Supplementary Information for

The genome of an arbuscular mycorrhizal fungus provides insights into the oldest plant symbiosis

This PDF file includes:

Supporting Online Material

Table of Contents

1. Material and Methods

1.1. Biological material

Fungal isolates and spore production. The isolate DAOM-197198 (= DAOM-181602) of *Rhizophagus irregularis* (Błaszk, Wubet, Renker & Buscot) C. Walker & A. Schüßler comb. nov. was used for genome sequencing in this study (1,2); see (1) for the history of the DAOM-197198 isolate). As AM fungi are obligate symbionts, spores were produced in monoxenic cultures maintained on *Agrobacterium rhizogenes*-transformed roots of carrot (*Daucus carota*, clone DC2) (3). In addition to DAOM-197198, a number of genetically divergent strains of *R. irregularis* (A4, B3, C2) (4), and one strain of *R. diaphanus* (MUCL43196) have also been used. The selected strains (A4, B3, C2) have been selected to assess genome polymorphism (C2, MUCL43196) and/or potential MAT gene 'idiomorphs' (A4, B3, C2) for the following reasons: (i) *in vitro* fungal cultures with transformed carrot roots are available, allowing the isolation of sufficient amount of DNA to perform PCR that are free of any obvious contaminants ; (ii) They represent different members of a single population of *R. irregularis* isolated in Taenikon, Switzerland (4); (iii) They are genetically different from each other, and are able to undergo anastomoses; and may thus be able of undergoing a parasexual life cycle (i.e. nuclear exchange); and (iv) Genome sequencing is on-going on these strains for further analyses. These isolates were cultivated within *in vitro* split-plates during eight weeks to produce ~8,000,000 pure spores (multisporal & multinucleate material) as previously described (4).

Time-course of spore germination. For the transcript profiling of germinating spores carried out by Illumina RNA sequencing (RNA-Seq), 80,000 to 125,000 sterile spores of *R. irregularis* DAOM-197198 (Agronutrition, Labège, France) were germinated in sterile liquid M medium. Spores were germinated in the dark at 30°C in 2% CO_2. After 2and 9 days of germination, the spores were removed by filtration, immediately frozen in liquid nitrogen and crushed with a pestle and mortar for further RNA extraction.

R. irregularis *gene expression in* Medicago truncatula. To increase the mycorrhizal colonization of *M. truncatula* roots, a nursery system was used as follows. Leek seeds (*Allium porrum* cv. Elboeuf) were germinated on vermiculite for two weeks and then transferred to 4 L pots with sterile charred clay (Oil-Dri, Klasmann, France) as substrate, mixed with 400 spores of *R. irregularis* per L. Leek plants were grown for 3 months at 25°C/22°C under a long day photoperiod (16h light/8h dark at 25°C/22°C) with 80% humidity. Seeds of *M. truncatula* ecotype A17 were surface sterilized with 3.2% sodium hypochlorite for 10 min, 95% ethanol for

2 min, and washed four times with sterile distilled water. Seeds were germinated on 1% water-agar for 5 days at 4°C in the dark. Germinated seedlings were transferred to pots containing 3-month-old mycorrhizal leeks (6 plantlets per pot) and grown under a long day photoperiod (16h light/8h dark at 22°C/20°C) with 40% humidity. To avoid any changes in gene expression related to the nursery growth system (e.g., nutrient and signal exchanges between plants via the fungal networks), plantlets were removed from the leek-nursery pot after three weeks and transferred into a 250-ml pot with sterile charred clay as substrate. After two weeks (22/20°C, 70% humidity under a long day photoperiod), *M. truncatula* roots were collected, thoroughly washed to remove the substrate, and frozen in liquid nitrogen. Random fragments of the mycorrhizal roots were isolated to estimate the colonization level (5). Colonization levels ranged between 65% and 68%. All plants were watered daily with half-strength Long Ashton Nutrient Solution (6) containing a final concentration of 7.5 μM P.

1.2. DNA extraction

For 454 pyrosequencing, high-molecular weight (HMW) genomic DNA (gDNA) was extracted from 1,000,000 sterile spores of *R. irregularis* DAOM-197198 (commercial inoculant Mycorhise® ASP; Premier Tech Biotechnologies, Rivière-Du-Loup, QC, CA) in a liquid suspension of 4,000 spores per mL. This suspension was filtered on a sterile plastic 35μm sieve. Under a binocular microscope, spores were checked for root contamination and root fragments were removed with forceps. Spores were then gently crushed in a 1.5 ml plastic microtube using a sterilized pestle at 4°C. DNA was extracted using DNeasy Plant Mini kit (Qiagen) at the Tree-Microbe Interactions Department at INRA-Nancy, according to manufacturer's instructions.

For Illumina genome sequencing, HMW DNA was prepared by the Laboratoire de Recherche en Sciences Végétales (LRSV, Toulouse, France) from 18,000,000 pure sterile spores of DAOM-197198 produced by Agronutrition (Labège, France) as described in section 1.1. Spores were gently crushed without freezing and DNA was extracted using a modified cetyltrimethylammonium bromide (CTAB) method (7). About 20 μg DNA was subsequently analyzed for quality via pulsed field gel electrophoresis using a BioRad CHEF Mapper system (Bio-Rad Laboratories, Hercules, California).

1.3. Genome Sequencing & Assembly

Sanger sequencing. Fosmids were sequenced using Sanger sequencing on ABI 3730XL capillary machines at the U.S. Department of Energy Joint Genome Institute (JGI). All general aspects of library construction and sequencing can be found at the JGI website

http://www.jgi.doe.gov/ and GenBank (accession n° AUPC00000000, BioProject: PRJNA208392).

454 pyrosequencing. HMW gDNA was sequenced at the Genome Quebec Innovation Centre (McGill University, Montreal) and at Beckman Genomics facilities (Danvers, MA, USA) by GS FLX Titanium (Roche 454 Life Science).

Illumina sequencing. HMW gDNA was high-molecular-weight DNA was used to generate libraries for Illumina Sequence by Synthesis (Illumina-SBS) genome sequencing using the standard Illumina TruSeq DNA protocol (http://genome.med.harvard.edu/documents/illumina/TruSeq_DNA_SamplePrep_Guide_15005180_A.pdf). Illumina sequencing was conducted using the services of a commercial provider (GATC Biotech AG, Konstanz, Germany) on a HiSeq 2000 sequencing platform. Illumina sequencing yielded 37,094,828 single end reads of 100 bp length.

PacBio sequencing. One single-molecule real-time (SMRT) cell library was prepared with PacBio C2 DNA preparation kit 2.0 (3-10 kb, Pacific Biosciences cat#001‐540‐835), using 3.5 µg of a purified DAOM-197198 HMW gDNA. The polymerase was bound to the library with the DNA/Polymerase Binding Kit XL 1.0 (Pacific Biosciences cat# 100-150-800). Subsequently, the library was sequenced on the PacBio RS using nine SMRT cells (version 2 cat#001-350-385) and C2 DNA Sequencing Kit 2.0 (cat #001-554-002) at the Lausanne University Genomic Technologies Facility. A one hundred and twenty minute length movie was taken for each SMRTcell. Sequencing yielded 766 Mb of quality-filtered data in 139,080 raw long reads, with an average length of 3 kb (maximum length: 22 kb). The read sequences represented a genome coverage of ~5X.

Genome assembly. All computational genome assemblies were conducted using an Apple Power Mac server (2 x 2.93 GHz 6-core Intel Xeon with 64 Go DDR3 RAM). Before genome assembly, the sequences were filtered for adapter sequences, low quality reads and bacterial and fungal contaminants (**Fig. S1**). Sequences with a GC%>45 were considered as contaminants and discarded (**Fig. S20**). We then used the CLC Genomic Workbench program v6.0 for the genome hybrid assembly using 30,067,076 filtered reads (3,78 Gb). Parameters were as followed: automatic word size; bubble size, 50; deletion cost, 3; insertion cost, 3; length fraction, 0.5; mapping mode, map reads back to contigs; minimum contig length, 1,000; mismatch cost, 2; similarity fraction, 0.8. This genome hybrid assembly, so-called Gloin1, resulted in 28,371 scaffolds, ranging in size from 1,000 to 57,883 nucleotides (N50 length of 4.19 kb). Due to the

high proportion of repetitive DNA, a substantial part of the whole-genome shotgun reads collapsed into relatively small contigs characterized by exceptionally high read depths (>2,000-fold) (**Fig. S2)**.

We sought to improve the assembly by using a hybrid SMRT-Illumina strategy. The Gloin1 hybrid assembly was combined with PacBio data using the PacBio AHA (A Hybrid Assembler) scaffolding algorithm to produce the assembly PacBio-Gloin1. This resulted in a ~101 million base-pair (Mb) genome assembly (N50 length of 15.16 kb). This assembly indicated that the DAOM-197198 genome is rich in AT bases (A + T content 72%) (**Fig. 2**). Each scaffold was screened against bacterial proteins, organelle sequences and GenBank using megablast against Genbank NR and BLASTP against a set of known microbial proteins. Bacterial sequences were discarded (**Fig. S20**). Additional scaffolds were identified as unanchored ribosomal DNA, mitochondrial DNA, and repetitive elements. Scaffolds <1 kbp were removed. The genome sequences have been deposited in the GenBank database (accession n° AUPC00000000, BioProject: PRJNA208392) and are available at the *Rhizophagus* JGI portal at http://jgi.doe.gov/Glomus/.

To estimate the size of the *R. irregularis* genome, we first estimated the real sequencing depth based on *k*-mer (17-mer) frequency distribution in the Illumina sequencing reads as described (11) (**Fig. S6**). The peak of 17-mer frequency (M) in reads is correlated with the real sequencing depth (N), read length (L), and *k*-mer length (K), their relations can be expressed as $M = N * (L – K + 1)/ L$. Then, we divided the total Illumina sequence length (2,741Mb) by the real sequencing depth (17.85x) and obtained an estimated genome size of 153 Mb.

The sequences with a GC%>45% were assembled using CLC Genomic Workbench v6.0 yielding 13,365 scaffolds; 90.5 % of them showed a BLASTN hit against bacterial genomes in the NCBI nr database. The most abundant bacterial sequences belonged to *Pseudomonas denitrificans* (19.3%), *Pseudomonas putida* (17.5%), *Klebsiella oxytoca* (16.4%) and *Enterobacter cloacae* (12.4%).

1.4. Single nucleotide polymorphisms (SNP) calling

Illumina genomic reads from *R. irregularis* DAOM-197198 and RNA-Seq reads from *R. irregularis* DAOM-197198, *R. irregularis* C2 and *R. diaphanum* were mapped to the *R. irregularis* Gloin1 assembly using the Burrows-Wheeler Aligner (BWA) software package (9) with the default settings. SNP calling was performed separately for the Illumina genome and transcriptome sequencing reads using the functions *mpileup* and *bcftools* of SAMtools v0.1.18 (10), with options –bvcg. The obtained SNP positions were filtered to remove homozygous positions, SNP quality <20 or mapping quality < 20 and SNP within 3 bp around a gap. Finally,

to remove potential artifactual SNPs caused by repetitive and paralogous sequences, SNP positions with coverage < 5 and coverage in the top 5% (>30, according to the respective sequencing coverage) were filtered for the genome data (**Table 1**). The SNP rate in the DAOM-197198 genome was compared to the SNP rate in the homokaryotic (haploid) genome of the ectomycorrhizal ascomycete *T. melanosporum* Mel28 and the dikaryotic (diploid) genome of the ectomycorrhizal basidiomycete *Laccaria bicolor* S238N. Illumina genomic reads were mapped to the respective reference genomes available at the JGI MycoCosm portal and SNPs were called as described above. SNPs in transcripts of *R. irregularis* DAOM-197198, *R. irregularis* C2 and *R. diaphanum* MUCL43196 were identified using SAMtools v2.

Genes with the highest number of high quality SNPs in the *R. irregularis* genome assembly were validated by mapping transcriptome reads (38.5 Gb) from germinating spores (**Table S2**) to the genome assembly by using SAMtools v2 and CLC Genomic Workbench v6.5. Both the 'Probabilistic variant detection' and the 'Quality based variant detection' algorithms in CLC Genomic Workbench v6.5 were used (see descriptions at (http://www.clcsupport.com/clcgenomicsworkbench/650/index.php?manual=Probabilistic_variant_detection.html and (http://www.clcsupport.com/clcgenomicsworkbench/650/index.php?manual=Quality_based_variant_detection.html#sec:SNPdetection).

1.5. Transposable elements, simple sequence repeats, mini-satellites and satellites

RepeatScout (12) was used to identify *de novo* repetitive DNA in the *R. irregularis* assembly. The default parameters (with l= 15) were used. RepeatScout generated a library of 4,198 consensus sequences. This library was then filtered as follows: 1) all the sequences less than 100 bp in size were discarded; 2) repeats having less than ten copies in the genome were removed (as they may correspond to protein-coding gene families) and 3) repeats having significant hits to known proteins in UNIPROT (13) other than proteins known as belonging to TEs were removed. The remaining 2,012 consensus sequences remaining were annotated manually by a TBLASTX search (14) against RepBase (15); 21 consensus sequences showed homologies with Class 1 retrotransposons and 87 with Class 2 transposons (**Table S1**). The remaining 1,904 consensus sequences were not categorized. To identify full length LTR retrotransposons, a second *de novo* search was performed with LTR_STRUC (16). The program yielded only two full-length candidate LTR retrotransposon sequences, which were checked for their homology using the TBLASTX against the sequences coming from the RepBase database. Neither element exhibited significant homology with known TE families. These sequences have been excluded from further analyses. The number of repeat element occurrences and the percent

of genome coverage were assessed by masking the genome assembly using RepeatMasker (17) with the 2,012 consensus sequences coming from the RepeatScout pipeline. RepeatMasker masked 11.30 % of the genome assembly; 10.30 % of the genome was masked by repeated elements belonging to unknown/uncategorized families, 0.29 % by Class 1 retrotransposons *Gypsy* and 0.71 by Class 2 transposons (**Table S1**). Notably, this proportion of repeated elements is expected to be a significant underestimate as over 16% of Illumina reads were not assembled and likely correspond to repeated sequences. This abundance of repeated structures was confirmed by Sanger sequencing of random fosmids. Not surprisingly, the sequenced fosmids showed a much higher abundance of TEs (36%) (**Fig. S4**). The fosmid regions that did not align to the genome assembly mainly coded for repeated elements (**Fig. S4**); 18 LTRs, including two complete *Gypsy* with similarity to *Rhizopus oryzae* LTRs and numerous LTR remnants were identified in fosmids. The fact that LTR retrotransposons are long (9 to 25 kb) (**Fig. S4**), highly repetitive and nested is the main explanation for the observed high fragmentation of the assembly.

MISA (http://pgrc.ipk-gatersleben.de/misa/download/misa. pl) with default parameters was used to identify mono- to hexanucleotide simple sequence repeat (SSR) motifs. A total of 73,276 SSRs have been identified in the *R. irregularis* genome corresponding to 62,641 mono-, 5,939 di-, 3,562 tri-, 879 tetra-, 180 penta- and 75 hexanucleotide motifs. The relative abundance of SSRs was calculated as the number of SSRs per Mb. For all 73,276 SSRs, the relative abundance was 804 SSRs/Mb. The *R. irregularis* genome is richer in SSR than any other fungal genome analyzed to date (18, 19). Looking at the relative abundance, the SSRs were more frequent in the UTR (1,389 SSR/Mbps) and introns (1,306 SSR/Mbps) than in intergenic regions (969 SSR/Mbps), TE (68 SSR/Mbps) and exons (65 SSR/Mbps). The abundance of SSRs in UTRs is not common in fungi and might represent an involvement of these repeated sequences in gene regulation as observed in human (20). Mini-satellites (motif of 7 to 100 bp) and satellites (motif >100bp) were searched for in the *R. irregularis* genome using the Tandem Repeats Finder software (21) with the following parameters: 2 7 7 80 10 50 500. A total of 34,040 mini-satellites and 2,026 satellites were identified with genome coverage of 2.43 % and 1.32%, respectively. These values are in the range found in other fungal genomes (19).

1.6. Gene calling and sequence analysis

Contigs larger than 1,000 bp produced by the assembly were used as input contigs for gene model generation and downstream analysis. Gene calling was conducted using the JGI Annotation Pipeline, which combines several gene predictors: i) protein-based gene models

were predicted using FGENESH+ (22) and GeneWise (23) seeded by BLASTx alignments of genomic sequence against sequences from the NCBI non-redundant protein set nr and from the JGI *Phycomyces* and *Mucor* annotations (http://jgi.doe.gov/mucoromycotina/; manuscript in preparation), ii) *ab initio* gene models were predicted using GeneMark-ES (24) and FGENESH, the latter trained on the set of putative full-length genes and reliable protein-based models, and iii) RNA-based gene models were derived either by assembling 14,749,180 RNA sequences into 52,409 contig sequences which were then modeled on genomic sequence, or by aligning 14,750,646 RNA sequences onto genomic sequence, and then assembling the alignments into gene models. GeneWise models were completed using scaffold data to find start and stop codons. RNA contig alignments to the genome were used to verify, complete, and extend the gene models. Because multiple gene models per locus were often generated, a single representative gene model for each locus was chosen based on protein similarity and RNA support, and used for further analysis. This led to a filtered set of 30,282 gene models with their properties and support by different lines of evidence summarized in **Table S3**. After filtering for bacterial contaminants (**Fig. S20**) and fragments of TEs, the final number of protein-coding genes was set to 28,232. Among these gene models, 23,561 genes have transcriptomic support (RNA-Seq) and/or showed sequence similarity to documented proteins and/or domains; these models were considered as '*high-confidence*' gene models.

All predicted gene models were functionally annotated by the JGI Annotation Pipeline using InterProScan (25) and hardware-accelerated double-affine Smith-Waterman alignments (http://www.timelogic.com/) against highly curated databases such as SwissProt (26), KEGG (Kyoto Encyclopedia of Genes and Genomes) (27) and PFAM (28). KEGG hits were used to map EC numbers (29), and InterPro, KEGG, and SwissProt hits were used to map GO (Gene Ontology) terms (30). In addition, predicted proteins were annotated according to KOG (EuKaryotic Orthologous Groups) classification (31). Protein targeting predictions were made with signalP (32) and TMHMM (33). All gene models and annotations may be accessed at the JGI *Rhizophagus irregularis* Portal (http://jgi.doe.gov/Glomus/; (34)).

The different species chosen for multigene families prediction cover representative species in the Basidiomycota, Ascomycota, and Zygomycota. Protein sets were retrieved from the JGI MycoCosm Portal and the Broad Institute Portal. Multigene families were predicted from 185,398 predicted proteins found in *R. irregularis* and the other representative fungal genomes using the MCL Markov clustering algorithm (35) with inflation parameter set to 2.0. As a result, 7,689 protein families containing at least five sequences were identified. Multigene families were analyzed for evolutionary changes in protein family size using the CAFE program (36) (**Fig. S13**). The program uses a random birth and death process to model gene gain and loss

across a user specified tree structure. The distribution of family sizes generated under the random model provides a basis for assessing the significance of the observed family size differences among taxa (*p*-value <0.001). CAFE estimates, for each branch in the tree, whether a protein family has not changed, has expanded or has contracted. The phylogenetic tree was constructed using maximum likelihood analyses of a concatenated alignment of 85 core genes representative of conserved fungal genes (FUNYBASE, http://genome.jouy.inra.fr/funybase/).

In *R. irregularis*, 1,014 families were expanded, 5,440 showed no change and 1,235 families had undergone contraction by comparison to a putative most recent common ancestor (MRCA). **Tables S5** and **S8** present contracted and expanded gene families with lower *p*-value in *R. irregularis* genome, respectively. The PFAM domains for these families were associated by homology searches using hmmscan from HMMER3 package (37) and the Pfam database (http://pfam.sanger.ac.uk/; (38)).

