Morris JS, Aldaz CM. 2003. Differential expression in SAGE: accounting for normal between-library variation. Bioinformatics **19**: 1477-1483.

Beaudet D, Chen ECH, Mathieu S, Yildirir G, Ndikumana S, Dalpé Y, Séguin S, Farinelli L, Stajich JE, Corradi N. **2018**. Ultra-low input transcriptomics reveal the spore functional content and phylogenetic affiliations of poorly studied arbuscular mycorrhizal fungi. DNA Research: **25**: 217-227.

Bécard G, Fortin JA. 1988. Early events of vesicular arbuscular mycorrhiza formation on Ri T-DNA transformed roots. New Phytologist **108**: 211–218.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B, Statistical Methodology **57**: 289–300.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**: 2114-20.

Bonfante P, Genre A. 2008. Plants and arbuscular mycorrhizal fungi: an evolutionary-developmental perspective. Trends in Plant Science **13**: 492-498.

Bravo A, Brands M, Wewer V, Dörmann P, Harrison MJ. 2017. Arbuscular mycorrhiza-specific enzymes FatM and RAM2 fine-tune lipid biosynthesis to promote development of arbuscular mycorrhiza. New Phytologist **214**: 1631-1645.

Brundrett MC, Tedersoo L. 2018. Evolutionary history of mycorrhizal symbioses and global host plant diversity. New Phytologist **220**: 1108–1115.

Bruns TD, Corradi N, Redecker D, Taylor JW, Öpik M. 2018. Glomeromycotina: what is a species and why should we care? New Phytologist **220**: 963–967.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution **17**: 540-552.

Chang Y, Wang S, Sekimoto S, Aerts AL, Choi C, Clum A, LaButti KM, Lindquist EA, Yee Ngan C, Ohm RA et al. 2015. Phylogenomic analyses indicate that early Fungi evolved digesting cell walls of algal ancestors of land plants. Genome Biology and Evolution **7**: 1590–1601.

Chang Y, Desirò A, Na H, Sandor L, Lipzen A, Clum A, Barry K, Grigoriev IV, Martin FM, Stajich JE, Smith ME, Bonito G, Spatafora JW. 2018. Phylogenomics of Endogonaceae and evolution of mycorrhizae within Mucoromycota. New Phytologist. doi.org/10.1111/nph.15613.

Chen ECH, Morin E, Beaudet D, Noel J, Yildirir G, Ndikumana S, Charron P, St-Onge C, Giorgi J, Krüger M et al. 2018a. High intraspecific genome diversity in the model arbuscular mycorrhizal symbiont Rhizophagus irregularis. New Phytologist **220**: 1161–1171.

Chen ECH, Mathieu S, Hoffrichter A, Sedzielewska-Toro K, Peart M, Pelin A, Ndikumana S, Ropars J, Dreissig S, Fuchs J, Brachmann A, Corradi N. 2018b. Single nucleus sequencing reveals evidence of inter-nucleus recombination in arbuscular mycorrhizal fungi. eLife **7**: e39813.

Corradi N, Brachmann A. 2017. Fungal mating in the most widespread plant symbionts? Trends in Plant Science **22**: 175–183.

Corrochano LM, Kuo A, Marcet-Houben M, Polaino S, Salamov A, Villalobos-Escobedo JM, Grimwood J, Álvarez MI, Avalos J, Bauer D et al. 2016. Expansion of signal transduction pathways in fungi by extensive genome duplication. Current Biology **26**: 1577-1584.

Coupe SA, Deikman, J. 1997. Characterization of a DNA-binding protein that interacts with 5′ flanking regions of two fruit-ripening genes. Plant Journal **11:** 1207-1218.

Field KJ, Pressel S. 2018. Unity in diversity: structural and functional insights into the ancient partnerships between plants and fungi. New Phytologist **220**: 996–1011.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. arXiv preprint: 1207.3907 [q-bio.GN]

Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences **108**: 1513–1518.

Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F et al. 2014. MycoCosm portal: gearing up for 1000 fungal genomes. Nucleic Acids Research **42**: D699–D704

Han MV, Thomas GW, Lugo-Martinez J, Hahn MW. 2013. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Molecular Biology and Evolution **30**: 1987-1997.