Secreted proteins were identified using a custom pipeline including the SignalP v4 (39),WolfPSort (40), TMHMM, TargetP (41), and PS-Scan algorithms (42). The secreted peptidases were identified in the genome using the MEROPS peptide database (http://merops.sanger.ac.uk/; (43); http://merops.sanger.ac.uk). Prediction of genes coding for polyketide synthases (PKS), modular nonribosomal peptide synthetases (NRPS), terpene cyclases, and dimethylallyl diphosphate tryptophan synthases (DMATS) was performed using the SMURF database (http://jcvi.org/smurf/index.php; (44)). The genes coding for protein kinases (kinome) from *R. irregularis* and other selected fungi were annotated and classified into one of the 12 major eukaryotic protein kinase groups using HMMSCAN searches from the HMMER3 package against multilevel HMM libraries from the Kinomer 1.0 database (http://www.compbio.dundee.ac.uk/kinomer; (45)) with a HMM bit score cutoff of 50.

Identification of carbohydrate active enzyme (CAZyme) genes in the *R. irregularis* genome, as well as in multiple genomes that were used for comparative analysis, was achieved using the CAZy database (www.cazy.org) annotation pipeline (46)). For each fungus, we have listed the number of representatives of each CAZy family and then performed a double clustering based on Bray-Curtis distances (i) between organisms according to their family distribution and (ii) between families according on their distribution pattern in the different genomes. Distances were computed using GINKGO (47) and the distance trees were constructed with FastME (http://www.atgc-montpellier.fr/fastme/).

MATA-HMGs proteins (MAT-HMG) representative of different fungal phyla were searched in the *R. irregularis* genome using reciprocal Blast procedures (i.e. BlastX, tBlastX, BlastP, tBlastN). For comparisons, similar searches were also performed across publicly available genome sequence data from members of the Zygomycota, Basidiomycota, and Ascomycota. All

potential MAT-HMGs identified in genome data from *Rhizophagus* spp. (this study) and in other available fungal genomes were further compared against the GenBank nr database in order to confirm their homology, and their sequences were manually inspected to avoid redundancy.

We used a set of rules as described previously (48) to annotate all transcription associated proteins (TAPs; comprising transcription factors, TF, and transcriptional regulators, TR) in the proteome. The total complement of TAPs consists of 1,439, which is the largest complement as compared with other fungi (**Table S7**). The number of HMG and TRAF TRs encoded by *R. irregularis* is exceptionally high as compared to most other organisms in this study (**Table S7**). HMG domain proteins are discussed in the main text. In mammals, the tumor necrosis factor receptor-associated factor (TRAF) TRs couple receptor proteins to signaling cascades. They interact with a variety of proteins that regulate receptor-induced cell fate decisions and are involved in innate immunity signaling.

Circular visualizations (**Fig. 2, Fig. S3 and S4**) were constructed using CIRCOS version 0.62 (49). Figures S5 and S7 were constructed using Integrative Genomics Viewer v2.2 (50). Phylogenetic reconstruction, expression profiling and protein domains of full-length *Rhizophagus*-specific tyrosine kinase-like proteins as shown in **Figure S14** was plotted using iTOL (51).

1.7. RNA extraction, RNA-Seq, clustering, and assembly

Illumina sequencing of RNA from spores and mycorrhizal roots. Total RNA was extracted using the RNeasy plant mini kit (Qiagen) following the manufacturer recommendations before quantification with the Nanodrop fluorospectrometer. RNA quality was determined by electrophoresis on an Agilent Bioanalyzer. RNA-Seq libraries were constructed using the standard Illumina TruSeq RNA protocol to sequence mRNA (P.N. RS-122-2001). mRNA were purified using oligodT-containing beads. Then, RNA was fragmented to generate double stranded cDNA for sequencing. Twelve cycles of PCR were applied to amplify libraries, and size selection was performed on E-gel (Invitrogen). Libraries were quantified by qPCR using the KAPA Library Quantification Kit (PN KK4824) to obtain an accurate quantification. Sequencing was carried out at the GeT-PlaGe facility (https://genomique.genotoul.fr, Toulouse, France) using the Illumina HiSeq 2000 instrument and the Illumina TruSeq SBS sequencing kit v3 (PN FC-401-3001) (100 bp-paired reads).

RNA-Seq for assessing the intra- and interspecific variability in the transcriptome. Total RNA was extracted from *in vitro* cultures of *R. irregularis* and *R. diaphanus* using the RNeasy plant mini kit (Qiagen) following the manufacturer recommendations before quantification with the

Nanodrop fluorospectrometer. RNA-Seq libraries were constructed using the standard Illumina TruSeq RNA protocol to sequence mRNA. mRNA were purified using oligodT-containing beads. Then, RNA was fragmented to generate double stranded cDNA for sequencing. Sequencing was carried out by Fasteris (Geneva, Switzerland) using one complete channel on the Illumina HiSeq 2000 instrument. The sequencing procedure yielded 201,051,108 and 176,382,504 100 bp-paired reads for *R. diaphanus* and *R. irregularis* C2, respectively. RNA-Seq reads are available at the NCBI Short Read Archive: Accession n° SRX312982 and SRX312214 for *R. diaphanum* and *R. irregularis* C2, respectively.

Reads were assembled using the Velvet Oases software (http://www.ebi.ac.uk/~zerbino/oases/), which builds a hash table of all possible *k*-mer (sequence of "k" bases) in the dataset and through de Bruijn graph construction and repeat resolution build de novo contigs. In opposition to VELVET contigs, OASES transcripts can share parts of their sequences between two variants of a given locus. Several values of *k*-mer (only odds number allowed) were used to optimize the assembly, with has values ranging from 89 to 95 used in the assembly transcriptome assembly process for both strains. The assembly resulting from an hash value of 93 was found to be the most optimal, and resulted in the acquisition of approximately 20,000 contigs for each strain. To validate the *de novo* assemblies and estimate the number of reads assembled, a BWA mapping was carried on the *de novo* contigs from the best Oases assembly with 100K pairs of each library. As could have been expected, the library map better on their respective assembly than on the combined one. Furthermore, the sum of the assembly is close to the statistics of the combined assembly.

Gene expression and differential expression analysis. To obtain expression (transcript) profiles for germinated spores in *R. irregularis* DAOM-197198, *R. irregularis* C2 and *R. diaphanum*, the RNA-Seq reads were separately mapped against the reference genome Gloin1 using TopHat v2.0.8 (52). The gene expression levels for each annotated gene model were then estimated as the number of Fragments Per Kilobase of exon per Million fragments mapped (FPKM) in exonic regions with Cuffdiff v2.0.2 (53) using upper quartile normalization and multi-mapped read correction. To detect symbiosis-related genes (**Table S11, S12, S14, S16**), transcript levels of *R. irregularis* genes expressed in *R. irregularis-M. truncatula* mycorrhizal roots and spores germinated for 2 or 9 days were assessed as described above and differential expression was obtained using the Cuffdiff method (54) with a FDR-adjusted *p*-value cutoff of 0.05. All experiments were carried out in triplicates.

1.8. Identification of large-scale genome duplications

In order to identify possible large scale duplication events, we employed the KeyS software as previously described (55). For the clustering of paralogous genes the minimal connectivity threshold used was 50% (half linkage). The distribution of the synonymous substitutions (K*s*) of the paralogues (**Fig. S8**) did not follow the expected steep exponential decay pattern *(56),* but rather exhibited a large right shoulder. This distribution pattern could be due to a much lower loss rate than typically observed (56), or due to several hidden large scale duplications contributing to the distribution. With regard to the former scenario, we carried out a regression analysis using y = a * exp(-x * d), where d is the decay or gene loss parameter (*56*). This resulted in $R^2 = 0.97$ and d = 0.8. Since typical d values are an order of magnitude larger (7.0-23.9) (56), we also investigated whether hidden distributions might contribute to the observed distribution, by using the EMMIX software *(57)* to fit a mixture model of normal distributed components to the gene cluster distribution data. The purpose of the mixture model is to identify the mixed normal distributions that best describe the observed K*s* distribution *(58).* The assumed mixed distributions were modeled with one to ten components and the EM algorithm was repeated 100 times with random starting values and 10 times with k-mean starting values. Additionally the algorithm was set to finish when the proposed model acquired an insignificant *p*-value. The criterion used to identify the best fitting model was the minimization of the Bayesian Information Criterion in addition to a significant *p*-value for the model. The proposed mixture model contains eight components, with average K*s* values between 0.05 and 4.74 (**Fig. S8**). The R^2 for the eight components is 0.94-0.96, i.e. in the same range as those of the exponential decay function. In the following, we analysed GO bias based on the eight components. However, it should be noted that exponential decay with d = 0.8 is also a possible explanation for the observed distribution, as mentioned above. Yet, the GO bias pointed out below for the Ks range ~ 0.18 to 4.74 is valid under either scenario.

To check whether the potential components were functionally biased, a GO bias analysis was conducted on the genes participating in the different components of the proposed mixture model, using the GOstats *(59)* R package. The *p*-value threshold was 0.01, the resulting *p*-values were false-discovery corrected *(60).* The Biological Process GO terms that were either over-represented per component were visualized (**Fig. S9**). The significantly biased GO terms for all three ontologies and the genes associated with these terms are shown in a series of Supplementary Tables that are downloadable from the INRA *Rhizophagus* web portal at: http://mycor.nancy.inra.fr/IMGC/GlomusGenome/index3.html. Also, all GO graphs are downloadable for all significantly biased terms; here, over-represented terms are shown in green, under-represented terms in red .

Annotation of genes phosphorus-related from the 8 EMMIX components. In the components 1, 2, 3, 4, 5, 7 and 8 (*i.e.*, in all but the component with the lowest Ks, 6) we observed that phosphorus-related terms are consistently over-represented. After isolating the genes from the corresponding components that cause the bias towards phosphorus-related terms, we counted the number of these genes that are annotated as kinases (GO:0004672, protein kinase activity; http://mycor.nancy.inra.fr/IMGC/GlomusGenome/index3.html). As can be seen from the tables, most genes are annotated as kinases. The remainder are either annotated under the term phosphorylation (some of them might also be kinases), or as phosphate-containing compound metabolic process (eight genes).

The GO terms that appear in the GO Fisher's test related with phosphorus are GO:0006468 (Protein Phosphorylation), GO:0016310 (Phosphorylation), GO:0006796 (Phospate-containing compound metabolic process) and GO:0006793 (Phosphorus metabolic process). Based on the GO and the KOG protein annotation, the file Gloin1_KOG_GO_annotation.txt was created, that combines the two annotations. The format of the file is tab-delimited, with 4 columns, the first having the protein id, the second the KOG annotation, the third the GO annotation and the fifth the number of GO terms that correspond to each gene.

In order to identify which genes out of each EMMIX component are related with the phosphorus-related terms, a python script was implemented that takes as input the genes with the phosphorus-related annotation and the genes from each of the components of the EMMIX output and produces files similar to the Gloin1_KOG_GO_annotation.txt, but only with the genes of interest. The table below contains the number of genes that are phosphorus-related for each component.

Component Number	**Phosphorus-related genes**
1	280
2	465
3	669
4	132
5	201
7	393
8	670

The GO term GO:0004672 (protein kinase activity) was used to identify the genes that are also annotated as kinases, since just by using the word 'kinase' in the description we might lose some genes. In the table below we have the number of phosphorus-related genes that are annotated as kinases.

Component Number	Phosphorus-related genes	Kinase-annotated genes
1	280	266
2	465	451
3	669	650
4	132	128
5	201	201
7	393	392
8	670	653

Paralog clustering for all genes showed that there were only nine near-identical paralogs (>=99.5% identical on n.a. level, without leading/trailing gaps). In total there were 7,954 paralogs in the assembly. A similar analysis identified 1,357 and 12,360 paralogs in *Arabidopsis* and *Physcomitrella*, respectively.

Table S1 to S16

Table S1. The diversity and distribution of class I and class II transposable elements in the *R. irregularis* genome assembly. Distribution in various TE families, number of copies and % genome assembly coverage are shown. LINE, long interspersed nuclear element;

Class	Superfamily	# consensus	# copies	% genome
Class I	**LINE**	11	232	0.18
	Copia	1	25	0.01
	Gypsy	9	164	0.10
Total Class I		21	421	0.29
Class II	**TIR**	69	1183	0.57
	Polinton/*Maverick*	10	157	0.10
	Helitron	8	78	0.04
Total Class II		87	1,418	0.71
Uncategorized		1,904	68,207	10.30
Total		2,012	70,046	11.30

Table S2. Genes with the highest number of high quality SNPs in the *R. irregularis* genome assembly. SNP calling was carried out by mapping genomic and transcriptome reads to the genome assembly by using SAMtools v2 and CLC Genomic Workbench v6.5.

Protein ID	**IPR domain**	**# SNPs** (SAMtools)	**# SNPs** (CLC*)	**# SNPs** (CLC**)
9966	Protein kinase-like	9	5	4
15612	Protein kinase, core	4	5	4
12595	Carbohydrate-binding-like fold	4	8	2
1986	BTB/POZ***	3	4	3
348164	Peptidase S10	3	4	2
14919	No hit	2	3	3
2038	BTB/POZ	2	3	3
1905	BTB/POZ	2	3	2
348638	Cytochrome P450	2	5	4

* Probabilistic variant detection algorithm in CLC Genomic Workbench v6.5 (http://www.clcsupport.com/clcgenomicsworkbench/650/index.php?manual=Probabilistic_variant_detection.html)

**Quality based variant detection in CLC Genomic Workbench v6.5. (http://www.clcsupport.com/clcgenomicsworkbench/650/index.php?manual=Quality_based_variant_detection.html#sec:SNPdetection)

*** The BTB/POZ: BTB (for BR-C, ttk and bab) or POZ (for Pox virus and Zinc finger) domain; the BTB/POZ domains from several zinc finger proteins have been shown to mediate transcriptional repression and to interact with components of histone deacetylase co-repressor complexes

Table S3. Properties of *Rhizophagus irregularis* gene models.

Property or number	Value
Avg. gene length	1,188 nt
Avg. transcript length	890 nt
Avg. protein length	270 aa
Avg. exon length	257 nt
Avg. intron length	123 nt
Avg. exon frequency	3.5 exons per gene
# multiexon genes	2,3251 (77%)
# genes with similarity to protein in nr	17,848 (59%)
# genes in DAOM-197198 multigene family	21,996 (73%)
# genes with RNA coverage	20,790 (69%)
# genes with Pfam domain	10,020 (33%)
# genes with signal peptide	1,995 (7%)
# genes with transmembrane domain	3,655 (12%)
# genes with EC number	2,750 (9%)
# genes with GO term	9,591 (32%)

Abbreviations: Avg., average; aa, amino acids; GO, Gene Ontology; nr, nonredundant database in GenBank; nt, nucleotides

Table S4. Number and percentage of expressed genes in *Rhizophagus irregularis* DAOM-197198, *R. irregularis* C2 and *R. diaphanum* MUCL43196. Expression profiles were obtained by mapping RNA-Seq reads of germinated spores from *R. irregularis* DAOM-197198, R. *irregularis* C2 and *R. diaphanum* MUCL43196 on the *R. irregularis* genome assembly.

	% mapped reads	**# expressed genes**	**% expressed genes**
Spores DAOM-197198	84.0	20,191	71.5
Spores *R. irregularis* C2	70.6	18,690	66.2
Spores *R. diaphanum* MUCL43196	58.4	16,555	58.6
All strains		*22,299*	*78.9*
Spores DAOM-197198 2d	76.9	17,883	63.3
Spores DAOM-197198 9d	78.1	18,324	64.9
Spores DAOM-197198 *in planta*	5.3	13,463	47.6
All		*22,647*	*80.2*

Table S5. The top 40 protein families in expansion (excluding transposon-related families) *in R. irregularis* genome as compared to representative fungi.

Family	Pucgr	Ustma	Copci	Lacbi	Tubme	Aspni	Botci	Neucr	Sacce	Morel	Rhiir	Rhior	PFAM description
0	0	0	0	0	0	0	0	0	0	2	1538	2	**Tyrosine protein kinase**
2	5	5	4	5	5	4	2	4	4	138	1066	21	**Sel1-like; Serine/threonine protein kinase**
5	0	1	0	0	0	0	1	0	0	1	915	0	**BTB/POZ**
17	0	0	0	0	0	0	0	0	0	0	389	0	
30	0	0	0	0	0	0	0	0	0	0	281	0	
44	0	0	0	0	0	0	0	0	0	0	202	0	
67	0	1	0	0	0	0	0	0	0	4	153	2	**Kelch repeat**
31	0	0	1	1	9	11	11	5	0	15	149	8	
81	0	0	0	0	0	0	0	0	0	2	134	0	**Sel1-like; Tyrosine protein kinase**
88	0	0	0	0	0	0	0	0	0	0	128	0	
91	0	0	0	0	0	0	0	0	0	0	125	0	**Cytochrome P450**
103	0	0	0	0	0	0	0	0	0	0	117	0	
112	0	0	0	0	0	0	0	0	0	0	111	0	
113	0	0	0	0	0	0	0	0	0	0	109	0	
115	0	0	0	0	0	0	0	0	0	0	108	0	
120	0	0	0	0	0	0	0	0	0	0	106	0	
127	0	0	0	0	0	0	0	0	0	0	103	0	
138	0	0	0	0	0	0	0	0	0	0	97	0	
145	0	0	0	0	0	0	0	0	0	0	93	0	
33	1	0	7	6	1	1	1	1	0	15	93	62	**Methyltransferase type 11**
132	5	0	0	4	0	0	0	0	0	0	92	0	
167	0	0	0	0	0	0	0	0	0	0	86	0	
176	0	0	0	0	0	0	0	0	0	0	84	0	**BTB/POZ**
183	0	0	0	0	0	0	0	0	0	0	83	0	
187	0	0	0	0	0	0	0	1	0	0	81	0	
218	0	0	0	0	0	0	0	0	0	0	74	0	
238	0	0	0	0	0	0	0	0	0	0	70	0	**Guanylate-binding protein**
258	0	0	0	0	0	0	0	0	0	0	67	0	
9	12	12	45	51	16	102	107	24	1	25	67	25	**Cytochrome P450**
266	0	0	0	0	0	0	0	0	0	0	65	0	**Zinc finger, C2H2-type**
286	0	0	0	0	0	0	0	0	0	0	61	0	
282	0	0	0	0	0	0	0	0	0	0	59	0	**Tetratricopeptide TPR1/TPR2**
298	0	0	0	0	0	0	0	0	0	0	59	0	
299	0	0	0	0	0	0	0	0	0	0	59	0	**Tyrosine protein kinase**
314	0	0	0	0	0	0	0	0	0	0	58	0	
59	5	4	6	18	8	5	6	7	6	14	58	13	**Ubiquitin; Ribosomal protein S27a**
339	0	0	0	0	0	0	0	0	0	0	55	0	
41	26	0	4	12	0	0	0	0	0	0	55	30	
362	0	0	0	0	0	0	0	0	0	0	53	0	**Tyrosine protein kinase**
363	0	0	0	0	0	0	0	0	0	0	53	0	**High mobility group box, HMG1/HMG2**

The table lists TRIBE-MCL families that are in expansion in the *R. irregularis* lineage (gray row, CAFE analysis, P<0.001) (Fig. S14). Annotations are based on searches of T. melanosporum protein sequences against the PFAM database. Abbreviations: Pucgr: *Puccinia graminis*; Ustma: *Ustilago maydis*; Copci: *Coprinopsis cinerea*; Lacbi: *Laccaria bicolor*; Tubme: *Tuber melanosporum*; Aspni: *Aspergillus niger*; Botci: *Botrytis cinerea*; Neucr: *Neurospora crassa*; Sacce: *Saccharomyces cerevisiae*; Rhior: *Rhizopus oryzae*; Morel: *Mortierella elongata*; Rhiir: *Rhizophagus irregularis.*

Table S6. The repertoire of protein kinases (kinome) in *R. irregularis* (gray row) compared to the kinomes of representative fungi. Protein kinase domains have been identified by HMMER search against the KINOMER database (http://www.compbio.dundee.ac.uk/kinomer, multilevel HMM library); score >50. Kinases are classified into major groups.