Idnurm A, Hood ME, Johannesson H, Giraud T. 2015. Contrasted patterns in mating-type chromosomes in fungi: Hotspots versus coldspots of recombination. Fungal Biology Reviews **29**: 220–229.

Kamel L, Tang N, Malbreil M, San Clemente H, Le Marquer M, Roux C, Frei dit Frey N. 2017. The comparison of expressed candidate secreted proteins from two

arbuscular mycorrhizal fungi unravels common and specific molecular tools to invade different host plants. Frontiers in Plant Science **8**: 124

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution **30**: 772-780.

Kobayashi Y, Maeda T, Yamaguchi K, Kameoka H, Tanaka S, Ezawa T, Shigenobu S, Kawaguchi M. 2018. The genome of Rhizophagus clarus HR1 reveals a common genetic basis for auxotrophy among arbuscular mycorrhizal fungi. BMC Genomics **19**: 465.

Krüger M, Krüger C, Walker C, Stockinger H, Schüssler A. 2012. Phylogenetic reference data for systematics and phylotaxonomy of arbuscular mycorrhizal fungi from phylum to species level. New Phytologist **193**: 970-984.

Lanfranco L, Fiorilli V, Gutjahr C. 2018. Partner communication and role of nutrients in the arbuscular mycorrhizal symbiosis. New Phytologist **220**: 1031–1046.

Lastovetsky OA, Gaspar ML, Mondo SJ, LaButti KM, Sandor L, Grigoriev IV, Henry SA, Pawlowska TE. 2016. Lipid metabolic changes in an early divergent fungus govern the establishment of a mutualistic symbiosis with endobacteria. Proceedings of the National Academy of Sciences of the USA **113**: 15102-15107.

Leggett RM, Clavijo BJ, Clissold L, Clark MD, Caccamo M. 2014. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. Bioinformatics **30**: 566-568.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics **25**: 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics **25**: 2078-2079.

Lin K, Limpens E, Zhang Z, Ivanov S, Saunders DGO, Mu D, Pang E, Cao H, Cha H, Lin T et al. 2014. Single nucleus genome sequencing reveals high similarity among nuclei of an endomycorrhizal fungus. PLoS Genetics **10**: e1004078.

Lombard V, Golaconda Ramalu H, Drula E, Coutinho PM, Henrissat B. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acid Research **42**: D490-D495.

Luginbuehl LH, Oldroyd GED. 2017. Understanding the arbuscule at the heart of endomycorrhizal symbioses in plants. Current Biology **27**: R952-R963.

Maeda T, Kobayashi Y, Kameoka H, Okuma N, Takeda N, Yamaguchi K, Bino T, Shigenobu S, Kawaguchi M. 2018. Evidence of non-tandemly repeated rDNAs and their intragenomic heterogeneity in Rhizophagus irregularis. Communications Biology **1**: 87.

Maherali H, Klironomos JN. 2007. Influence of phylogeny on fungal community assembly and ecosystem functioning. Science **316** :1746-1748.

Mathieu S, Cusant L, Roux C, Corradi N. 2018. Arbuscular mycorrhizal fungi: intraspecific diversity and pangenomes. New Phytologist 220: 1129–1134.

Martin FM, Uroz S, Barker DG. 2017. Ancestral alliances: Plant mutualistic symbioses with fungi and bacteria. Science **356**: eaad4501.

Mittl PRE, Schneider-Brachert W. 2007. Sel1-like repeat proteins in signal transduction. Cellular Signalling **19**: 20-31.

Payen T, Murat C, Martin F. 2016. Reconstructing the evolutionary history of gypsy retrotransposons in the Périgord black truffle (Tuber melanosporum Vittad.). Mycorrhiza **26**: 553–563.

Pellegrin C, Morin E, Martin FM, Veneault-Fourrey C. 2015. Comparative analysis of secretomes from ectomycorrhizal fungi with an emphasis on small-secreted proteins. Frontiers in Microbiology **6**: 1278.

Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. Bioinformatics **21**: i351–i358.

Redecker D, Schüssler A, Stockinger H, Stürmer SL, Morton JB, Walker C. 2013. An evidence-based consensus for the classification of arbuscular mycorrhizal fungi (Glomeromycota). Mycorrhiza **23**: 515-531

Riley R, Corradi N. 2013. Searching for clues of sexual reproduction in the genomes of arbuscular mycorrhizal fungi. Fungal Ecology **6**: 44-49.

Ropars J, Toro KS, Noel J, Pelin A, Charron P, Farinelli L, Marton T, Krüger M, Fuchs J, Brachmann A et al. 2016. Evidence for the sexual origin of heterokaryosis in arbuscular mycorrhizal fungi. Nature Microbiology. **1:** 16033.

Salvioli A, Ghignone S, Novero M, Navazio L, Venice F, Bagnaresi P, Bonfante P. 2016. Symbiosis with an endobacterium increases the fitness of a mycorrhizal fungus, raising its bioenergetic potential. ISME Journal **10**: 130-44.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single copy orthologs. Bioinformatics **31**: 3210–3212.

Spanu PD. 2012. The genomics of obligate (and nonobligate) biotrophs. Annual Review of Plant Pathology **50**: 91-109.

Spatafora JW, Chang Y, Benny GL, Lazarus K, Smith ME, Berbee ML, Bonito G, Corradi N, Grigoriev I, Gryganskyi A et al. 2016. A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data. Mycologia **108**: 1028-1046.

St-Arnaud M, Hamel C, Vimard B, Caron M, Fortin JA. 1996. Enhanced hyphal growth and spore production of the arbuscular mycorrhizal fungus Glomus intraradices in an in vitro system in the absence of host roots. Mycological Research **100**: 328–332.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H et al. 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Research **12**: 1611–1618.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**: 2688-2690.

Strullu-Derrien C, Selosse MA, Kenrick P, Martin FM. 2018. The origin and evolution of mycorrhizal symbioses: from palaeomycology to phylogenomics. New Phytologist **220**: 1012-1030.

Sun X, Chen W, Ivanov S, MacLea, AM, Wight H, Ramaraj T, Mudge J, Harrison MJ, Fei Z. 2018. Genome and evolution of the arbuscular mycorrhizal fungus Diversispora epigaea (Glomus versiforme) and its bacterial endosymbionts. New Phytologist. doi: 10.1111/nph.15472.

Tang N, San Clemente H, Roy S, Becard G, Zhao B, Roux C. 2016. A survey of the gene repertoire of Gigaspora rosea unravels conserved features among glomeromycota for obligate biotrophy, Frontiers in Microbiology **7:** 233.

Van der Heijden MGA, Martin FM, Selosse MA, Sanders IR. 2015. Mycorrhizal ecology and evolution: the past, the present, and the future. New Phytologist **205**: 1406–1423

Tisserant E, Kohler A, Dozolme-Seddas P, Balestrini R, Benabdellah K, Colard A, Croll D, Da Silva C, Gomez SK, Koul R et al. 2011. The transcriptome of the arbuscular mycorrhizal fungus Glomus intraradices (DAOM 197198) reveals functional tradeoffs in an obligate symbiont. New Phytologist **193**: 755–769.

Tisserant E, Malbreil M, Kuo A, Kohler A, Symeonidi A, Balestrini R, Charron P, Duensing N, Frei dit Frey N, Gianinazzi-Pearson V et al. 2013. Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. Proceedings of the National Academy of Sciences of the USA **110:** 20117-20122

Uehling J, Gryganskyi A, Hameed K, Tschaplinski T, Misztal PK, Wu S, Desirò A, Vande Pol N, Du Z, Zienkiewicz A, Zienkiewicz K et al. 2017. Comparative genomics of Mortierella elongata and its bacterial endosymbiont Mycoavidus cysteinexigens. Environmental Microbiology **19**: 2964–2983.

Voets L, de la Providencia IE, Declerck S. 2006. Glomeraceae and Gigasporaceae differ in their ability to form hyphal networks. New Phytologist **172**: 185-188.

Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix D, Kenyon R, Machi D et al. 2014. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Research **42**: D581-D591.

Wewer V, Brands M and Dörmann P. 2014. Fatty acid synthesis and lipid metabolism in the obligate biotrophic fungus Rhizophagus irregularis during mycorrhization of Lotus japonicus. Plant Journal **79**: 398–412.

Zeng T, Holmer R, Hontelez J, Te Lintel-Hekkert B, Marufu L, de Zeeuw T, Wu F1, Schijlen E, Bisseling T, Limpens E. 2018. Host- and stage-dependent secretome of the arbuscular mycorrhizal fungus Rhizophagus irregularis. Plant Journal **94**: 411–425.