	AGC	**CAMK**	**CK1**	**CMGC**	**RGC**	**STE**	**TK**	**TKL**	**PIKK**	**PDHK**	**RIO**	**Alpha**
Puccinia graminis	25	32	7	22	0	17	0	3	3	3	2	**17**
Ustilago maydis	16	31	2	20	0	15	0	2	2	3	2	0
Coprinopsis cinerea	24	30	5	29	0	21	0	15	3	3	2	0
Laccaria bicolor	25	38	12	45	0	18	0	42	3	3	2	0
Saccharomyces cerevisiae	22	40	4	25	0	15	0	1	4	2	1	0
Tuber melanosporum	19	28	2	23	0	13	0	1	3	3	2	0
Aspergillus niger	18	27	3	43	0	16	0	0	**5**	3	2	0
Botrytis cinerea	17	25	1	20	0	14	0	2	4	3	2	0
Neurospora crassa	18	29	3	23	0	14	0	2	4	3	2	2
Rhizopus oryzae	**58**	**93**	**18**	**50**	0	**55**	0	3	**5**	**6**	2	0
Rhizophagus irregularis	25	41	3	24	0	22	0	**832***	4	3	2	13
Dictyostelium discoideum	27	41	3	35	0	45	0	72	4	0	2	6

*88 TKL with whole catalytic domain; 812 expressed TKLs.

Abbreviations: TKL, tyrosine kinase-like proteins. AGC, protein kinases A, G, and C; CAMK, calcium/calmodulin-dependent kinases; CK1, casein kinase 1; CMGC, cyclin- dependent kinases, mitogen-activated, glycogen-synthase, and CDK-like kinases; STE, sterile phenotype kinases; RGC, receptor guanylate cyclase kinases ; TK, tyrosine kinases ; PIKK, phosphatidyl inositol 30 kinase-related kinases ; PDHK, pyruvate dehydrogenase kinases ; RIO, right open reading frame kinases ; Alpha, Alpha kinases.

Table S7. Distribution of genes involved in transcription in the genomes of *Rhizophagus irregularis* and representative fungi.

Transcription Factors	*Puccinia graminis*	*Ustilago maydis*	*Coprinopsis cinerea*	*Laccaria bicolor*	*Saccharomyces cerevisiae*	*Tuber melanosporum*	*Aspergillus niger*	*Botrytis cinerea*	*Neurospora crassa*	*Rhizopus oryzae*	***Rhizophagus irregularis***	*Dictyostelium discoideum*
AP2/EREBP	0	0	0	0	0	0	0	0	0	0	0	0
Argonaute	2	0	7	6	0	3	1	1	2	2	26	5
ARID	2	2	2	2	2	2	2	4	3	2	2	2
bHLH	7	12	10	16	7	8	13	9	15	52	11	0
BSD domain containing	2	1	2	1	2	2	2	2	2	2	2	3
bZIP	8	7	10	9	12	27	16	11	12	48	17	19
C2C2_GATA	5	9	10	12	10	6	6	6	6	32	19	20
C2H2	72	30	64	62	39	58	65	61	68	121	95	27
C3H	9	15	21	26	7	12	14	15	11	24	19	20
CCAAT_Dr1	1	1	1	1	1	1	1	0	1	2	1	1
CCAAT_Hap2	1	1	1	0	1	1	1	1	1	2	3	1
CCAAT_Hap3	1	2	2	2	1	2	2	3	2	2	1	2
CCAAT_Hap5	2	4	3	2	3	3	2	2	5	3	4	5
Coactivator p15	0	0	1	1	1	0	1	1	1	1	1	2
CPP	0	0	0	0	0	0	0	0	0	0	0	1
CSD	3	0	0	0	0	0	1	0	0	0	6	0
DDT	0	1	1	1	2	0	1	1	1	2	0	1
Dicer	1	0	0	0	0	0	0	0	0	0	1	0
DUF547 domain containing	0	0	0	0	0	0	0	0	0	0	0	1
DUF833 domain containing	1	1	1	1	1	1	1	0	1	0	1	0
E2F/DP	0	0	0	0	0	0	0	0	0	3	2	3
FHA	4	4	9	10	14	9	9	9	9	18	10	20
GIF	0	0	0	0	0	0	0	0	0	0	0	0
GNAT	10	19	27	23	12	17	34	30	25	28	21	35
HB	8	6	10	10	8	4	4	6	5	37	9	14
HMG	11	8	25	15	7	8	7	8	10	27	126	4
HSF	6	4	3	4	5	3	3	3	3	17	4	1

IWS1	1	1	1	1	1	1	1	1	1	1	0	1
Jumonji	1	1	1	1	1	1	1	1	1	1	1	0
LIM	5	5	5	6	3	4	4	4	4	9	4	15
MADS	2	2	1	3	4	2	2	1	2	9	2	4
MBF1	1	0	1	1	1	1	1	1	1	2	1	0
MED6	1	1	1	1	1	1	1	1	1	1	2	1
MED7	0	1	1	1	1	2	1	1	1	1	1	1
MYB-related	4	4	7	5	10	8	7	10	9	15	24	16
MYB	2	5	6	8	5	7	8	5	8	10	19	13
PcG_FIE	0	0	0	0	0	1	0	0	0	0	3	0
PcG_VEFS	0	0	0	0	0	1	0	0	0	0	1	0
PHD	8	12	14	13	10	9	15	11	14	22	16	11
RB	0	0	0	0	0	0	0	0	0	0	0	1
Rcd1-like	1	1	1	1	1	1	1	1	1	1	1	2
RF-X	1	1	1	1	1	1	1	0	1	0	0	0
RRN3	1	1	1	1	1	1	1	1	1	1	1	1
RWP-RK	0	0	0	0	0	0	0	0	0	0	0	2
SET	8	5	7	15	5	7	8	7	9	12	12	11
Sin3	2	1	4	4	1	1	1	0	1	6	2	1
Sir2	4	4	8	8	5	12	6	6	7	4	5	5
SOH1	0	1	1	1	1	0	1	1	1	0	1	1
SWI/SNF_BAF60b	2	1	2	3	3	2	2	2	2	5	2	2
SWI/SNF_SNF2	48	17	24	19	17	17	23	24	25	25	15	22
SWI/SNF_SWI3	0	0	0	0	2	2	1	1	2	4	2	0
TEA	0	0	0	0	1	1	1	0	0	1	0	0
TFb2	2	1	1	2	1	1	1	1	1	1	1	1
TRAF	1	2	15	10	1	6	5	13	6	15	536	20
Trihelix	0	0	0	0	0	0	0	0	0	0	3	0
TUB	0	0	0	0	0	0	0	0	0	1	1	0
WRKY	0	0	0	0	0	0	0	0	0	3	1	1
Zinc finger, AN1 & A20 type	0	0	1	1	2	1	1	1	1	4	5	4

Zinc finger, MIZ type	4	1	1	2	3	2	3	2	2	1	0	7
Zinc finger, ZPR1	1	1	1	1	1	1	1	1	1	1	1	1
Zn_cluster	24	85	41	65	53	58	235	103	95	30	20	2

Table S8. The top 30 protein families in contraction (excluding TE-related families) *in R. irregularis* genome as compared to other fungi.

Family ID	Pucgr	Ustma	Copci	Lacbi	Tubme	Aspni	Botci	Neucr	Sacce	Morel	Rhiir	Rhior	PFAM description
1	74	52	76	84	59	58	59	60	75	140	75	214	**Serine/threonine protein kinase**
3	42	37	53	150	35	49	43	37	33	224	54	73	**WD40 repeat**
37	16	7	5	12	3	11	16	17	6	36	9	33	**Peptidase A1**
47	17	6	16	11	6	3	5	3	2	29	8	34	**Polysaccharide deacetylase**
58	5	6	7	7	6	9	5	6	11	23	7	31	**Zinc finger, C2H2-type**
166	0	0	0	0	0	0	0	0	0	77	2	1	**Leucine-rich repeat ; Cyclin-like F-box**
179	0	1	2	4	0	8	11	0	2	45	2	1	
123	3	3	6	9	1	17	8	3	2	8	1	13	**NADH:flavin oxidoreductase**
154	16	2	8	8	2	10	4	6	1	8	1	8	**Glycoside hydrolase, family 18**
447	0	1	1	1	0	0	1	0	0	25	1	6	**Peptidase M13**
52	0	0	0	0	1	10	5	2	0	172	0	0	**Monooxygenase, FAD-binding**
32	0	0	0	0	0	0	0	0	0	11	0	35	**Zinc finger, CCHC-type**
48	0	0	0	0	0	0	0	0	0	11	0	41	**Zinc finger, SWIM-type**
279	0	0	0	0	0	0	0	0	0	22	0	12	
63	0	7	10	13	2	41	24	9	4	17	0	16	**Major facilitator superfamily MFS-1**
68	12	8	6	11	6	24	19	10	18	24	0	10	**Amino acid permease**
172	0	0	0	0	0	0	0	0	0	85	0	0	
206	0	0	0	0	0	0	0	0	0	76	0	0	**Kelch repeat**
93	4	6	4	9	5	34	21	4	4	8	0	9	**Amino acid permease**
225	0	0	0	0	0	0	0	0	0	72	0	0	**Pentapeptide repeat; WD40 repeat**
249	0	0	0	0	0	0	0	0	0	68	0	0	**Cyclin-like F-box**
270	0	0	0	0	0	0	0	0	0	64	0	0	
62	6	8	19	15	3	44	39	9	0	27	0	1	**Carboxylesterase, type B**
333	0	0	0	0	0	0	0	0	0	55	0	0	
361	0	0	0	0	0	0	0	0	0	53	0	0	**Cyclin-like F-box**
384	0	0	0	0	0	0	0	0	0	51	0	0	**Cyclin-like F-box**
442	0	0	0	0	0	0	0	0	0	8	0	15	**Spore coat protein CotH**
842	0	0	1	1	1	0	0	1	0	7	0	8	

The table lists TRIBE-MCL families that are in contraction in the *R. irregularis* lineage (gray row, CAFE analysis, P<0.001) (Fig. S14). Abbreviations: Pucgr: *Puccinia graminis* ; Ustma: *Ustilago maydis* ; Copci: *Coprinopsis cinerea* ; Lacbi: *Laccaria bicolor* ; Tubme: *Tuber melanosporum* ; Aspni: *Aspergillus niger* ; Botci: *Botrytis cinerea* ; Neucr: *Neurospora crassa* ; Sacce: *Saccharomyces cerevisiae* ; Rhior: *Rhizopus oryzae* ; Morel: *Mortierella elongata* ; Rhiir: *Rhizophagus irregularis*

Table S9. *Saccharomyces cerevisiae* proteins missing in the obligate biotrophic pathogen, *Blumeria graminis*, the obligate symbiont *Rhizophagus irregularis* and other representative saprotrophic or symbiotic fungi. BLASTP e-value, 1^{e}-5. Yeast proteome was obtained via Saccharomyces Genome Database (http://downloads.yeastgenome.org/sequence/S288C_reference/orf_protein). This missing ascomycete core genes (MACGs) procedure has been adapted from Spanu *et al.* (61). Blugr, *Blumeria graminis*; Pucgr, *Puccinia graminis*; Ustma, *Ustilago maydis*; Copci, *Coprinopsis cinerea*; Lacbi, *Laccaria bicolor*; Tubme; *Tuber melanosporum*; Aspni, *Aspergillus niger*; Botci, *Botrytis cinerea*; Neucr, *Neurospora crassa*; Mucci, *Mucor circinelloides*; Phybl, *Phycomyces blakesleeanus*; Rhior: *Rhizopus oryzae*; Batde, *Batrachochytrium dendrobatidis;* Rhiir, *Rhizophagus irregularis*.

Name	Description	Blugr	Pucgr	Ustma	Copci	Lacbi	Tubme	Aspni	Botci	Neucr	Mucci	Phybl	Rhior	Batde	Rhiir
	Thiamine metabolism/transport														
YGR144W	Thiamine thiazole synthase	No	yes	yes	yes	No	No	yes	yes	yes	yes	No	yes	No	No
YPL214C	Thiamine biosynthetic bifunctional enzyme	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	No
YLR237W	Thiamine transporter	No	No	yes	No	yes	yes	yes	yes	yes	No	No	No	No	No
YOL055C	Hydroxymethylpyrimidine/phosphomethylpyrimidine kinase THI20	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	No	No
YPL258C	Hydroxymethylpyrimidine/phosphomethylpyrimidine kinase THI21	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	No	No
YPR121W	Thiamine biosynthesis protein THI22	No	yes	yes	yes	yes	yes	yes	No	yes	yes	yes	yes	No	No
YOR192C	Thiamine transporter THI72	No	No	yes	No	yes	yes	yes	yes	yes	No	No	No	No	No
YOR071C	Nicotinamide riboside transporter 1	No	No	yes	No	yes	yes	yes	yes	yes	No	No	No	No	No
	Allantoine metabolism/transport														
YIR027C	Allantoinase	No	yes	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	No	yes
YIR029W	Allantoicase	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
YIR028W	Allantoin permease	No	No	yes	No	yes	yes	yes	yes	yes	No	No	No	No	No
YIR023W	Transcriptional activator protein DAL81	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
YHL016C	Urea active transporter	No	No	yes	yes	yes	yes	yes	yes	yes	yes	No	yes	No	yes
	Methionine metabolism and (siro-)heme biosynthesis														
YKR069W	Uroporphyrinogen-III C-methyltransferase	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	No	yes

YJR010W	Sulfate adenylyltransferase	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	No	yes
YBR213W	Siroheme biosynthesis protein MET8	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	No	yes
YKL001C	Adenylyl-sulfate kinase	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	No	yes
YPR167C	Phosphoadenosine phosphosulfate reductase	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	No	yes
YOR278W	Uroporphyrinogen-III synthase	No	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Alcohol metabolism/fermentation														
YGL256W	Alcohol dehydrogenase 4	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	No	No
YCR107W	Putative aryl-alcohol dehydrogenase AAD3	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
YDL243C	Probable aryl-alcohol dehydrogenase AAD4	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
YFL056C	Putative aryl-alcohol dehydrogenase AAD6	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
YJR155W	Putative aryl-alcohol dehydrogenase AAD10	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
YNL331C	Putative aryl-alcohol dehydrogenase AAD14	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
YOL165C	Putative aryl-alcohol dehydrogenase AAD15	No	No	yes	yes	yes	No	yes	yes	yes	yes	No	yes	No	yes
YFL057C	Putative aryl-alcohol dehydrogenase AAD16	No	No	yes	yes	yes	No	yes	yes	yes	yes	yes	yes	No	yes
YPL088W	Putative aryl-alcohol dehydrogenase YPL088W	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
YLR044C	Pyruvate decarboxylase isozyme 1	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	No	yes
YOL086C	Alcohol dehydrogenase 1	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Glutamate metabolism														
YOR375C	NADP-specific glutamate dehydrogenase 1	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
YAL062W	NADP-specific glutamate dehydrogenase 2	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Uracil metabolism/transport														
YBL042C	Uridine permease	No	No	yes	No	yes	yes	yes	yes	yes	No	No	No	No	No

YBR021W	Uracil permease	No	No	yes	No	yes	yes	yes	yes	yes	No	No	No	No	No
YKL216W	Dihydroorotate dehydrogenase	No	No	yes	yes	yes	yes	No	yes	yes	yes	yes	yes	yes	yes
	Glutathione metabolism														
YGR154C	Glutathione S-transferase omega-like 1	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	No	yes
YMR251W	Glutathione S-transferase omega-like 3	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	No	yes
YKR076W	Glutathione S-transferase omega-like 2	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	No	yes
YLR299W	Gamma-glutamyltransferase	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Detoxification/stress response														
YER185W	Protoporphyrin uptake protein 1	No	No	yes	No	No	No	yes	yes	yes	No	No	No	No	No
YGR213C	Protein RTA1	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	No	yes	No	No
YJR104C	Superoxide dismutase	No	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
YGR234W	Flavohemoprotein	No	No	No	No	No	No	yes	yes	yes	No	No	No	No	No
YIL053W	(DL)-glycerol-3-phosphatase 1	No	No	yes	yes	yes	yes	yes	No	yes	yes	yes	No	No	No
YPR201W	Arsenical-resistance protein 3	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	No	No
YGL196W	D-serine dehydratase	No	No	No	yes	No	yes	yes	yes	yes	yes	No	yes	No	No
YHR044C	2-deoxyglucose-6-phosphate phosphatase 1	No	No	yes	yes	yes	yes	yes	No	yes	yes	yes	No	No	No
YHR043C	2-deoxyglucose-6-phosphate phosphatase 2	No	No	yes	yes	yes	yes	yes	No	yes	yes	yes	No	No	No
	Arabinono-1,4-lactone biosynthesis														
YML086C	D-arabinono-1,4-lactone oxidase	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	No	yes
YMR041C	D-arabinose 1-dehydrogenase	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Chaperones														
YBR227C	Mitochondrial clpX-like chaperone MCX1	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	No

YMR038C	Superoxide dismutase 1 copper chaperone	No	No	No	No	No	yes	yes	yes	yes	yes	yes	yes	No	yes
	Nitrate metabolism														
XP_752655	Nitrate transporter	No	No	yes	yes	yes	yes	yes	yes	yes	No	No	No	No	yes
CAD28426	Nitrite reductase	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
AAL85636	Nitrate reductase	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
	Proteases/peptidases														
YBR286W	Aminopeptidase Y	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
YHR132C	Metallocarboxypeptidase	No	No	yes	yes	yes	yes	yes	yes	yes	No	No	No	yes	No
YIL108W	Zinc metalloproteinase	No	No	No	No	No	yes	yes	yes	yes	yes	No	yes	No	yes
	Aromatic amino acid metabolism														
YGL202W	Aromatic/aminoadipate aminotransferase 1	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	No
YHR137W	Aromatic amino acid aminotransferase 2	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	No
	Channels/transporters														
YJL093C	Outward-rectifier potassium channel TOK1	No	No	yes	yes	yes	yes	yes	yes	yes	No	yes	No	No	No
YBR296C	Phosphate permease PHO89	No	No	yes	yes	No	yes	yes	yes	yes	yes	yes	No	No	yes
YIL023C	Zinc transporter YKE4	No	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	No	yes
YKL221W	Probable transporter MCH2	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	No
	Mating type/cell cycle/budding														
YBR276C	Dual specificity protein phosphatase PPS1	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
YGL056C	Protein SDS23	No	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
YBR214W	Protein SDS24	No	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
YIL140W	Axial budding pattern protein 2	No	yes	yes	yes	yes	No	yes	No	yes	No	No	No	No	No

	ER quality control														
YPL096W	Peptide-N(4)-(N-acetyl-beta-glucosaminyl)asparagine amidase	No	No	No	No	No	yes	yes	yes	yes	yes	yes	yes	yes	No
YHR176W	Thiol-specific monooxygenase	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
YBR015C	Alpha-1,2-mannosyltransferase MNN2	No	No	No	No	No	No	yes	yes	yes	yes	yes	yes	yes	No
YJL186W	Alpha-1,2-mannosyltransferase MNN5	No	No	No	No	No	No	yes	yes	yes	yes	yes	yes	yes	No
	Others														
YLL057C	Alpha-ketoglutarate-dependent sulfonate dioxygenase	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	No	No
YDR465C	Arginine N-methyltransferase 2	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
YNL229C	Transcriptional regulator URE2	No	No	yes	yes	yes	yes	yes	yes	yes	No	No	No	yes	yes
YOR388C	Formate dehydrogenase 1	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
YDR242W	Probable amidase	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	No	yes
YMR302C	Mitochondrial escape protein 2	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
YJL145W	Phosphatidylinositol transfer protein SFH5	No	No	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
YLR047C	Probable ferric reductase transmembrane component 8	No	No	yes	yes	yes	yes	yes	yes	yes	No	No	yes	No	No
YLR278C	Uncharacterized transcriptional regulatory protein	No	No	yes	No	No	yes	yes	yes	yes	No	No	No	No	No
YIL162W	Invertase 2	No	No	yes	No	No	yes	yes	yes	yes	No	yes	No	No	No
YDR030C	Radiation-sensitive protein 28	No	No	yes	No	yes	yes	yes	yes	yes	yes	yes	yes	yes	No

Table S10. Proteins containing a carbohydrate-binding module, carbohydrate esterases, expansin-related proteins, and glycoside hydrolases in *R. irregularis* DAOM-197198. The enzyme families are represented by their class and family number according to the Carbohydrate-Active enZYmes Database database (http://www.cazy.org/).