The following Supporting Information is available for this article:

Supplementary Table S1 Summary statistics of predicted gene repertoires for Glomeromycotina, Mortierella elongata, and selected Mucoromycotina used in this study.

Supplementary Table S2 Number of fungal gene markers identified by BUSCO in the genome assemblies used in the present study.

Supplementary Table S3 Gene orthology for the the four sequenced Glomeromycotina and eight sequenced Mucoromycota species.

Supplementary Table S4 Summary statistics for predicted proteome, core and dispensable genes (core-disp) and species-specific genes (specs) of the eight Mucoromycota species.

Supplementary Table S5 Most abundant protein families in expansion in Rhizophagus irregularis compared to the other Glomeromycotina, selected Mucoromycotina and Mortierella elongata.

Supplementary Table S6 Most abundant protein families in expansion in Rhizophagus cerebriforme compared to the other Glomeromycotina, selected Mucoromycotina and Mortierella elongata.

Supplementary Table S7. Most abundant protein families in expansion in Rhizophagus diaphanus compared to the other Glomeromycotina, selected Mucoromycotina and Mortierella elongata.

Supplementary Table S8. Most abundant protein families in expansion in Gigaspora rosea compared to the other Glomeromycotina, selected Mucoromycotina and Mortierella elongata.

Supplementary Table S9. Distribution of genes coding for signaling/transduction pathways in Glomeromycotina species analyzed in this study.

Supplementary Table S10. List of orthologous genes showing evidence of rapid sequence evolution in the Glomeromycotina genomes.

Supplementary Table S11. Distribution of genes coding for secreted carbohydrate-active enzymes (CAZymes), total CAZymes and CAZymes acting on plant or fungal cell walls.

Supplementary Table S12. Presence of meiosis-specific gene orthologues in Glomeromycotina species with sequenced genomes and single nucleotide polymorphisms in Glomeromycotina.

Supplementary Table S13. Secretome, including secreted CAZymes, secreted lipases, secreted proteases and small secreted proteins (SSP) for all Mucoromycota species in this study.

Supplementary Table S14. Orthogroups of small secreted proteins without annotation (unknown proteins).

Supplementary Table S15. Presence and sequence similarity of upregulated genes from R. irregularis interacting with Brachypodium distachyon and of upregulated genes from G. rosea interacting with Brachypodium distachyon in genomes of sequenced Mucoromycota (linked to Fig. 5B).

Supplementary Table S16. Clusters of all Mucoromycota species genomes.

Supplementary Table S17. Pfam protein domains counts in genomes for all Mucoromycota species in this studies and five isolates of R. irregularis (linked to Fig.3A)

Supplementary Fig. S1 Macrosynteny between Rhizophagus irregularis and R. diaphanus scaffolds.

Supplementary Fig. S2 Expansion and contraction of gene families as identified by CAFÉ analysis in sequenced Glomeromycotina, Mortierella elongata and selected Mucoromycotina.

Supplementary Fig. S3 Sequence divergence of conserved orthogroups in sequenced Glomeromycotina in this study.

Supplementary Fig. S4 Presence and abundance of genes encoding secreted plant cell wall degrading enzymes in the genome of the eight Mucoromycota species.

Supplementary Fig S5 Distribution of allele frequency (as SNP) in the genome of Rhizophagus irregularis, R. diaphanus, R. cerebriforme, and Gigaspora rosea.

Supplementary Fig. S6 Schematic representation of the putative MAT-locus in Glomeromycotina.

Table 1 Summary statistics for genome assemblies of the sequenced Glomeromycotina and selected saprotrophic Mucoromycota used in this study.