JGI Protein ID	Defline	CAZy module(s)
255	Carbohydrate-Binding Module Family 13 protein	CBM13
6978	Carbohydrate-Binding Module Family 13 protein	CBM13
16808	Carbohydrate-Binding Module Family 13 protein	CBM13
16818	Carbohydrate-Binding Module Family 13 protein	CBM13
63968	Carbohydrate-Binding Module Family 13 protein	CBM13
74095	Carbohydrate-Binding Module Family 13 protein	CBM13
212254	Carbohydrate-Binding Module Family 13 protein	CBM13
84949	Carbohydrate-Binding Module Family 18 protein	CBM18
45161	Carbohydrate-Binding Module Family 18 / Carbohydrate Esterase Family 4 protein	CBM18-CE4
34561	Carbohydrate-Binding Module Family 21 protein	CBM21
335833	Carbohydrate-Binding Module Family 48 protein	CBM48
70859	Carbohydrate-Binding Module 48 / Glycoside Hydrolase Family 13 protein	CBM48-GH13
341482	Carbohydrate-Binding Module Family 50 protein	CBM50
348911	Carbohydrate-Binding Module Family 50 protein	CBM50
26526	Carbohydrate Esterase Family 11 protein	CE11
20441	Carbohydrate Esterase Family 4 protein	CE4
25223	Carbohydrate Esterase Family 4 protein	CE4
89895	Carbohydrate Esterase Family 4 protein	CE4
205470	Carbohydrate Esterase Family 4 protein	CE4
343321	Carbohydrate Esterase Family 4 protein	CE4
345355	Carbohydrate Esterase Family 4 protein	CE4
346280	Carbohydrate Esterase Family 4 protein	CE4
335796	Carbohydrate Esterase Family 9 protein	CE9
58723	Distantly related to plant expansins	EXPN
70056	Distantly related to plant expansins	EXPN
82162	Distantly related to plant expansins	EXPN
84458	Distantly related to plant expansins	EXPN
346642	Distantly related to plant expansins	EXPN
27716	Glycoside Hydrolase Family 108 protein	GH108
81947	Glycoside Hydrolase Family 125 protein	GH125
272160	Glycoside Hydrolase Family 125 protein	GH125
82975	Glycoside Hydrolase Family 13 protein	GH13
94949	Glycoside Hydrolase Family 13 protein	GH13
256840	Glycoside Hydrolase Family 13 protein	GH13
135429	Glycoside Hydrolase Family 15 protein	GH15
40274	Glycoside Hydrolase Family 18 protein	GH18
30711	Glycoside Hydrolase Family 19 protein	GH19
86511	Glycoside Hydrolase Family 20 protein	GH20

91889	Glycoside Hydrolase Family 23 protein	GH23
10338	Glycoside Hydrolase Family 24 protein	GH24
92943	Glycoside Hydrolase Family 24 protein	GH24
324701	Glycoside Hydrolase Family 24 protein	GH24
326114	Glycoside Hydrolase Family 24 protein	GH24
35329	Glycoside Hydrolase Family 25 protein	GH25
92712	Glycoside Hydrolase Family 26 protein	GH26
19913	Glycoside Hydrolase Family 27 protein	GH27
137379	Glycoside Hydrolase Family 27 protein	GH27
77961	Glycoside Hydrolase Family 31 protein	GH31
340187	Glycoside Hydrolase Family 31 protein	GH31
268674	Glycoside Hydrolase Family 35 protein	GH35
348856	Glycoside Hydrolase Family 35 protein	GH35
19827	Glycoside Hydrolase Family 36 protein	GH36
314624	Glycoside Hydrolase Family 37 protein	GH37
342201	Glycoside Hydrolase Family 38 protein	GH38
19084	Glycoside Hydrolase Family 47 protein	GH47
33551	Glycoside Hydrolase Family 47 protein	GH47
40562	Glycoside Hydrolase Family 47 protein	GH47
230436	Glycoside Hydrolase Family 47 protein	GH47
342669	Glycoside Hydrolase Family 47 protein	GH47
64734	Glycoside Hydrolase Family 5 protein	GH5
76762	Glycoside Hydrolase Family 5 protein	GH5
338770	Glycoside Hydrolase Family 5 protein	GH5
33915	Glycoside Hydrolase Family 63 protein	GH63
35188	Glycoside Hydrolase Family 63 protein	GH63
29553	Glycoside Hydrolase Family 9 protein	GH9
45754	Glycoside Hydrolase Family 9 protein	GH9
318916	Glycoside Hydrolase Family 9 protein	GH9
32509	Glycoside Hydrolase Family 91 protein	GH91

Table S11. The most highly upregulated transcripts in *Rhizophagus irregularis-Medicago truncatula* symbiotic roots

Protein ID	FPKM Spores 2d	FPKM Spores 9d	FPKM *in planta*	InterPro domain	Secreted protein
344204*	0.0	0.0	24600	specific protein	S
21945*	0.0	0.0	5609	hypothetical protein	S
22556	0.0	0.0	4040	specific protein	
343985	0.0	0.0	3493	specific protein	S
54643*	0.0	0.2	3163	GPR1/FUN34/yaaH	
7716	0.0	2.0	2948	specific protein	S
347085	0.0	27.9	2887	specific protein	S
336160	0.0	0.0	2686	specific protein	
334000	0.0	0.0	2659	specific protein	
248224	0.0	0.0	2524	specific protein	S
30459*	0.0	0.0	2169	alpha/beta hydrolase	
11377	0.0	0.0	1716	specific protein	S
347932	0.0	0.0	1491	specific protein	
6890	0.0	0.1	1422	specific protein	S
11536*	0.0	0.2	1373	specific protein	S
24444	0.0	0.0	1318	specific protein	S
326976	0.0	0.0	1309	specific protein	
96997	0.0	0.0	1123	specific protein	
347485	0.0	7.5	1105	specific protein	
330295*	0.0	0.0	1057	peptidase A1	S
2831	0.0	0.0	1052	hypothetical protein	
30684*	0.0	0.0	1014	cytochrome P450	
337599*	0.0	0.0	950	hypothetical protein	S
61727	0.0	0.0	911	cytochrome P450	
322915*	0.0	0.0	887	specific protein	S
343687*	0.0	7.2	882	specific protein	
348622	0.0	0.0	871	MD-2-related lipid-recognition	S
349607*	0.0	0.0	857	specific protein	
326387*	0.0	0.0	847	specific protein	
348750*	0.0	0.0	841	specific protein	
85283*	0.0	0.2	771	high mobility group box	
6745	0.0	0.0	750	esterase	S
3358	0.0	0.0	742	specific protein	S
24442	0.0	0.0	682	specific protein	S
182238	0.0	3.6	663	hypothetical protein	S
21350*	0.0	0.0	553	cytochrome P450	
65037*	0.0	0.1	526	major facilitator superfamily	
84430*	0.0	0.4	507	specific protein	
307589*	0.0	0.0	488	specific protein	
81390*	0.0	0.1	473	GPR1/FUN34/yaaH	

Upregulation in *R. irregularis-M. truncatula* symbiotic roots (= *in planta*) is assessed by comparing RNA-Seq transcript profiles to those from germinated spores at two and nine days (= Spores 2d and Spores 9d). Mycorrhizal roots were sampled from plantlets in contact with *R. irregularis* for two weeks. Values are the means of three biological replicates. They are expressed as FPKM, i.e. fragments per kilobase of exon per million RNA-Seq fragments mapped to *R. irregularis* genes (54). Based on the statistical analysis, a gene was considered significantly upregulated if it met two criteria: (1) False Discovery Rate (FDR)-adjusted *p*-value cutoff of 0.05 and (2) mycorrhiza versus germinated spore fold change ≥2. See Supplementary Information section 1.1 and 1.7 for details. * transcripts expressed in microdissected arbuscules obtained as describe in (62). Abbreviations: GPR1/FUN34/yaaH: a transmembrane protein related to ammonia exporter; S, proteins with a predicted signal peptide; specific protein, protein encoded by an orphan gene with no hit in the GenBank nr database.

Table S12. Most abundant InterPro (IPR) domains in symbiosis up-regulated genes in the *Rhizophagus irregularis-Medicago truncatula* symbiosis. Predicted proteins are ranked by the abundance of IPR categories

IPR	Description	#
No IPR	No IPR (*Rhizophagus*-specific proteins)	430
IPR001128	Cytochrome P450	36
IPR000719	Protein kinase, core	30
IPR013069	BTB/POZ	16
IPR011701	Major facilitator superfamily MFS-1	12
IPR006597	Sel1-like	11
IPR001199	Cytochrome b5	10
IPR013216	Methyltransferase type 11	10
IPR009071	High mobility group box, HMG	8
IPR001841	Zinc finger, RING-type	8
IPR001395	Aldo/keto reductase	7
IPR002213	UDP-glucuronosyl/UDP-glucosyltransferase	6
IPR007087	Zinc finger, C2H2-type	6
IPR002403	Cytochrome P450, E-class, group IV	6
IPR000209	Peptidase S8 and S53, subtilisin	6
IPR003172	MD-2-related lipid-recognition	5
IPR006025	Peptidase M, neutral zinc metallopeptidases	5
IPR004046	Glutathione S-transferase, C-terminal	5
IPR002198	Short-chain dehydrogenase/reductase	5
IPR003439	ABC transporter-like	5
IPR000073	Alpha/beta hydrolase fold-1	5
IPR001849	Pleckstrin-like	5
IPR003100	Argonaute and Dicer protein, PAZ	5
IPR001752	Kinesin, motor region	5
IPR008030	NmrA-like	5
IPR002347	Glucose/ribitol dehydrogenase	5
IPR010257	Fatty acid desaturase	5
IPR007855	RNA-dependent RNA polymerase	5
IPR001254	Peptidase S1 and S6, chymotrypsin/Hap	5
IPR001680	WD40 repeat	4
IPR003877	SPla/RYanodine receptor SPRY	4
IPR002921	Lipase, class 3	4
IPR004088	K Homology, type 1	4
IPR001117	Multicopper oxidase, type 1	4
IPR013094	Alpha/beta hydrolase fold-3	4
IPR002401	Cytochrome P450, E-class, group I	4
IPR006603	Cystinosin/ERS1p repeat	4
IPR001087	Lipase, GDSL	4
IPR001757	ATPase, P-type	4
IPR000910	High mobility group box, HMG1/HMG2	4
IPR014778	Myb, DNA-binding	4
IPR009003	Peptidase	4

Upregulation in *R. irregularis-M. truncatula* symbiotic roots (= *in planta*) is assessed by comparing RNA-Seq transcript profiles to those from germinated spores at two and nine days (= Spores 2d and Spores 9d). Mycorrhizal roots were sampled from plantlets in contact with *R. irregularis* for two weeks. Based on the statistical analysis, a gene was considered significantly upregulated if it met two criteria: (1) FDR-adjusted *p*-value cutoff of 0.05 and (2) mycorrhiza versus germinated spore fold change ≥2; 1,068 genes (4.7% of the 22,647 expressed genes) showed an upregulated expression. See Supplementary Information section 1.1 and 1.7 for details.

Table S13. Features of proteins with a predicted signal peptide.

Protein types/Motif	**Number**	**Number induced *in planta***
Proteins with a signal peptide	376	79
Small secreted proteins (SSPs) < 150 AA	153	29
Small cysteine-rich SSPs (SCRs: <150 AA, Cyst >3%)	59	12
Secreted proteins with sequence identity in GenBank nr	154	38
Secreted proteins with IPR domains	101	27
Secreted proteins with a GPI anchor	14	4
SSP with nuclear localization signals (NLS)	17	3
SSP with effector signature motifs *		
SSP with [LI]xAR motif	1	0
SSP with [YFW]xC motif	1	0
SSP with G[IFY][ALST]R motif	2	1
SSP with R[X]LR motif	0	0

* Only proteins with no motif identified after their sequences have been permuted 100 times by using the SHUFFLESEQ software in EMBOSS (http://emboss.sourceforge.net/) have been counted

Table S14. Upregulated transcripts coding for putative secreted proteins in *Rhizophagus-Medicago* symbiosis

Protein ID	Description	Length (AA)	# Cyst	GPI anchor	NLS	SCR	FPKM Spores 2d	FPKM Spores 9d	FPKM *in planta*
344204	specific protein	120	0				0.0	0.0	246000
21945	hypothetical protein	275	10				0.0	0.0	5609
343985	specific protein	52	2			SCR	0.0	0.0	3493
7716	specific protein	71	7			SCR	0.0	2.0	2948
347085	specific protein	94	6			SCR	0.0	27.9	2887
248224	specific protein	122	4			SCR	0.0	0.0	2524
11377	specific protein	71	9			SCR	0.0	0.0	1716
6890	specific protein	159	4				0.0	0.1	1422
11536	specific protein	154	4				0.0	0.2	1373
24444	specific protein	208	8				0.0	0.0	1318
330295	Peptidase A1	374	5				0.0	0.0	1057
337599	hypothetical protein	151	3				0.0	0.0	950
322915	specific protein	163	6				0.0	0.0	887
348622	MD-2-related lipid-recognition	167	4				0.0	0.0	871
6745	Esterase	99	1				0.0	0.0	750
3358	specific protein	128	2				0.0	0.0	742
24442	specific protein	209	8				0.0	0.0	682
182238	hypothetical protein	102	6			SCR	0.0	3.6	663
30117	specific protein	134	4				0.0	1.6	435
342269	specific protein	149	2				0.0	4.1	425
349450	MD-2-related lipid-recognition	151	4				0.0	0.0	382
23049	Peptidase S1 and S6	295	5				0.0	0.0	370
29400	Peptidase S1 and S6	291	7				0.0	0.0	367
176092	specific protein	147	3				0.0	0.3	350
9388	specific protein	167	3				0.0	0.0	301
18077	Peptidase	315	4				0.0	0.0	298
26749	specific protein	154	3				0.0	1.7	291
347635	specific protein	60	2			SCR	0.0	0.5	251
4655	specific protein	111	5			SCR	0.0	0.0	249
12926	specific protein	205	8				0.0	0.0	205
30130	specific protein	134	4				0.0	0.0	168
322704	specific protein	221	10				0.1	0.0	1114
27343	MD-2-related lipid-recognition	170	4				0.2	0.0	2205
18384	hypothetical protein	213	5		NLS		0.0	0.0	138
6820	Multicopper oxidase	622	8				0.0	0.0	117
84850	specific protein	297	4				0.0	0.0	95
18190	Peptidase A1	378	5				0.0	0.0	84
4928	specific protein	196	9				0.0	0.1	68
34922	specific protein	214	9				0.0	0.0	63
28370	Peptidase	315	4				0.0	0.0	61
321933	Peptidase S1 and S6	298	6				0.0	0.0	31
320886	Peptidase M	383	3				0.2	0.1	470
25518	specific protein	137	4				0.5	2.0	1069
22081	Peptidase S1 and S6	271	5				0.1	0.0	136
349254	hypothetical protein	245	6				0.6	0.5	783
8744	specific protein	71	8			SCR	3.0	0.0	2229
337108	hypothetical protein	256	4				0.5	0.1	258
334310	hypothetical protein	135	3				1.6	13.3	608
31657	specific protein	119	6			SCR	0.3	0.3	102
2724	hypothetical protein	451	7				0.1	0.0	22

336365	specific protein	132	13			SCR	0.4	4.2	57
342539	hypothetical protein	172	8				0.5	0.4	67
10008	specific protein	121	2				3.7	4.2	443
345761	specific protein	150	4				18.8	261.0	2144
31003	specific protein	100	0				7.5	58.2	536
343800	hypothetical protein	228	0	GPI			63.6	47.6	4070
36449	specific protein	95	1				3.0	1.2	82
29380	MD-2-related lipid-recognition	164	4				1.8	7.0	49
19507	Metallophosphoesterase	354	3				1.9	0.9	24
349684	specific protein	179	9				3.9	6.4	42
340872	Carbonic anhydrase	240	2				13.6	3.5	150
3298	hypothetical protein	178	9	GPI			37.3	86.5	344
33189	Kelch repeat	312	1				14.2	36.4	130
214362	hypothetical protein	172	4				5.1	12.3	44
319075	specific protein	138	8			SCR	1.9	2.4	16
161262	Glyoxal oxidase	562	6				3.9	0.8	32
216408	Multicopper oxidase	615	8				9.3	8.5	73
225359	specific protein	121	0				412.7	859.4	3228
25979	specific protein	80	0				919.9	970.7	6609
342887	hypothetical protein	205	2				4.7	36.6	29
8806	Metallophosphoesterase	349	3				16.4	24.0	85
31827	Alpha/beta hydrolase	354	1		NLS		5.7	8.4	27
103897	Multicopper oxidase	139	2				14.9	11.3	70
346611	Carbonic anhydrase	245	3				61.1	19.8	284
347071	hypothetical protein	262	10	GPI			274.1	294.7	1252
339235	Polysaccharide deacetylase	487	18				140.1	160.2	609
342046	hypothetical protein	479	21				14.8	7.4	50
342775	Heat shock protein	522	7		NLS		72.8	62.7	214

Upregulation in *R. irregularis-M. truncatula* symbiotic roots (= *in planta*) is assessed by comparing RNA-Seq transcript profiles to those from germinated spores at two and nine days (= Spores 2d and Spores 9d). Mycorrhizal roots were sampled from plantlets in contact with *R. irregularis* for two weeks. Based on the statistical analysis, a gene was considered significantly upregulated if it met two criteria: (1) FDR-adjusted *p*-value cutoff of 0.05 and (2) mycorrhiza versus germinated spore fold change ≥2; 1,068 genes (4.7% of the 22,647 expressed genes) showed an upregulated expression. See Supplementary Information section 1.1 and 1.7 for details. Abbreviations: AA, amino acids; Cyst, number of cysteine residues; FPKM, fragments per kilobase of exon per million RNA-Seq fragments mapped; GPI, GPI anchor; NLS, nuclear targeting signal; SCR, small cystein-rich proteins; specific protein, lineage-specific proteins.

Table S15. PFAM domains identified in proteins with a predicted signal peptide

PFAM domain	#
No Pfam domain	274
PF01344 Kelch motif	21
PF02221 ML domain	15
PF00082 Subtilase family	9
PF07732 Multicopper oxidase	5
PF00026 Eukaryotic aspartyl protease	5
PF01522 Polysaccharide deacetylase	5
PF07250 Glyoxal oxidase N-terminus	4
PF02298 Plastocyanin-like domain	3
PF00149 Calcineurin-like phosphoesterase	3
PF00188 Cysteine-rich secretory protein family	2
PF10342 Developmentally Regulated MAPK Interacting Protein	2
PF00194 Eukaryotic-type carbonic anhydrase	2
PF01565 FAD binding domain	2
PF00098 Zinc knuckle	1
PF05938 Plant self-incompatibility protein S1	1
PF01593 Flavin containing amine oxidoreductase	1
PF04185 Phosphoesterase family	1
PF02815 MIR domain	1
PF00085: Thioredoxin	1
PF02265 S1/P1 Nuclease	1
PF02128 Fungalysin metallopeptidase	1
PF00201 UDP-glucoronosyl and UDP-glucosyl transferase	1
PF00450 Serine carboxypeptidase S10	1
PF03572 Peptidase family S41	1
PF01391 Collagen triple helix repeat	1
PF03200 Mannosyl oligosaccharide glucosidase	1
PF00187 Chitin recognition protein	1
PF00023 Ankyrin repeat	1
PF04389 Peptidase family M28	1
PF03254 Xyloglucan fucosyltransferase	1
PF06814 Lung seven transmembrane receptor	1
PF05577 Serine carboxypeptidase S28	1
PF07719 Tetratricopeptide repeat	1
PF10137 Predicted nucleotide-binding protein containing TIR-like domain	1
PF00561 alpha/beta hydrolase fold	1

Table S16. Upregulated transcripts coding for CAZymes in *Rhizophagus-Medicago* symbiosis

Protein_ID	CAZy module	FPKM Spores 2d	FPKM Spores 9d	FPKM *in planta*	Description
348911	CBM50	0.3	27.6	3933.1	Peptidoglycan-binding LysM
75100	GT1	2.1	0.3	286.5	UDP-glucuronosyl/UDP-glucosyltransferase
70859	CBM48-GH13	102.4	155.2	223.3	α-amylase
340699	GT3	79.4	220.5	215.6	Glycogen synthase
344899	GT3	51.2	68.7	210.6	Glycogen synthase
45161	CBM18-CE4	0.0	0.0	144.0	Chitin-binding motif-chitin deacetylase
334983	GT15	54.6	42.7	135.5	Glycolipid 2-α-mannosyltransferase
121158	GT2	20.2	13.9	135.3	Chitin synthase
66145	GT1	0.8	0.5	112.1	UDP-glucuronosyl/UDP-glucosyltransferase
253398	GT39	47.9	39.7	109.7	Dol-P-Man: protein α−mannosyltransferase
337139	GT1	9.4	4.6	101.8	UDP-glucuronosyl/UDP-glucosyltransferase
167899	GT2	5.3	3.1	94.8	Chitin synthase
348856	GH35	18.0	20.0	88.4	Exo-β-glucosaminidase
326114	GH24	0.9	2.3	77.1	Lysozyme

Upregulation in *R. irregularis-M. truncatula* symbiotic roots (= *in planta*) was assessed by comparing RNA-Seq transcript profiles to those from germinated spores at two and nine days (= Spores 2d and Spores 9d). Mycorrhizal roots were sampled from plantlets in contact with *R. irregularis* for two weeks. Based on the statistical analysis, a gene was considered significantly upregulated if it met two criteria: (1) FDR-adjusted *p*-value cutoff of 0.05 and (2) mycorrhiza versus germinated spore fold change ≥2. FPKM, fragments per kilobase of exon per million RNA-Seq fragments mapped. Only CAZyme genes with an expression level *in planta* >70 FPKM were listed.

Table S17. Predicted MATA-HMG domain-containing proteins found in *R. irregularis* DAOM-197198 genome. The sequence accession number of the best reciprocal BLAST hit (BRH), the BLAST e-value, the predicted protein domain name, and the protein description of the first blast hit retrieved from the NCBI nr database are listed for each *R. irregularis* MAT-HMG gene. Green cells indicate the presence of these genes in isolates A4, B3, and C2 of *R. irregularis*, while red cells indicate their absence. See Supplementary Information section 1.1 and 1.6 for details and Fig. S19.

Genomic location	A4	B3	C2	Accession #	BLAST e-value	Protein domain	Protein description
scaffold 6161				BAE94382.1	3,00E-03	MATA_HMG	MAT1-2-1
scaffold 11528				EFY86728.1	8,00E-05	MATA_HMG	HMG transcription factor
scaffold 24332 1				EFX05114.1	2,00E-05	MATA_HMG	HMG box protein
scaffold 5899 2				XP_003007798.1	4,00E-06	MATA_HMG	predicted protein
scaffold 5899 1				EGF99649.1	9,00E-07	MATA_HMG	hypothetical protein
scaffold 25287				ABC68485.1	3,00E-03	MATA_HMG	MAT1-2-1
scaffold 10252				XP_002564591.1	4,00E-05	MATA_HMG	Hypothetical
scaffold 23703				CCE33026.1	3,00E-10	MATA_HMG	mating type gene
scaffold 24420				AFA26123.1	4,00E-10	MATA_HMG	SexM
scaffold 15128				EIW55118.1	5,00E-04	MATA_HMG	HMG box protein
scaffold 8926				AAT48651.1	1,00E-04	MATA_HMG	MAT1-2
scaffold 9271 2				CAD62166.1	3,00E-04	MATA_HMG	HMG box protein
scaffold 22967				AFM85245.1	1,00E-17	MATA_HMG	MatMc
scaffold 13782 2				CCM00606.1	2,00E-03	MATA_HMG	predicted protein
scaffold 11572				EHK50111.1	6,00E-03	MATA_HMG	hypothetical protein
scaffold 14496				ABB83710.1	3,00E-10	MATA_HMG	MAT1-2
scaffold 6509				CAD62166.1	2,00E-10	MATA_HMG	HMG box protein
scaffold 13073				CCM01306.1	2,00E-02	MATA_HMG	predicted protein
scaffold 18730				CAD62166.1	3,00E-04	MATA_HMG	HMG box protein
scaffold 11099				EIW63176.1	2,00E-09	MATA_HMG	Hypothetical
scaffold 24332 2				CCA67490.1	5,00E-11	MATA_HMG	Hypothetical
scaffold 8838				AEI83491.1	3,00E-05	MATA_HMG	MAT1-2
scaffold 25670				CCF52951.1	2,00E-11	MATA_HMG	Prf1
scaffold 13782 1				XP_002152469.1	1,00E-04	MATA_HMG	MAT1-2-1
scaffold 10425				EIW55066.1	1,80E+00	MATA_HMG	hypothetical protein
scaffold 5393 2				NP_595867.1	5,00E-05	MATA_HMG	mc 2
scaffold 9387				CCM01306.1	8,00E-05	MATA_HMG	Hypothetical

scaffold 23044		GAA93066.1	2,40E+00	MATA_HMG	Hypothetical
scaffold 1489		XP_003855030.1	4,00E-04	MATA_HMG	Hypothetical
scaffold 26230		AAK83344.1	1,00E-07	MATA_HMG	MAT1-1-3
scaffold 24639		XP_002488738.1	2,00E-06	MATA_HMG	MAT1-2-1
scaffold 18459 1		XP_003029891.1	3,00E-05	MATA_HMG	Hypothetical
scaffold 26755		CCA72393.1	1,00E-04	MATA_HMG	Hypothetical
scaffold 20035		AET35404.1	9,30E-02	MATA_HMG	SexP
scaffold 23687		CAI59768.2	2,00E-03	MATA_HMG	MAT1-2-1
scaffold 2943		CAD62166.1	2,00E-09	MATA_HMG	HMG box protein
scaffold 26602		ABC68485.1	3,00E-07	MATA_HMG	MAT1-2-1
scaffold 13111		EFY86728.1	3,00E-07	MATA_HMG	HMG transcription factor
scaffold 6122		BAE93596.1	5,00E-03	MATA_HMG	MAT1-1-3
scaffold 8872		AFP89369.1	5,80E-02	MATA_HMG	MAT1-2-1
scaffold 7299		XP_003297060.1	1,00E-16	MATA_HMG	Hypothetical
scaffold 23005		EGD75508.1	3,40E-02	MATA_HMG	Hypothetical
scaffold 4041		CCA72393.1	1,00E-04	MATA_HMG	Hypothetical
scaffold 24024		AAB28876.1	7,50E-01	MATA_HMG	matMc
scaffold 28019		CCF38267.1	1,60E-02	MATA_HMG	HMG box protein
scaffold 21404		ABX27909.1	9,00E-06	MATA_HMG	SexM
scaffold 22418		AEA29200.1	9,60E-02	MATA_HMG	MAT1-1-2
scaffold 8935		EJT79613.1	2,00E-03	MATA_HMG	Hypothetical
scaffold 26574		EGY15843.1	8,00E-06	MATA_HMG	Hypothetical
scaffold 22770		AFQ90566.1	1,00E-03	MATA_HMG	MAT1-2-1
scaffold 18308		EMC98166.1	2,00E-05	MATA_HMG	MAT1-2-1
scaffold 10137		ACJ70020.1	5,00E-06	MATA_HMG-box	mating type protein 1-2-1
scaffold 10207		ACV60399.1	8,00E-04	MATA_HMG-box	MAT1-2-1
scaffold 10219		EIW55066.1	2,00E-06	HMG-box	hypothetical protein
scaffold 1059		AEB33764.1	3,00E-05	MATA_HMG-box	HMG domain mating-type protein MAT1-2-1
scaffold 10953		CCF38267.1	6,00E-08	MATA_HMG-box	HMG box protein
scaffold 11056		CAI59768.2	3,00E-10	MATA_HMG-box	mating-type protein MAT1-2-1
scaffold 11982 1		EIW55066.1	3,00E-07	HMG-box	hypothetical protein
scaffold 11982 2		EAU36401.1	4,00E-06	MATA_HMG-box	predicted protein
scaffold 12031		CAB63345.1	1,00E-06	MATA_HMG-box	mating-type protein Mat a-1

scaffold 12373	AAT48651.1	6,00E-06	MATA_HMG-box	mating-type protein MAT1-2
scaffold 12516	AAB28876.1	9,00E-06	MATA_HMG-box	mat-Mc product
scaffold 13738	EFQ94844.1	6,00E-11	MATA_HMG-box	hypothetical protein
scaffold 13912	ADJ38503.1	4,00E-05	MATA_HMG-box	MAT1-2-1
scaffold 14082	NP_595867.1	9,00E-14	MATA_HMG-box	m-specific polypeptide mc 2
scaffold 14378	ADD92633	1,00E-04	HMG-box	MAT1-2-1
scaffold 1469	EIM88371	3,00E-06	HMG-box	hypothetical protein
scaffold 15072	EHK50111.1	2,00E-05	HMG-box	hypothetical protein
scaffold 15202	ABC68485.1	3,00E-04	MATA_HMG-box	MAT1-2-1
scaffold 15346 1	EGP92003.1	2,00E-14	MATA_HMG-box	high-mobility group protein
scaffold 15346 2	EEY15877.1	5,00E-07	HMG-box	predicted protein
scaffold 16096	BAE93753	4,00E-05	MATA_HMG-box	mating type gene
scaffold 16420	EIW55066.1	3,00E-07	HMG-box	hypothetical protein
scaffold 16771	EGF76365	2,00E-07	HMG-box	hypothetical protein
scaffold 17668	EIW61440.1	7,00E-13	HMG-box	hypothetical protein
scaffold 17872	EHA21813	9,00E-08	MATA_HMG-box	hypothetical protein
scaffold 18118	CAD59614.3	7,00E-15	MATA_HMG-box	putative mating type protein MAT-2
scaffold 18459 2	EFI94988.1	3,00E-08	HMG-box	hypothetical protein
scaffold 18713	ADB28879.1	1,00E-07	MATA_HMG-box	MAT1-2-1
scaffold 18989	BAE93759.1	1,00E-06	MATA_HMG-box	mating type gene
scaffold 19131	ADJ38503	2,00E-09	HMG-box	MAT1-2-1
scaffold 19402	EEA20220.1	1,00E-25	MATA_HMG-box	HMG box transcriptional regulator
scaffold 19862	EED11557.1	2,00E-08	MATA_HMG-box	mating type protein MAT1-2-1
scaffold 20416	CAI59768.2	2,00E-05	MATA_HMG-box	mating-type protein MAT1-2-1
scaffold 20689	AAK83344.1	3,00E-12	MATA_HMG-box	mating type protein MAT1-1-3
scaffold 20827	CAI45464.3	1,00E-13	HMG-box	mating type protein
scaffold 20832	BAA28611	7,00E-22	HMG-box	MAT-2 protein
scaffold 21001	EFX04658.1	1,00E-05	MATA_HMG-box	hmg box protein
scaffold 21210	EFY88585.1|	3,00E-08	MATA_HMG-box	MAT1-2-1 like protein
scaffold 2189	BAA28611.1	5,00E-09	HMG-box	MAT-2 protein
scaffold 21949	CAB63346.1	3,00E-05	MATA_HMG-box	mating-type protein Mat a-1
scaffold 22340	XP_661271.1	4,00E-13	MATA_HMG-box	hypothetical protein
scaffold 22341	EAW24898.1	8,00E-04	MATA_HMG-box	HMG box protein
scaffold 22419	AAL28013.1	3,00E-05	MATA_HMG-box	mating-type protein Mat a-1
scaffold 22834	ABX27909.1	1,00E-05	HMG-box	SexM
scaffold 22855	A49103	9,00E-08	MATA_HMG-box	mat-Mc

scaffold 22982		AAQ82722.1	1,00E-03	MATA_HMG-box	matMc
scaffold 22994		XP_002152469.1	1,00E-16	MATA_HMG-box	mating type protein MAT1-2-1
scaffold 230		BAE93753.1	3,00E-06	MATA_HMG-box	mating type gene
scaffold 23137		EGU84400.1	4,00E-07	MATA_HMG-box	hypothetical protein
scaffold 23489		BAA28611.1	1,00E-19	HMG-box	MAT-2 protein
scaffold 23584		A49103	1,00E-02	MATA_HMG-box	unnamed protein product
scaffold 23866		EEA19532.1	1,00E-08	MATA_HMG-box	mating type protein MAT1-2-1
scaffold 23936		EIE88156.1	1,00E-12	MATA_HMG-box	hypothetical protein
scaffold 23954		EFY86728	2,00E-09	MATA_HMG-box	HMG box transcription factor
scaffold 23995		CCF38267.1	3,00E-07	MATA_HMG-box	HMG box protein
scaffold 24083		CAA30481.1	2,00E-15	MATA_HMG-box	unnamed protein product
scaffold 24153		EFQ29499.1	9,00E-15	MATA_HMG-box	HMG box protein
scaffold 24446		AEB33764.1	7,00E-07	MATA_HMG-box	HMG domain mating-type protein MAT1-2-1
scaffold 24771		EGF84214.1	4,00E-12	HMG-box	hypothetical protein
scaffold 24949		CAI59768.2	2,00E-05	MATA_HMG-box	mating-type protein MAT1-2-1
scaffold 25118		CAD59612.3	2,00E-10	MATA_HMG-box	putative mating type protein MAT-2
scaffold 25720		AEB33764.1	2,00E-08	MATA_HMG-box	HMG domain mating-type protein MAT1-2-1
scaffold 26025		AEA29202.1	2,00E-06	HMG-box	mating type protein 1-1-2
scaffold 26382		ABN11480.1	7,00E-15	HMG-box	STE11-like transcription factor
scaffold 26760		A49103	5,00E-10	MATA_HMG-box	mat-Mc product
scaffold 27041		EIW65046.1	8,00E-09	HMG-box	HMG-box
scaffold 28122		AAK83344.1	1,00E-12	MATA_HMG-box	mating type protein MAT1-1-3
scaffold 2948		CAB63226.1	2,00E-07	MATA_HMG-box	mating-type protein Mat a-1
scaffold 3337		ACR78246.1	7,00E-08	MATA_HMG-box	MAT1-1-3
scaffold 3371		BAD01149.1	1,00E-04	MATA_HMG-box	MAT1-2-1
scaffold 4373		NP_595867.1	8,00E-18	MATA_HMG-box	m-specific polypeptide mc 2
scaffold 4671		ACV60399.1	8,00E-06	MATA_HMG-box	MAT1-2-1
scaffold 4700		XP_001645206.1	8,00E-10	HMG-box	hypothetical protein
scaffold 4807		A49103	1,00E-04	MATA_HMG-box	mat-Mc
scaffold 5230		EGP90006.1	4,00E-06	HMG-box	hypothetical protein
scaffold 5393 1		XP_001230155.1	4,00E-10	HMG-box	hypothetical protein
scaffold 5653		EAU35789.1	8,00E-14	MATA_HMG-box	predicted protein
scaffold 6025		EED11557.1	1,00E-05	MATA_HMG-box	mating type protein MAT1-2-1
scaffold 6032		AEA29202.1	3,00E-08	HMG-box	mating type protein 1-1-2

scaffold 6084		GAA95608.1	1,00E-04	HMG-box	hypothetical protein
scaffold 6363		AEB33764.1	4,00E-06	MATA_HMG-box	HMG domain mating-type protein MAT1-2-1
scaffold 6674		AFA26123.1	4,00E-07	MATA_HMG-box	SexM
scaffold 6688		EAL92951.2	1,00E-06	MATA_HMG-box	mating type protein MAT1-2-1
scaffold 7661		GAA98854.1	2,00E-08	HMG-box	hypothetical protein
scaffold 7897		AEI72616.1	7,00E-07	MATA_HMG-box	mating type A-1-3 HMG1/2
scaffold 8367		EIW55118.1	1,00E-11	HMG-box	HMG box protein
scaffold 8757		EAQ86935.1	1,00E-15	MATA_HMG-box	hypothetical protein
scaffold 8838		EEA19532.1	7,00E-09	MATA_HMG-box	mating type protein MAT1-2-1
scaffold 95 1		BAD72610.2	5,00E-12	MATA_HMG-box	MAT1-1-3
scaffold 95 2		BAE93748.1	5,00E-12	MATA_HMG-box	mating type gene
scaffold 9511		AAF00498	1,00E-05	MATA_HMG-box	MAT2-1
scaffold 9528		EED82531.1	2,00E-05	HMG-box	predicted protein
scaffold 21103		XP_003958640.1	1,00E-06	MATA_HMG	Hypothetical
scaffold 17607		AET35422.1	2,00E-03	HMG-box	SexP
scaffold 13559		XP_003334333.1	4,00E-14	MATA_HMG	Hypothetical

Figures S1 to S20

Figure S1. Strategy used for assembling the Sanger, 454, Illumina and PacBio genomic reads from *R. irregularis* DAOM-197198, the Illumina RNA sequences from *R. irregularis* DAOM-197198, and the annotation procedures. The so-called 'high confidence' gene models are those supported by RNA-Seq expression data or having a sequence similarity with known genes in the databases. Bacterial sequences were filtered as described in Fig. S20. N50: the length N for which 50% of all bases are in a sequence of length L<N. Mb, megabase; kb, kilobase. Spores DAOM: germinated spores from DAOM-197198; Spores C2: germinated spores from *R. irregularis* C2; Spores *R. diaph.*: germinated spores from *R. diaphanum*; *In Planta* DAOM: *R. irregularis* DAOM197198-*Medicago truncatula* symbiotic roots.

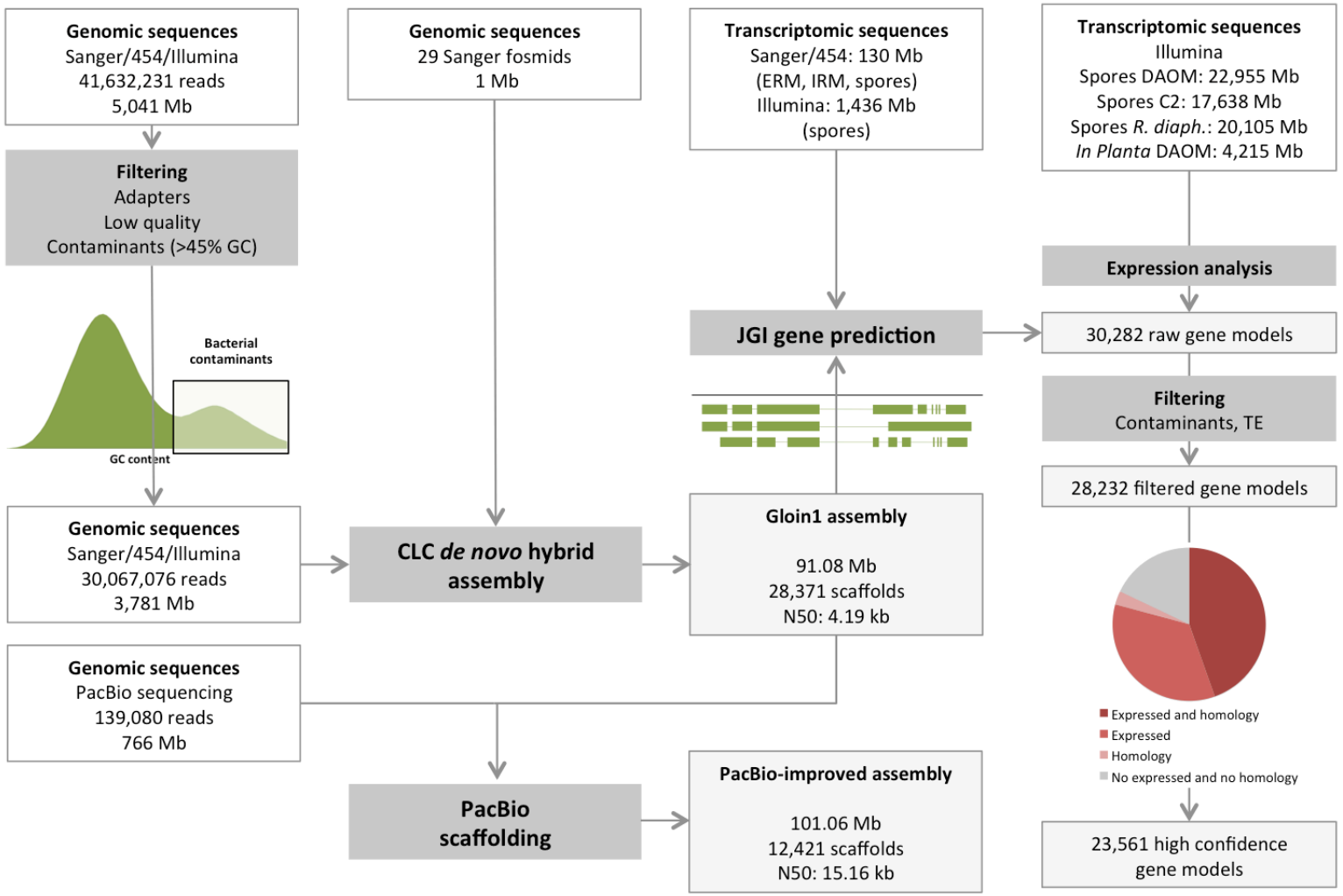


Figure S2. Sequence coverage of the *R. irregularis* DAOM19-7198 genome assemblies. Scaffold sequence length *versus* sequencing average coverage for the Gloin1 and PacBio-improved Gloin1 assemblies.