Species	Assembly size (Mbp)	Contig no.	Contig N50 (no.)	Contig L50 (kbp)	Scaffolds	Scaffold N50	Scaffold L50 (kbp)	Scaffold (kbp) min - max		Total gap length (%)	Total repeat (%)	GC content (%)
Gigaspora rosea	597.95	28,997	3,991	37.7	7,526	734	232.08	0.92	1,204.75	7.92	63.44	28.81
Mortierella elongata	49.86	742	77	219.8	473	31	517.14	1.00	1,526.29	0.30	4.63	48.05
Mucor circinelloides	36.59	26	4	4318.34	26	4	4,318.34	2.29	6,050.25	0.00	20.38	42.17
Phycomyces blakesleeanus	53.94	350	41	370.4	80	11	1,515.58	2.96	4,452.46	1.06	9.74	35.78
Rhizophagus cerebriforme	136.89	14,636	1,679	18.5	2,592	266	147.87	0.90	709.02	17.60	24.77	26.55
Rhizophagus diaphanus	125.87	11,501	1,354	22.9	2,764	269	137.49	0.88	686.31	12.52	20.18	27.19
Rhizophagus irregularis	136.80	5,810	768	52.03	1,123	129	336.38	0.96	1,375.86	5.06	26.38	27.53
Rhizopus microsporus	25.97	823	111	69.4	131	8	1,118.34	1.02	2,782.17	2.41	4.68	37.48

Fig. 1 Distribution of transposable element (TE) families in genomes of sequenced Mucoromycota. (a) TE coverage (%) in genome assemblies. (b) Copy number per TE family.

Fig. 2 Organismal phylogeny of the eight Mucoromycota species, plus one representative of Basidiomycota (Laccaria bicolor), one representative of Ascomycota (Tuber melanosporum) and two basal fungi, Conidiobolus coronatus and Rozella allomycis. We identified 784 gene clusters with only one protein-coding gene per species by clustering protein sequences using FastOrtho (Wattam et al., 2013). Each cluster was then aligned with MAFFT (Katoh & Standley, 2002), and a maximum likelihood inference was performed with RAxML (PROTGAMMAWAG model) and 1000 bootstrap replicates (Stamatakis, 2014).

Fig. 3 Gene conservation and innovation in Glomeromycotina, Mortierellamycotina, Mucoromycotina species. (a) Organismal phylogeny. (b) Bar graphs represent sets of conserved proteins shared among species (dark blue), sets of duplicated conserved proteins shared among species (light blue), sets of dispensable proteins (purple), sets of duplicated dispensable proteins (light purple), species-specific (orange) and duplicated specific-specific (light orange) proteins. Note that some of the species-specific genes found by comparing the eight Mucoromycota genomes have orthologues in other fungi (yellow). Protein ID and sequences for each FastOrtho orthogroups (i.e. gene families) are listed in Supporting Information Table S16.

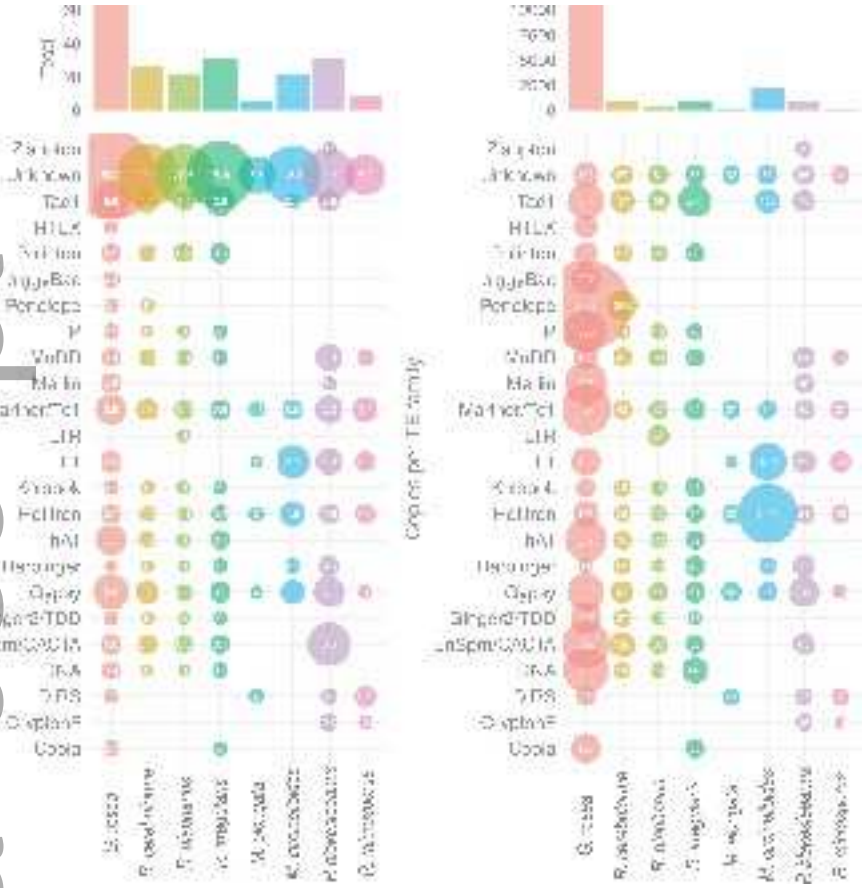
Fig. 4 Functional diversity encoded by Mucoromycota genomes. (a) Presence and abundance of the different Pfam domain-containing proteins in the eight Mucoromycotina species (this study) and Rhizophagus irregularis isolates A1, A4, A5, B3 and C2 (Chen et al., 2018). The heat map depicts absolute Pfam domain counts in each of the sampled genomes, according to the color scale (only the top most frequent 100 domains are shown). The abundance values were then transformed into z-scores, which are measure of relative enrichment (red) and depletion (green); the hierarchical clustering was done with a Euclidian distance metric and average linkage clustering method. The data were visualized and clustered using MultiExperiment Viewer