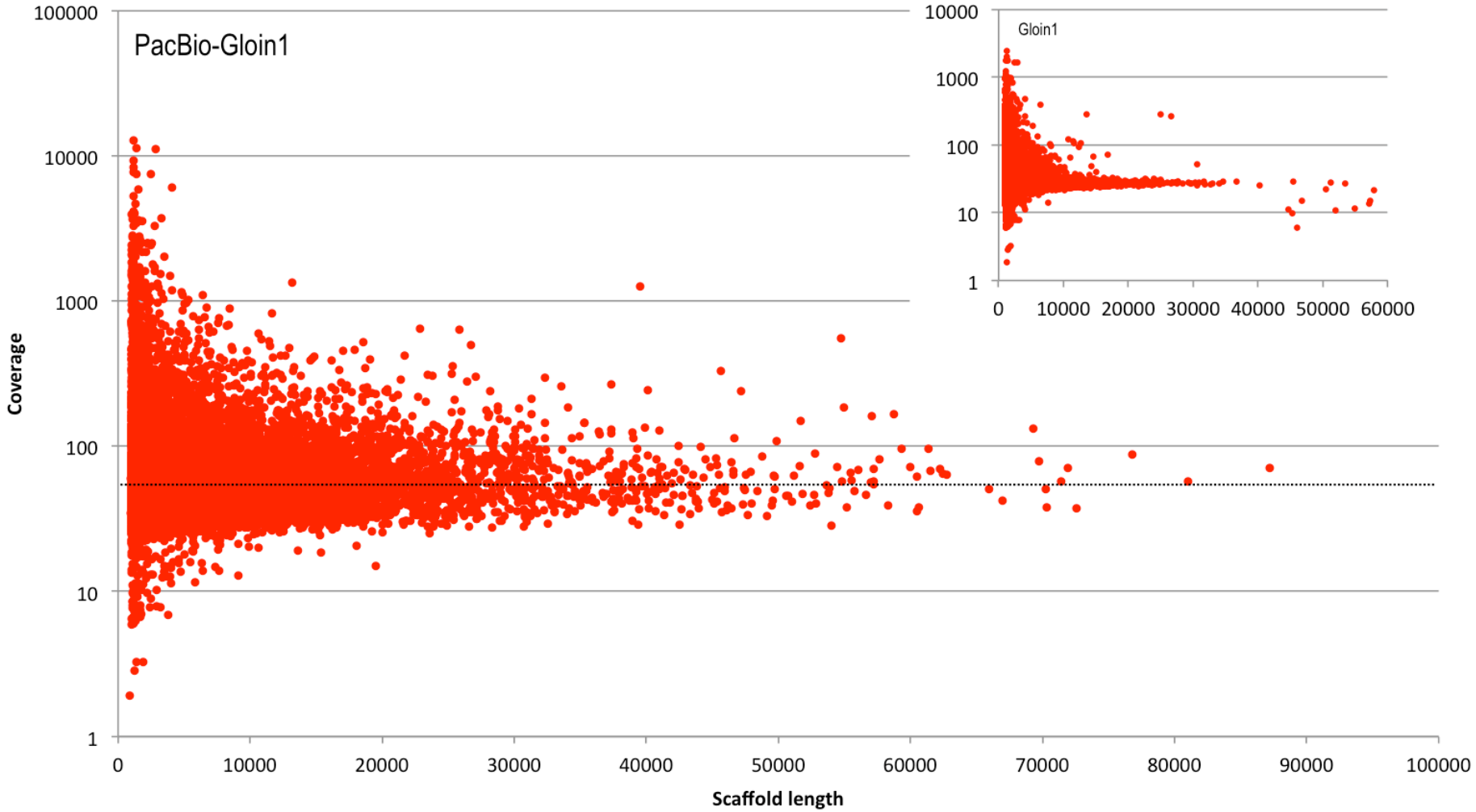


Figure S3. Macrosynteny between the three largest scaffolds of the PacBio-improved Gloin1 genome assembly (right panel) and scaffolds of the genome assembly Gloin1 (left panel). PacBio-improved Gloin1 genome assembly scaffolds are depicted by the colored blocks and Gloin1 scaffolds are represented by gray blocks. Comparison between the two assemblies was performed with MUMMER. Only regions larger than 1 kb and percent identity higher than 97% are connected with links of colors matching those used for the PacBio-improved Gloin1 genome assembly scaffolds. The location of protein-coding regions (blue), repeats regions (red) and gaps regions (grey) are showed on innermost circle. Outermost circle represents GC content based on a sliding window of 100 bp (red > 40% ; green < 20% ; midline: 33%).

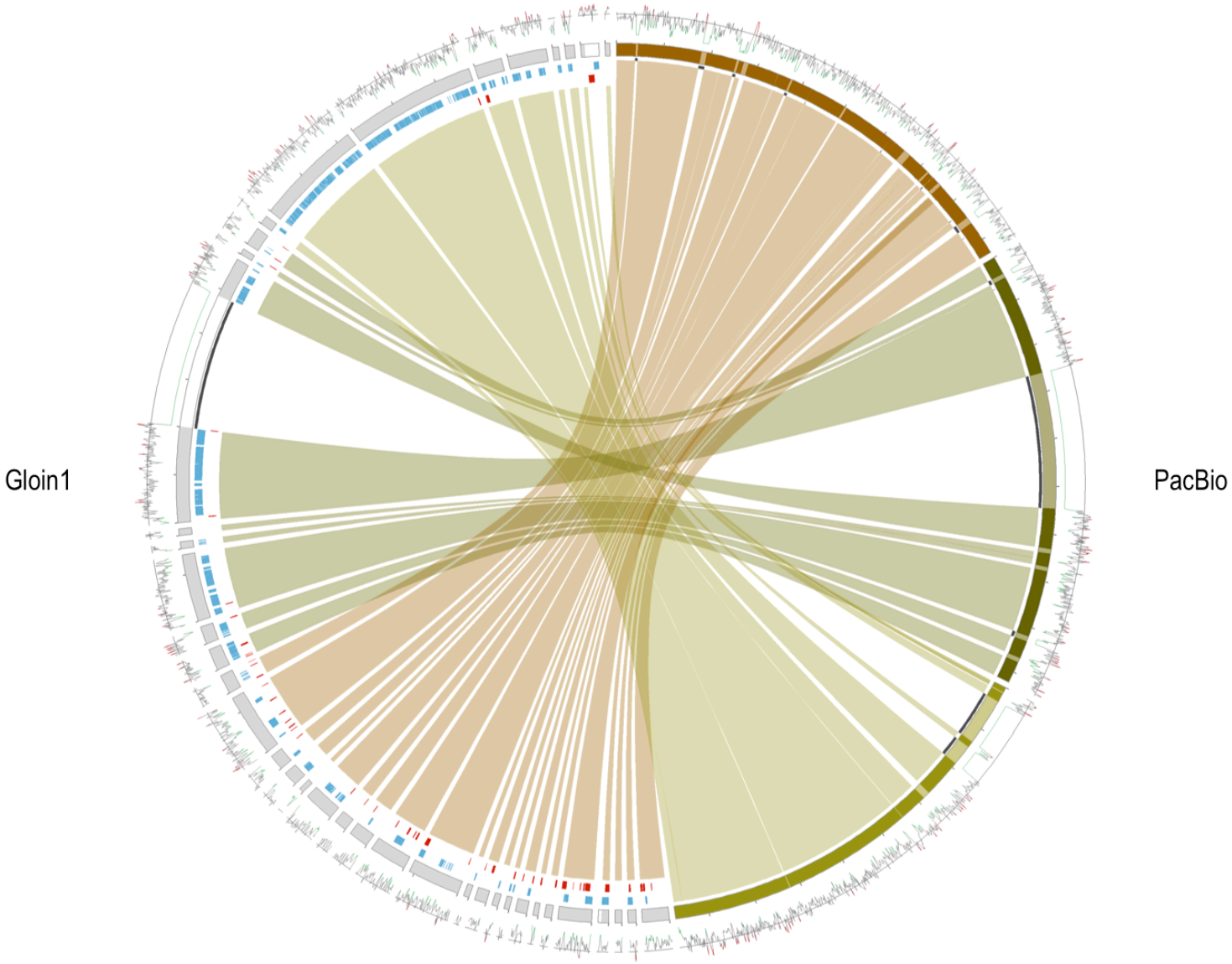


Figure S4. Locations of protein-coding genes, Polinton transposable elements and LTR retrotransposons on *R. irregularis* fosmids. **(a)** the 29 fosmids regions which aligned to Gloin1 scaffolds are highlighted in dark grey. (**b**) Locations of gene models (blue), LTRs retrotransposons (dark red), LTR retrotransposons fragments (light red) and Polinton transposable elements (green). (**c**) Genome read coverage (scale from 0 to 100). (**d**) Transcriptome read coverage, germinated spores (scale from 0 to 1000). (**e**) Transcriptome read coverage, *in planta* (scale from 0 to 200). (**f**) GC content based on a sliding window of 100 bp (red > 40% ; green < 20%; midline: 33%).

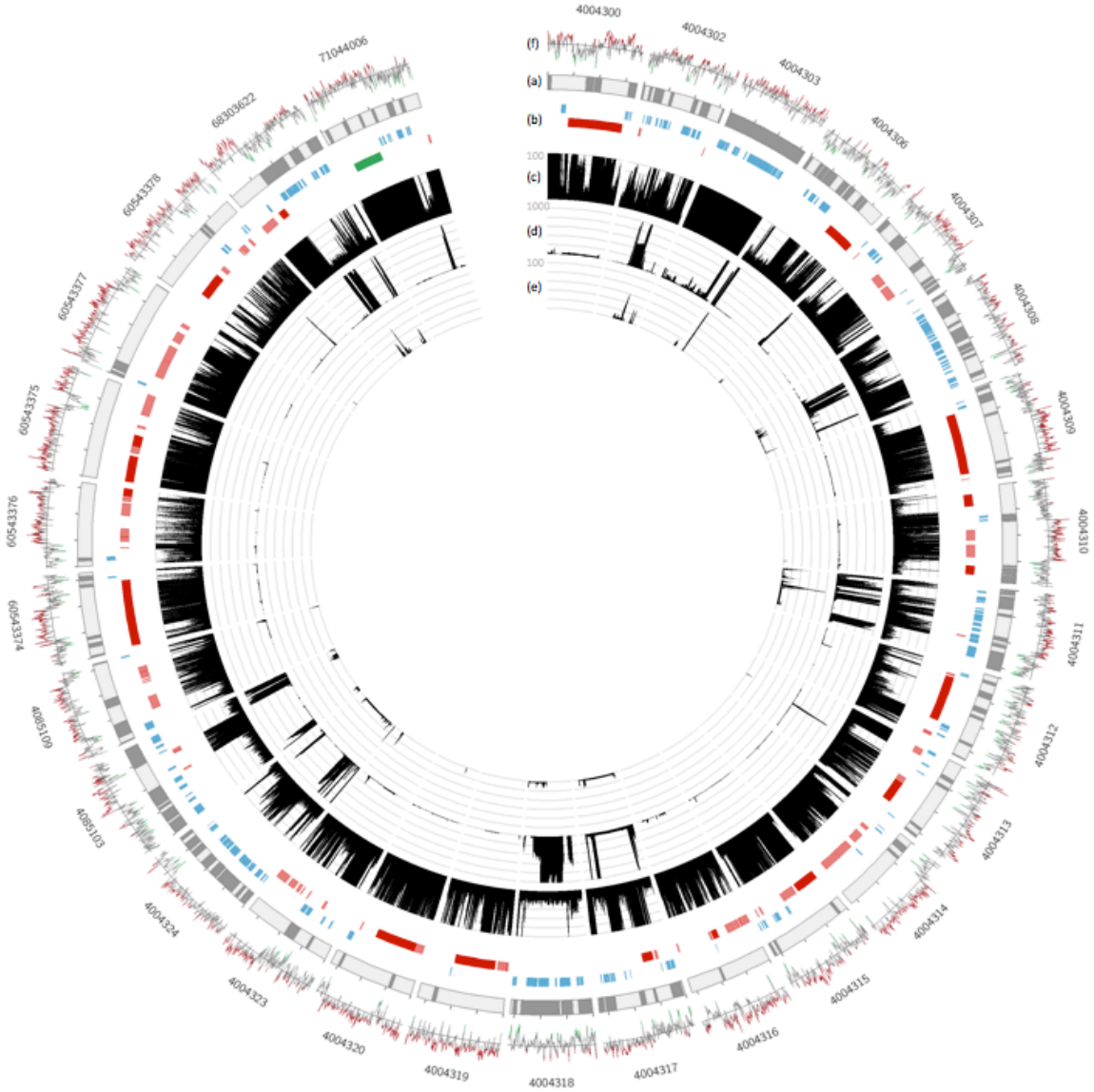


Figure S5. Features of the largest scaffolds from the *R. irregularis* genome assembly. (**A**) scaffold_1649. (**B**) scaffold_2742. (**a**) GC content based on a sliding window of 100 bp (red > 33%, green < 33%). (**b**) Location of predicted protein-coding gene models. (**c**) Location of repeated elements. (**d**) Genome read coverage (scale from 0 to 100). (**e**) Location of genomic SNPs. (**f**) Expressed gene read coverage, germinated spores (scale from 0 to 1000). (**g**) Expressed gene read coverage, *in planta* (scale from 0 to 100). (**h**) Location of SNPs in expressed DAOM-197198 genes identified using mapped RNA-Seq reads from DAOM-197198 germinated spores. (**i**) Location of SNPs in expressed genes, RNA-Seq reads from C2 germinated spores. (**j**) Location of SNPs in expressed genes, RNA-Seq reads from *R. diaphanum* germinated spores.

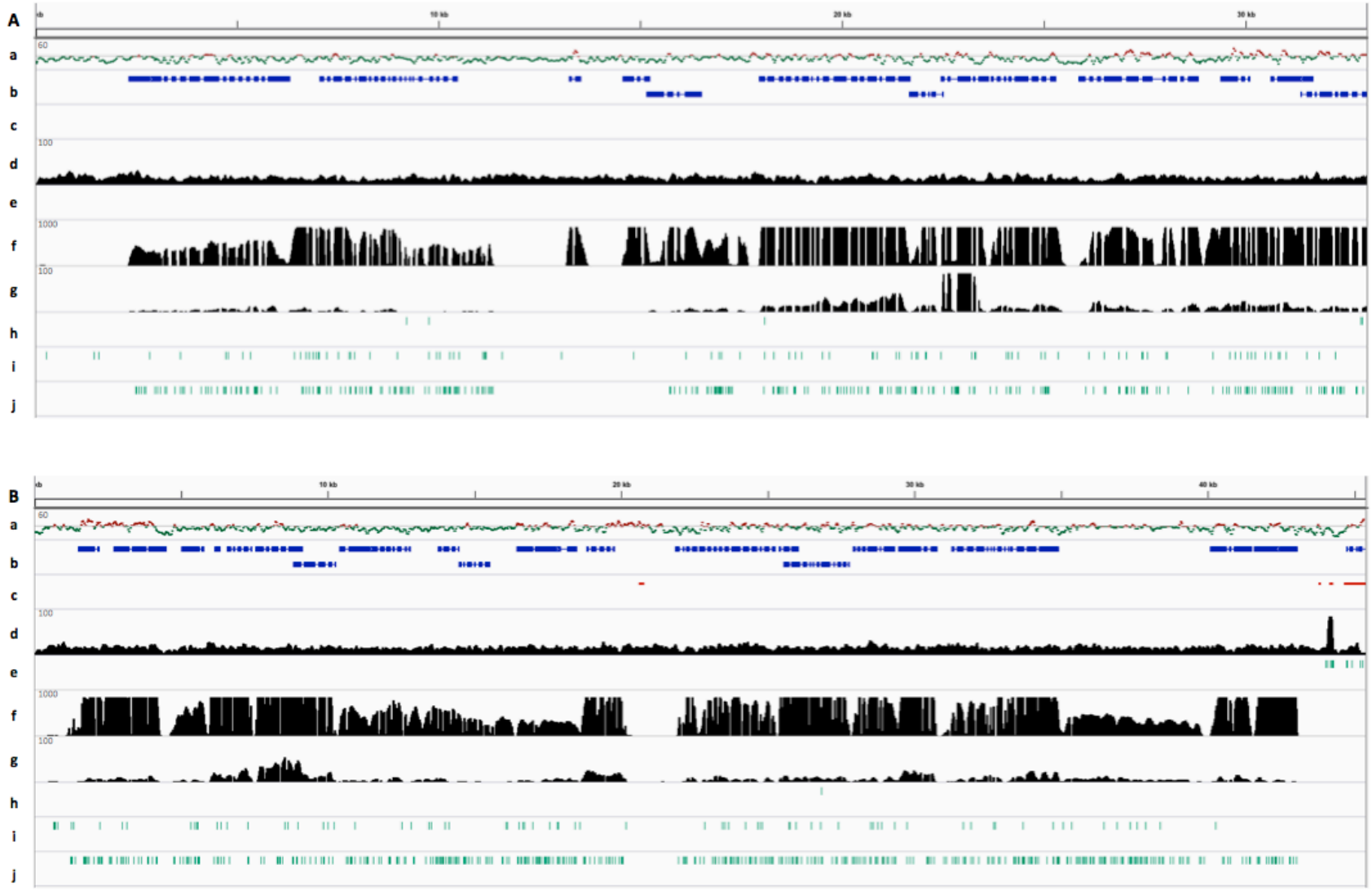


Figure S6. Size of the *R. irregularis* DAOM-197198 genome based on *k*-mer analysis. Distribution of 17-mer frequency in the raw Illumina genomic reads is showed. The peak depth is at 15X. The peak of 17-mer frequency (M) in reads is correlated with the real sequencing depth (N), read length (L), and kmer length (K), their relations can be expressed in a experienced formula: M = N * (L – K + 1) / L. Then, we divided the total sequence length by the real sequencing depth and obtained an estimated genome size of 153 Mb.