(http://www.tm4.org/mev.html). (b) Diversity of KEGG pathways in Mucoromycota genomes. Pearson correlation matrix was calculated based on profile of protein-coding genes assign to KEGG modules; to perform hierarchical clustering the correlation matrix is converted into a distance matrix. The hierarchical clustering was done with a Euclidian distance metric and complete linkage clustering method. Colors are coded from dark red representing high correlation to white representing lower correlation. Counts of selected Pfam-domain are listed in Supporting Information Table S17.

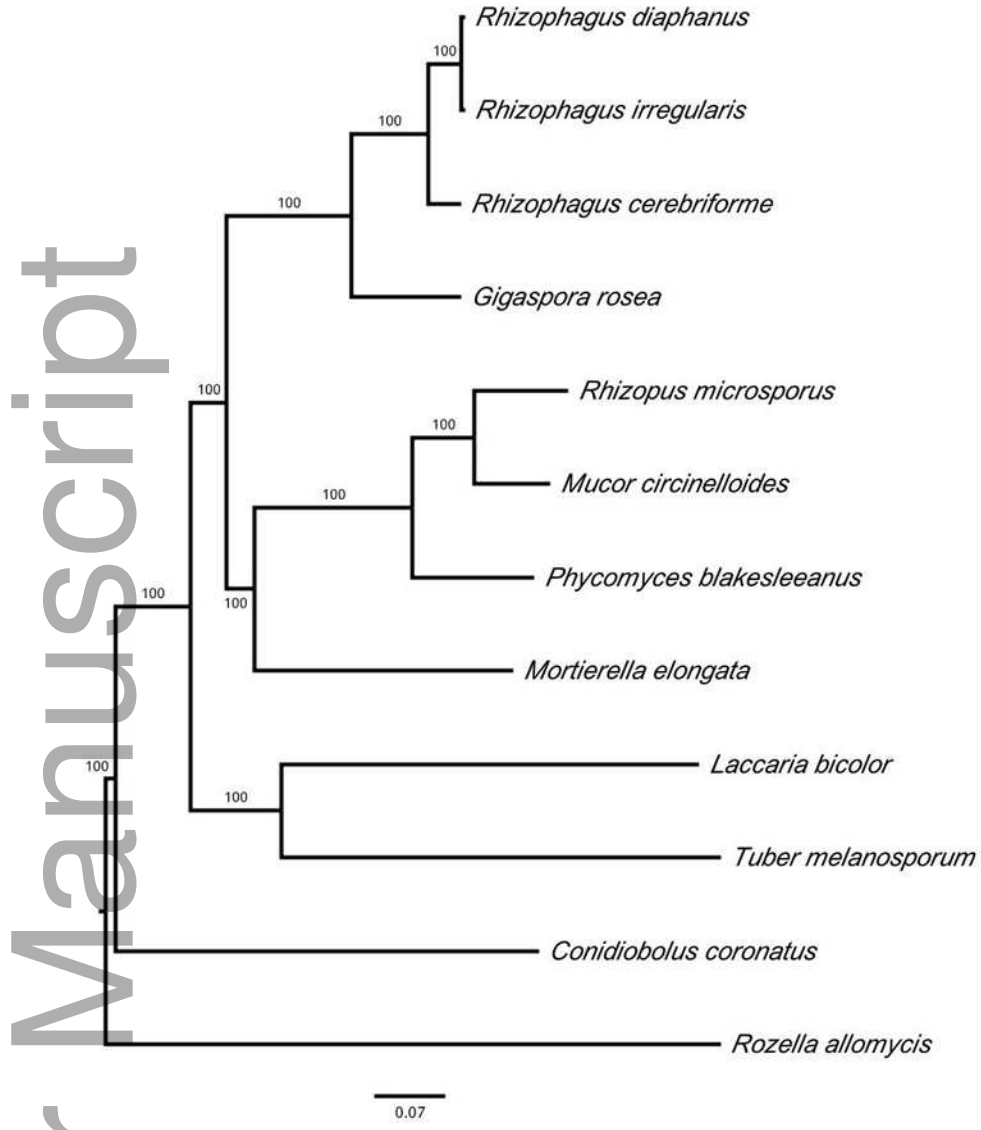
Fig. 5 Presence and abundance of genes encoding for secreted enzymes involved in the degradation of plant, fungal and bacterial cell wall polysaccharides in the eight Mucoromycota species. The bubble plot depicts absolute counts for genes encoding secreted CAZymes involved in the degradation of polysaccharides and lignin derivatives. The bar plots depicts the numbers and ratio of secreted and nonsecreted enzymes acting on plant (PCWDE) or microbial (MCWDE) polysaccharides (http://www.cazy.org). AA, auxiliary activities; CBM, carbohydrate-binding modules; CE, carbohydrate esterases; EXPN, distantly related to plant expansins; GH, glycoside hydrolases; PL, polysaccharide lyases.