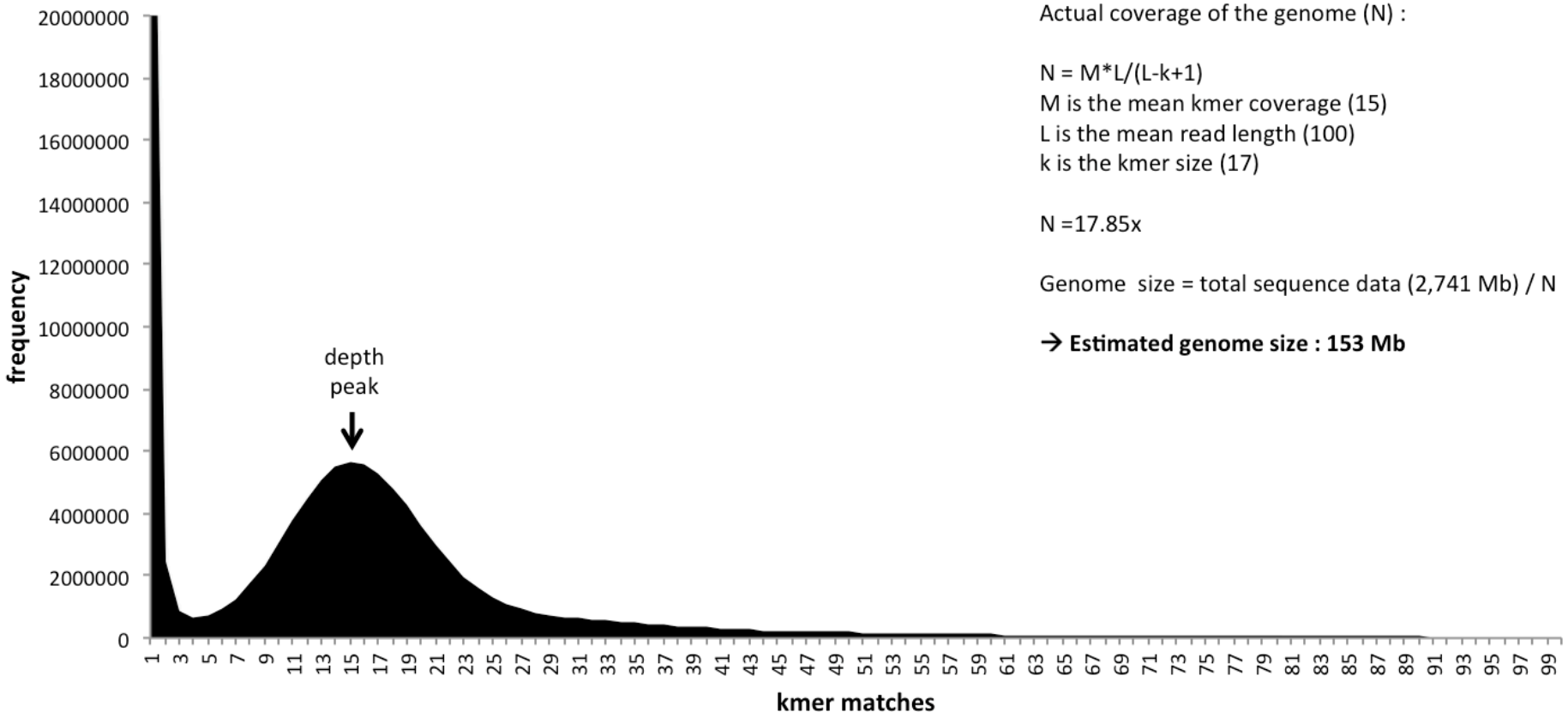


Figure S7. Venn diagram showing the number of SNPs identified in the *R. irregularis* DAOM-197198 genome by mapping Illumina RNA-Seq reads from *R. irregularis* DAOM-197198 transcriptome (8,643 SNPs), *R. irregularis* C2 transcriptome (5,144 SNPs) or *R. diaphanum* MUCL43196 (4,469 SNPs) and SNPs shared between isolates. Only 291 SNPs are shared by all isolates.

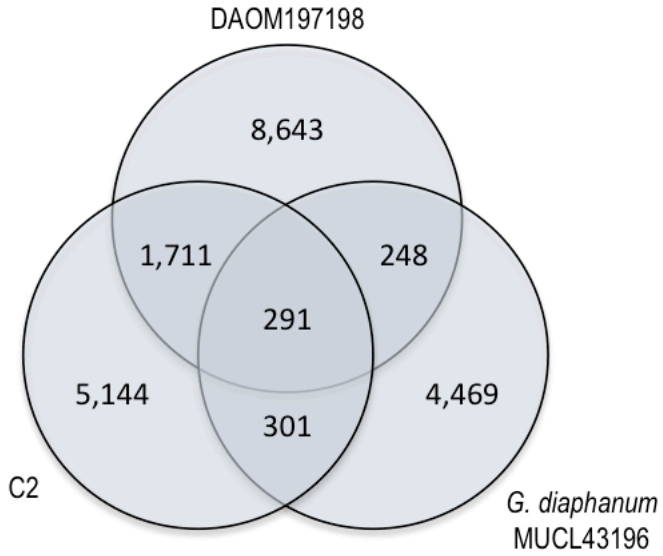


Figure S8. The distribution of duplication age (as measured by numbers of synonymous substitutions) of the *R. irregularis* paralogs. (**A**) Scaled normal probability density function (pdf) plot for each component detected by EMMIX and (**B**) histogram of the *Ks* values and the different components, where y-axis is the number of duplication nodes located in the *Ks* range of the corresponding bin.

.

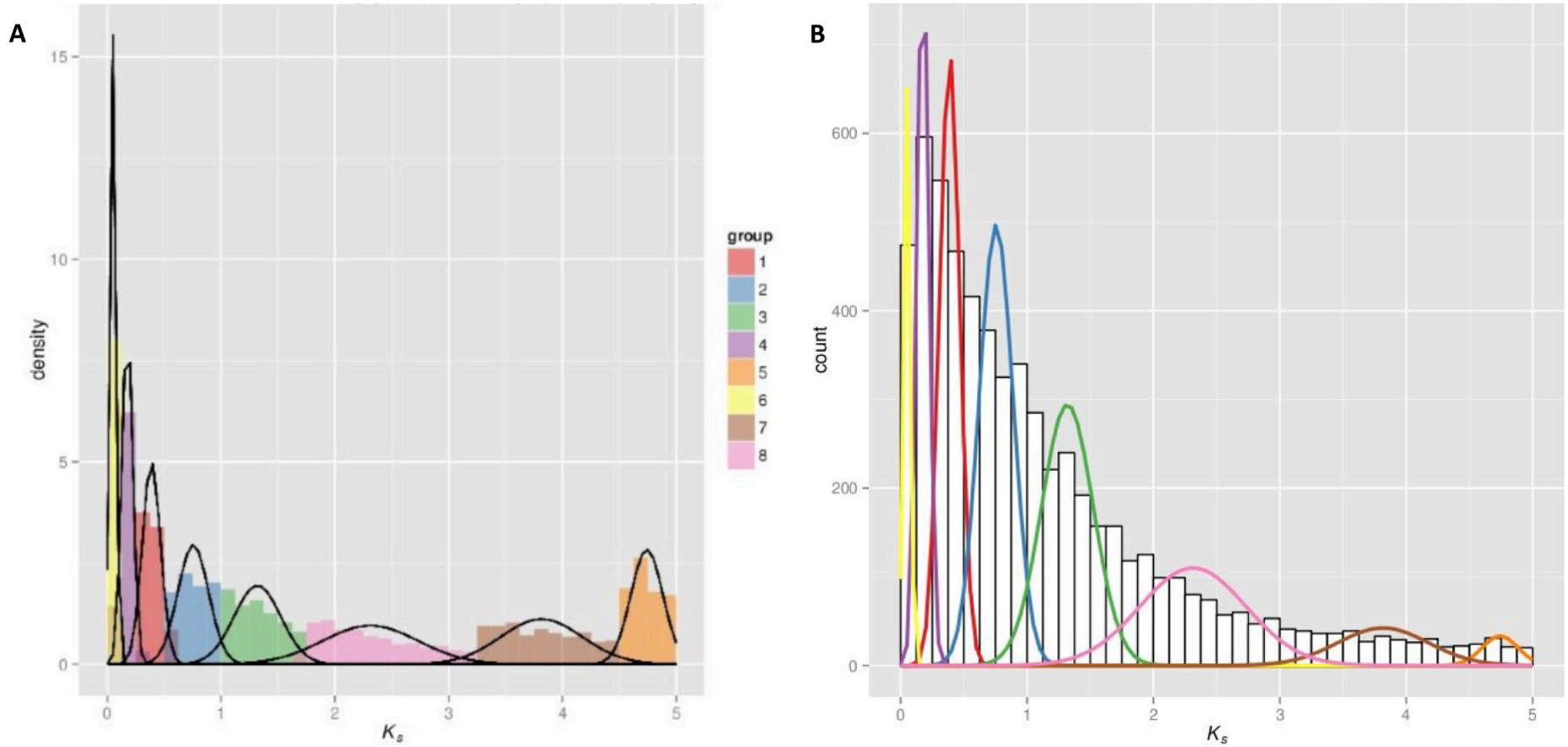
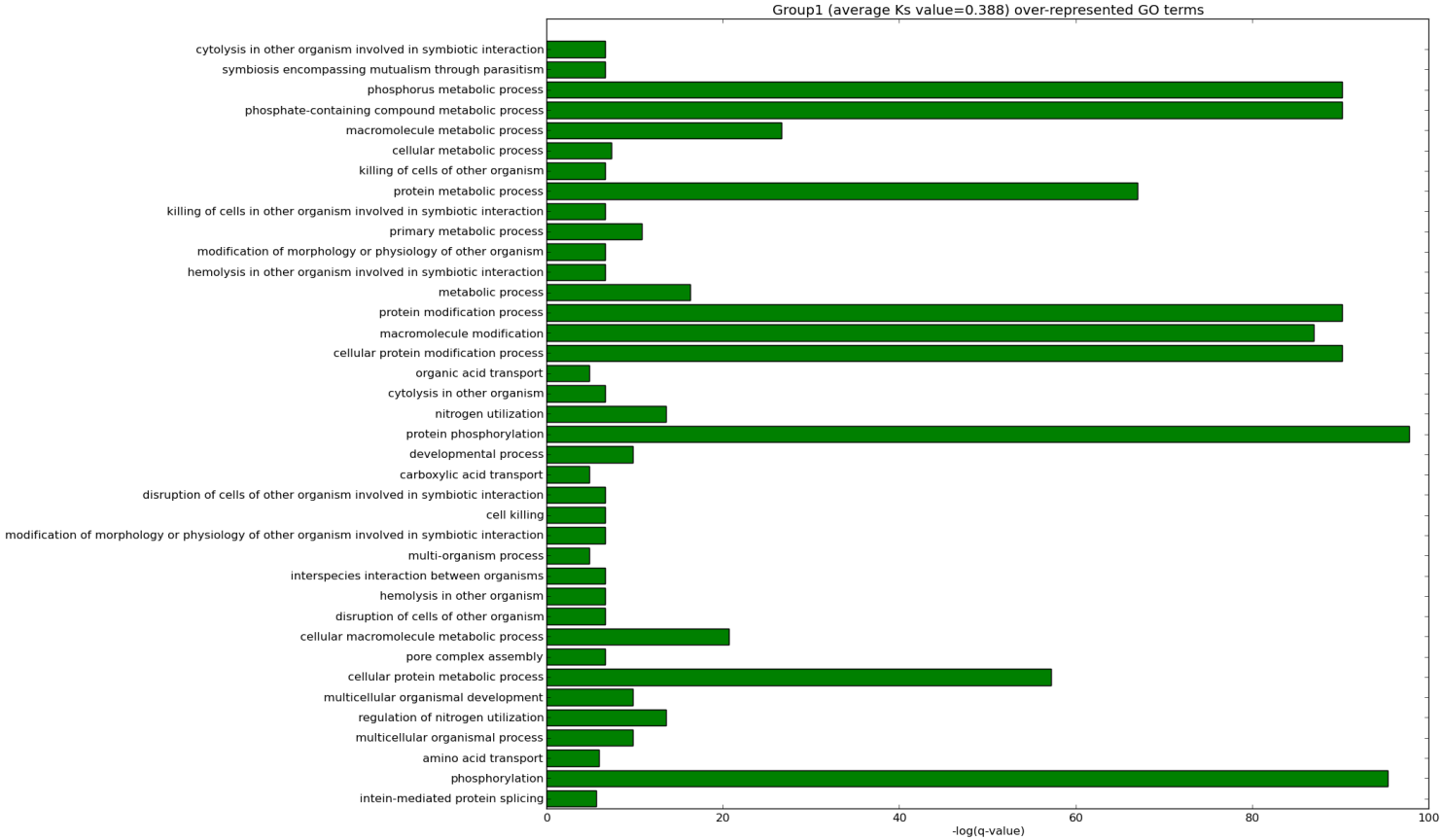
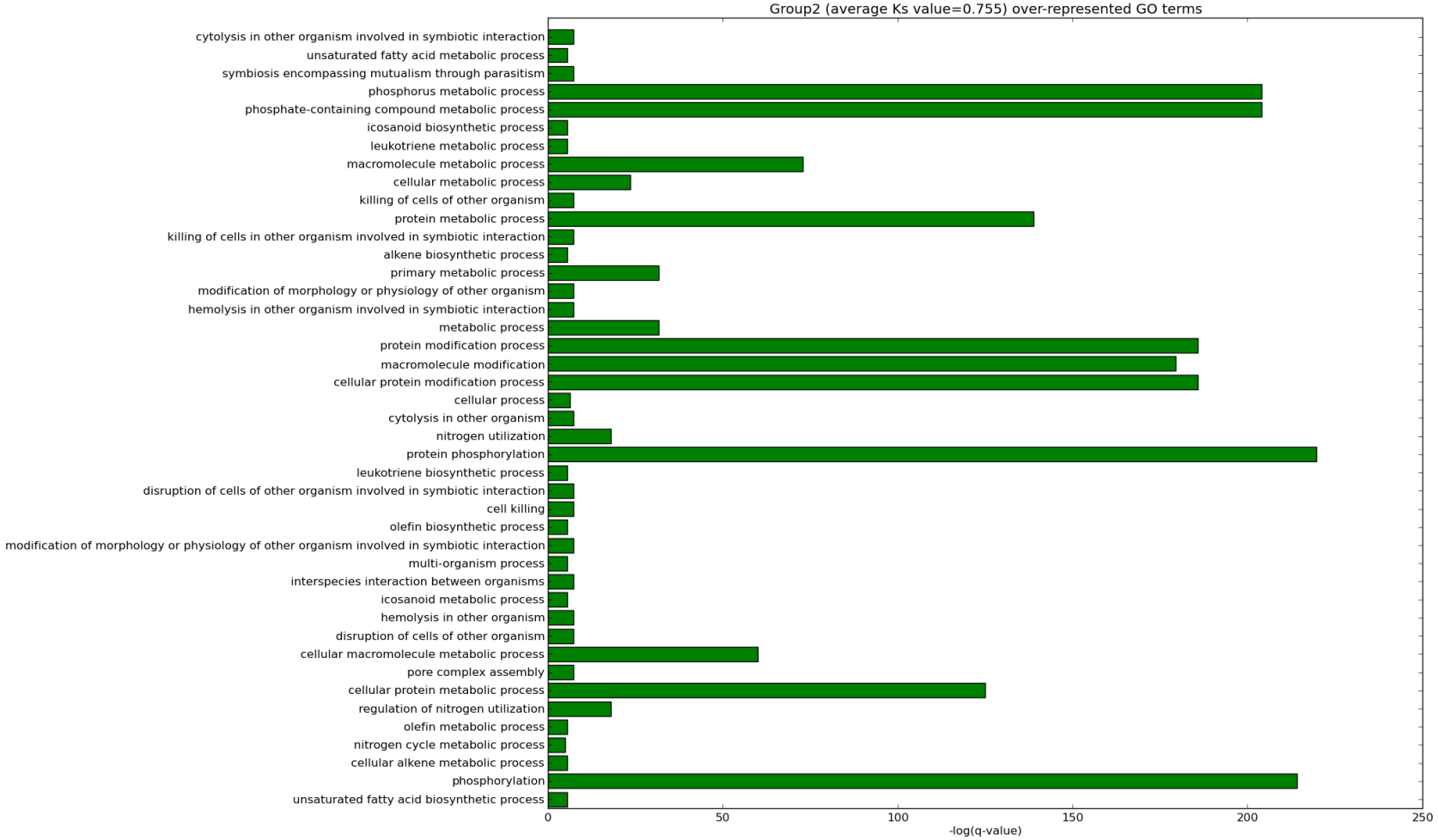


Figure S9. Duplicated paralogs in the *R. irregularis* genome are enriched for genes annotated to contribute to biological processes related to phosphorus metabolism and signaling via phosphorylation (metabolism and phosphorylation). Fisher's exact test based on the GO annotation for each of the 8 identified duplication groups are showed in Figure S8. The x-axis of the plots has the -log(q-value) of each identified term. Only over-represented terms are shown, whereas plots with under-represented terms can be downloaded from the INRA Rhizophagus web portal at: http://mycor.nancy.inra.fr/IMGC/GlomusGenome/index3.html. The number of genes participating in the different groups are 3,074 genes for group 1 (**A**), 5,015 for group 2 (**B**), 6,543 for group 3 (**C**), 1,764 for group 4 (**D**), 611 for group 5 (**E**), 700 for group 6 (**F**), 2,338 for group 7 (**G**) and 7,024 genes for group 8 (**H**). A series of tables listing the distribution of protein IDs for each GO categories, including all phosphorus-related terms and the kinases only (GO:0004672, protein kinase activity) for each group of paralogs, can be downloaded from the INRA *Rhizophagus* web portal at: http://mycor.nancy.inra.fr/IMGC/GlomusGenome/index3.html.

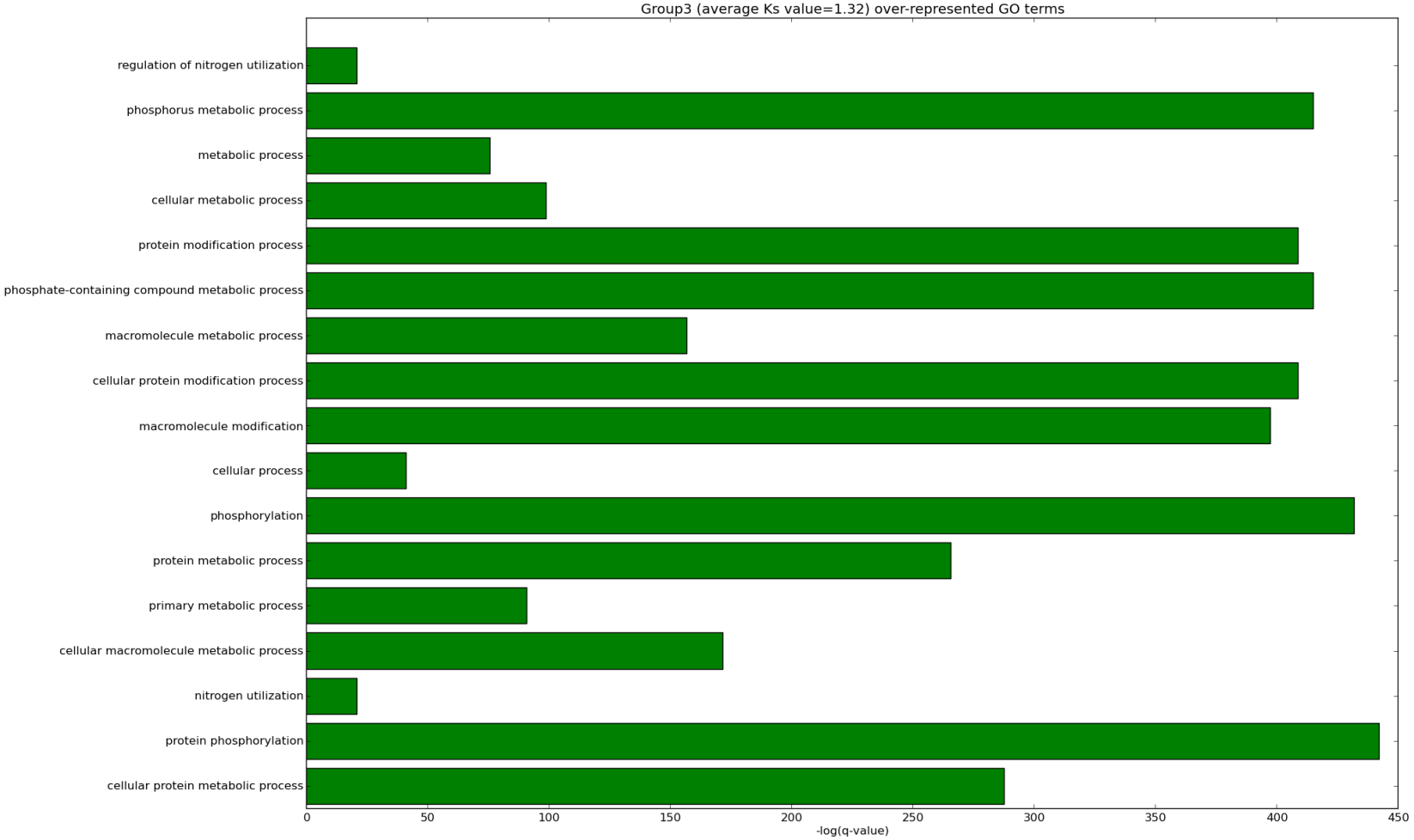
A.