Fig. 6 Presence and sequence similarity of symbiosis-upregulated genes from Rhizophagus irregularis (a) and Gigaspora rosea (b) interacting with Brachypodium distachyon in the genome of the eight Mucoromycota species. The heatmap depicts a double-hierarchical clustering of (a) 426 symbiosis-upregulated R. irregularis genes (rows, fold change ≥5 in symbiotic tissues compared to germinating hyphae from spores, false discovery rate-corrected $P \leq 0.05$; Supporting Information Table S15a) based on their percentage sequence identity, 0 to 100% (color scale at left) with their orthologues (if any) in selected taxa (columns). Right panel, functional categories (KOG) are given for each transcript cluster in percentage as bargrams and the number and percentage of genes in each cluster are shown. Data were visualized and clustered using R (package HeatPlus). The hierarchical clustering was done by using a Euclidian distance metric and Ward clustering method. The bottom heatmap (b) depicts a double-hierarchical clustering of 989 symbiosis-upregulated G. rosea genes (Table S15b) based on their percentage sequence identity with their orthologues (if any) in selected taxa.

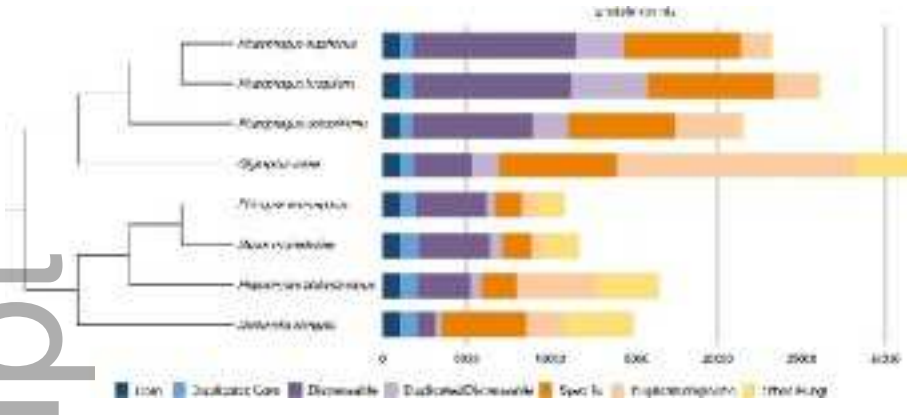
List of symbiosis-upregulated genes and their distribution by clusters is provided in Tables S15(a, b).



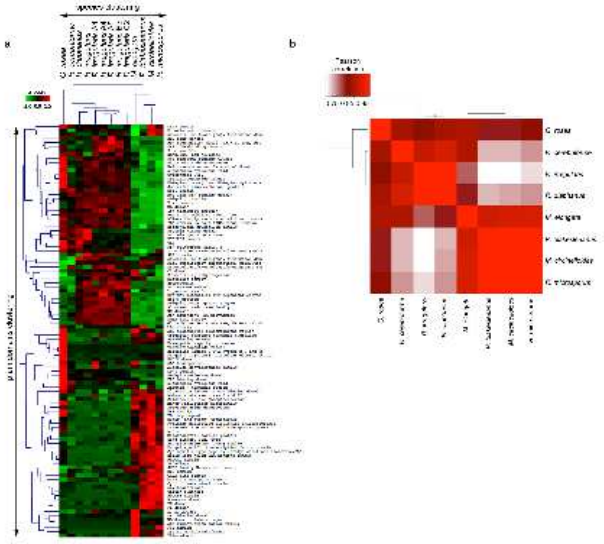
nph_15687_f1.tif



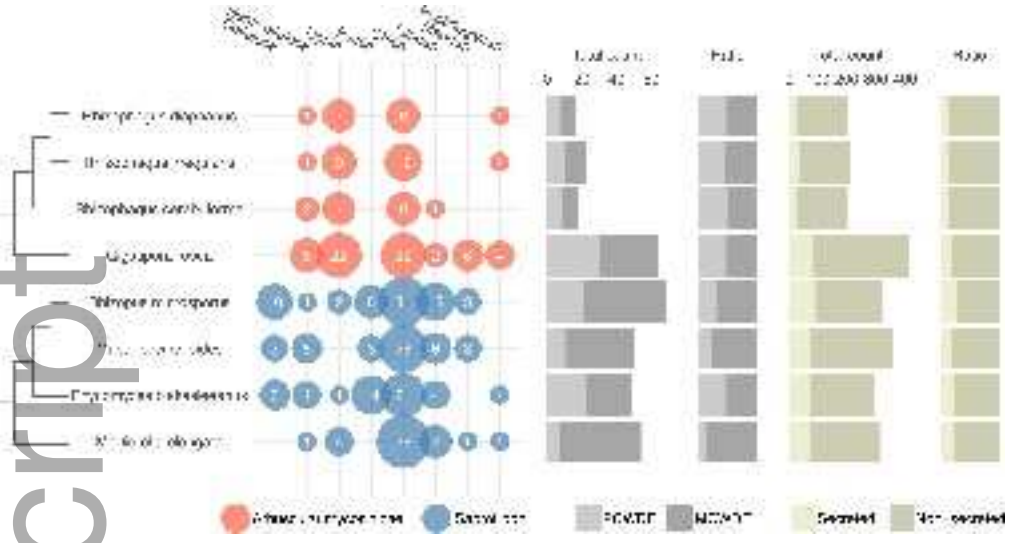
nph_15687_f2.tif



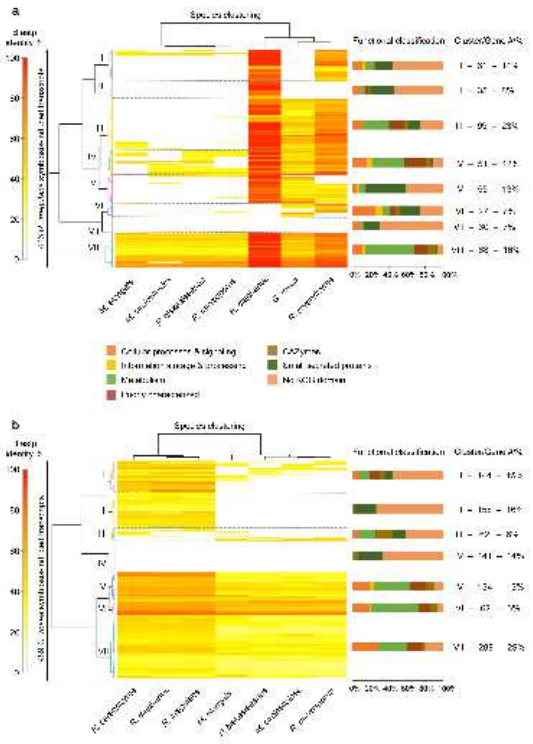
nph_15687_f3.tif



nph_15687_f4.tif



nph_15687_f5.tif



nph_15687_f6.tif