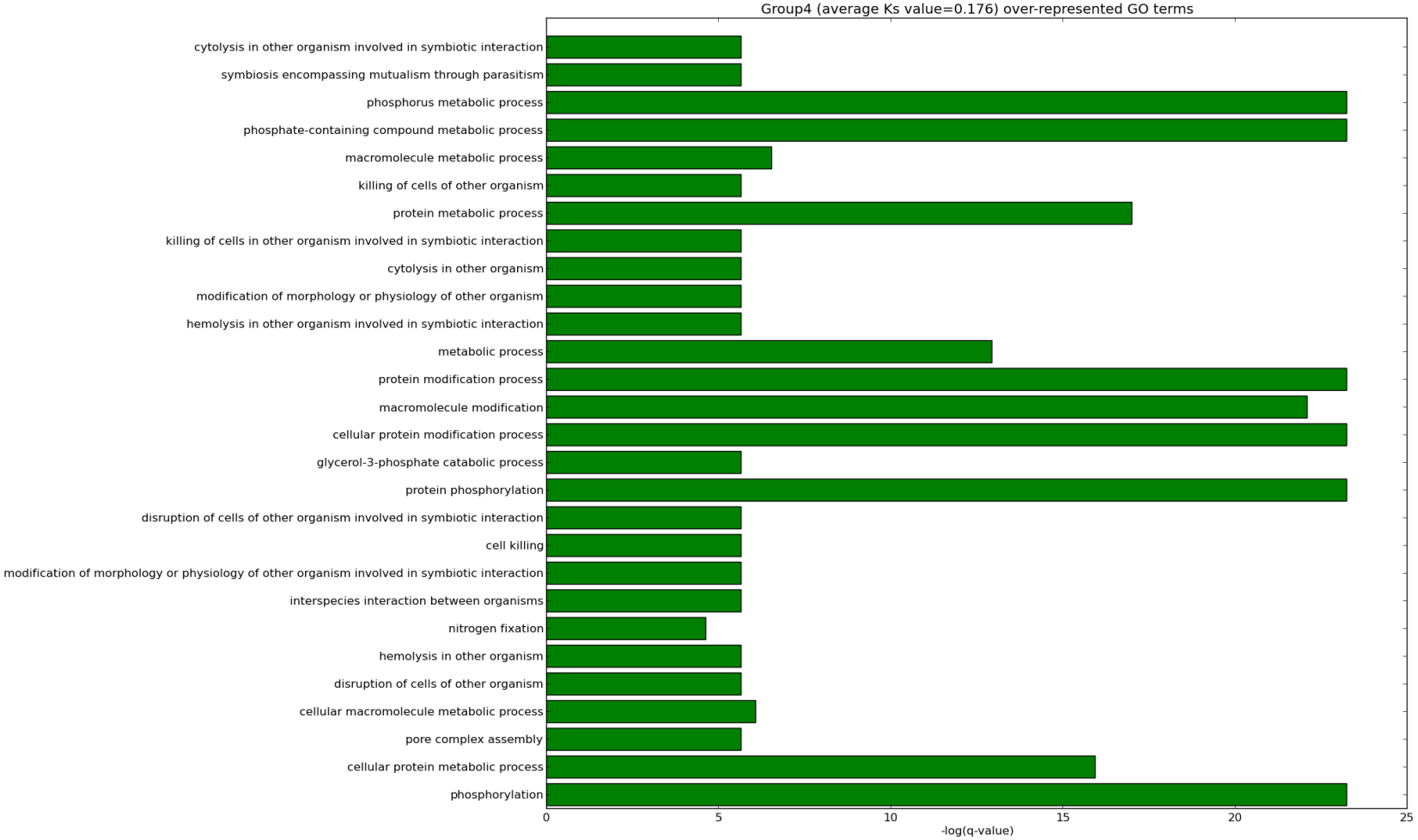
B.


Group2 (average Ks value=0.755) over-represented GO terms
cytolysis in other organism involved in symbiotic interaction
unsaturated fatty acid metabolic process
symbiosis encompassing mutualism through parasitism
phosphorus metabolic process
phosphate-containing compound metabolic process
icosanoid biosynthetic process
leukotriene metabolic process
macromolecule metabolic process
cellular metabolic process
killing of cells of other organism
protein metabolic process
killing of cells in other organism involved in symbiotic interaction
alkene biosynthetic process
primary metabolic process
modification of morphology or physiology of other organism
hemolysis in other organism involved in symbiotic interaction
metabolic process
protein modification process
macromolecule modification
cellular protein modification process
cellular process
cytolysis in other organism
nitrogen utilization
protein phosphorylation
leukotriene biosynthetic process
disruption of cells of other organism involved in symbiotic interaction
cell killing
olefin biosynthetic process
modification of morphology or physiology of other organism involved in symbiotic interaction
multi-organism process
interspecies interaction between organisms
icosanoid metabolic process
hemolysis in other organism
disruption of cells of other organism
cellular macromolecule metabolic process
pore complex assembly
cellular protein metabolic process
regulation of nitrogen utilization
olefin metabolic process
nitrogen cycle metabolic process
cellular alkene metabolic process
phosphorylation
unsaturated fatty acid biosynthetic process
0
50
100
150
200
250
-log(q-value)

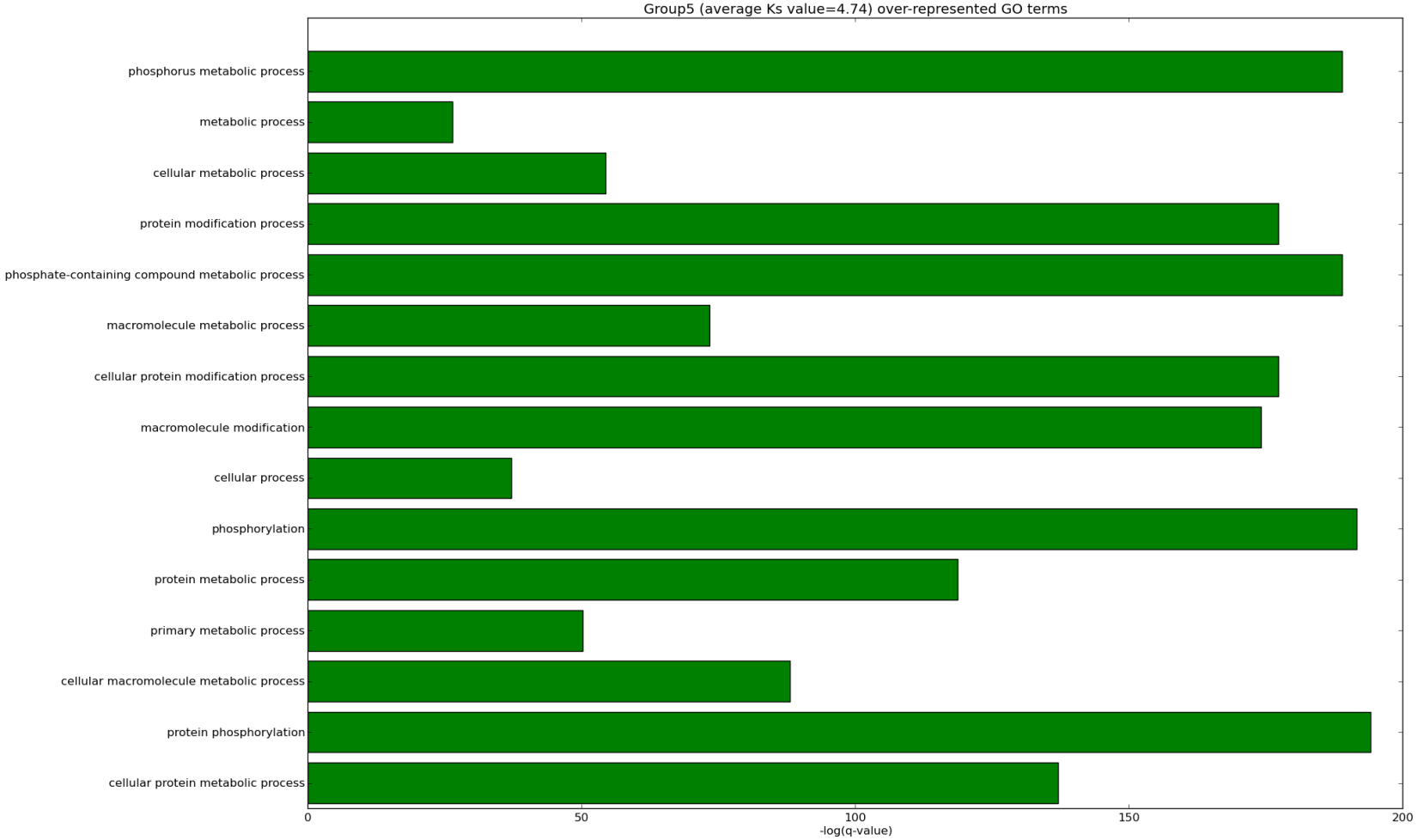
C.


Group3 (average Ks value=1.32) over-represented GO terms
regulation of nitrogen utilization
phosphorus metabolic process
metabolic process
cellular metabolic process
protein modification process
phosphate-containing compound metabolic process
macromolecule metabolic process
cellular protein modification process
macromolecule modification
cellular process
phosphorylation
protein metabolic process
primary metabolic process
cellular macromolecule metabolic process
nitrogen utilization
protein phosphorylation
cellular protein metabolic process
0
50
100
150
200
250
300
350
400
450
-log(q-value)

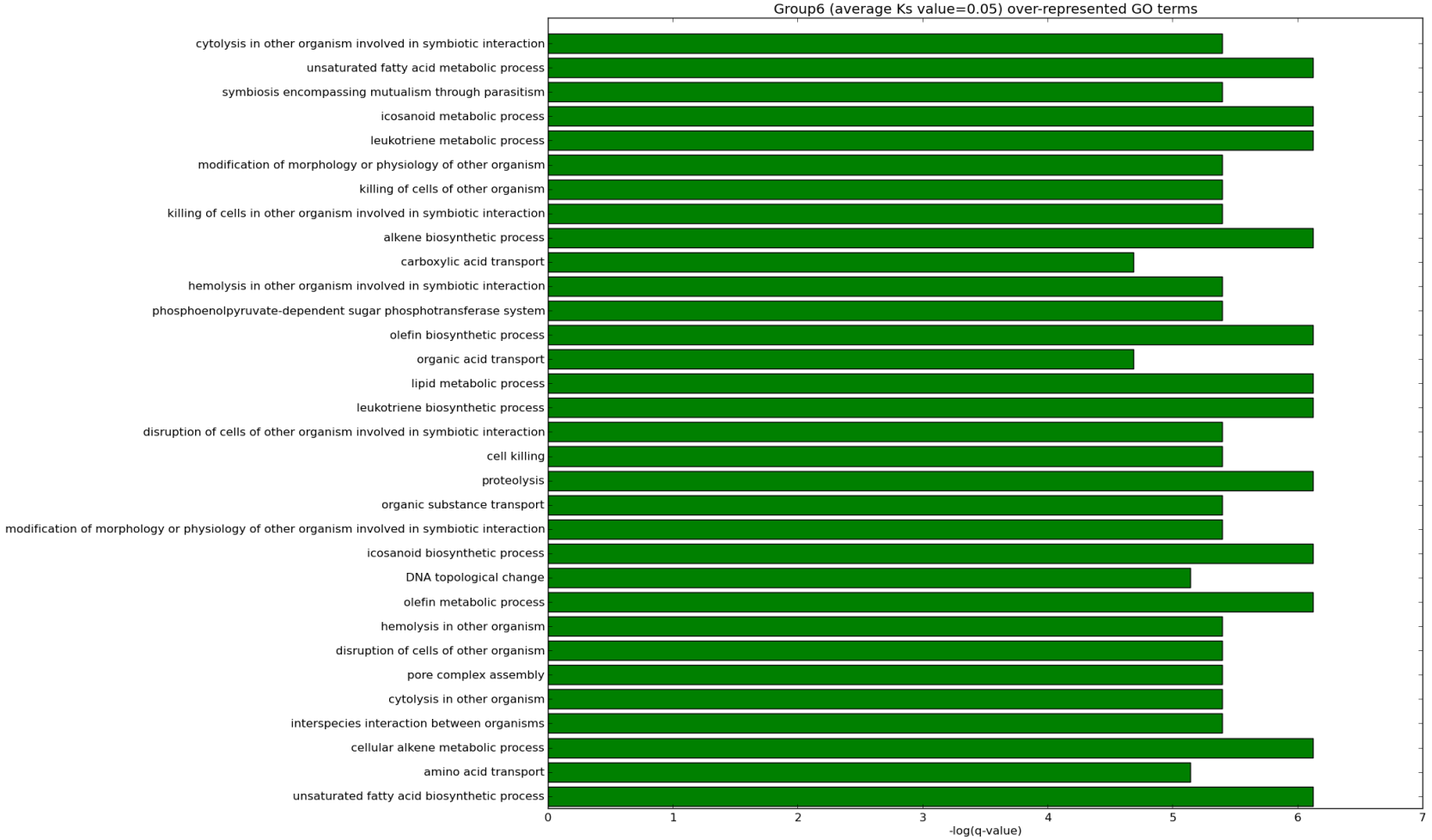
D.


Group4 (average Ks value=0.176) over-represented GO terms
cytolysis in other organism involved in symbiotic interaction
symbiosis encompassing mutualism through parasitism
phosphorus metabolic process
phosphate-containing compound metabolic process
macromolecule metabolic process
killing of cells of other organism
protein metabolic process
killing of cells in other organism involved in symbiotic interaction
cytolysis in other organism
modification of morphology or physiology of other organism
hemolysis in other organism involved in symbiotic interaction
metabolic process
protein modification process
macromolecule modification
cellular protein modification process
glycerol-3-phosphate catabolic process
protein phosphorylation
disruption of cells of other organism involved in symbiotic interaction
cell killing
modification of morphology or physiology of other organism involved in symbiotic interaction
interspecies interaction between organisms
nitrogen fixation
hemolysis in other organism
disruption of cells of other organism
cellular macromolecule metabolic process
pore complex assembly
cellular protein metabolic process
phosphorylation
0
5
10
15
20
25
-log(q-value)

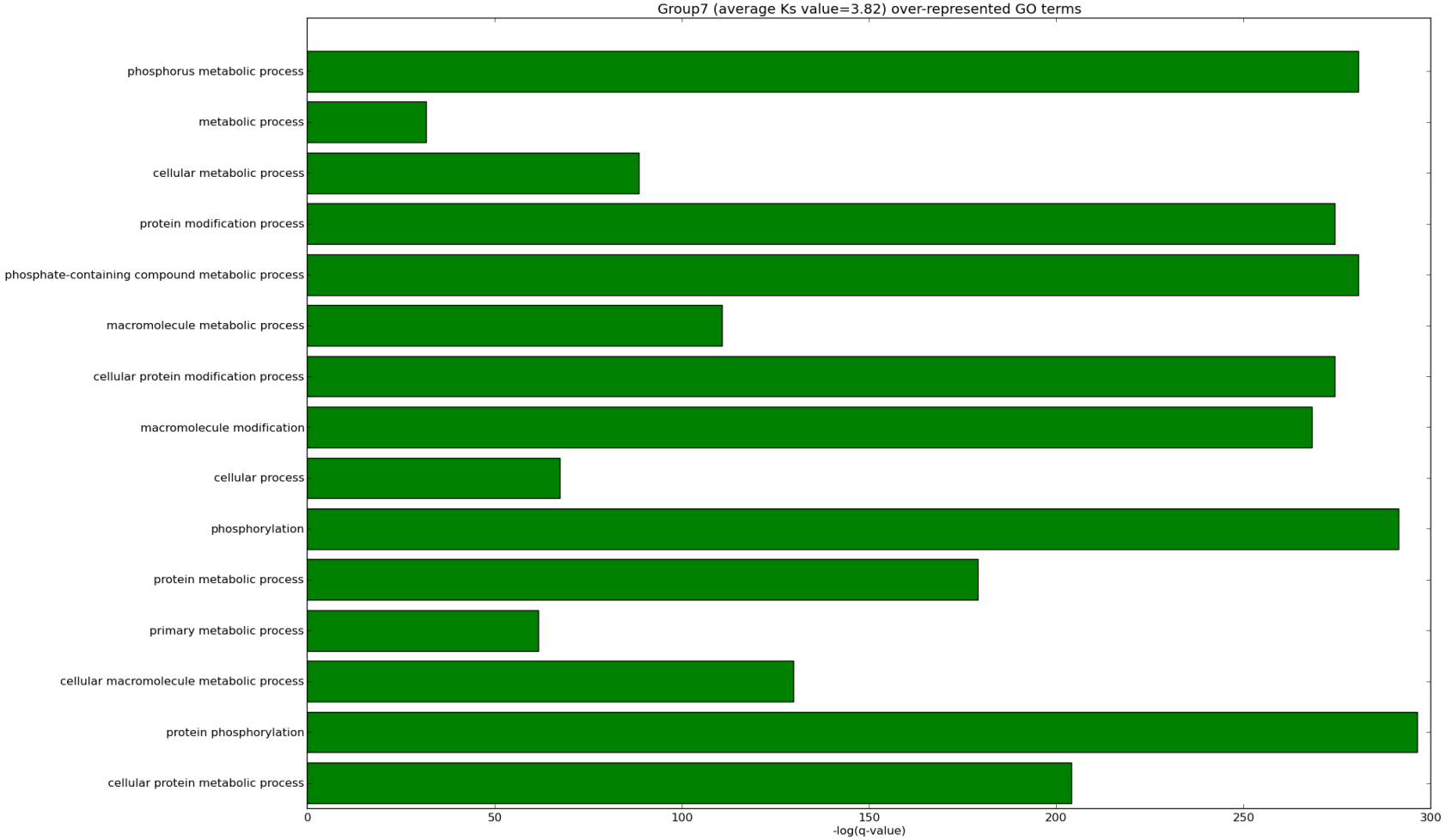
E.


Group5 (average Ks value=4.74) over-represented GO terms
phosphorus metabolic process
metabolic process
cellular metabolic process
protein modification process
phosphate-containing compound metabolic process
macromolecule metabolic process
cellular protein modification process
macromolecule modification
cellular process
phosphorylation
protein metabolic process
primary metabolic process
cellular macromolecule metabolic process
protein phosphorylation
cellular protein metabolic process
0
50
100
150
200
-log(q-value)

F.


Group6 (average Ks value=0.05) over-represented GO terms
cytolysis in other organism involved in symbiotic interaction
unsaturated fatty acid metabolic process
symbiosis encompassing mutualism through parasitism
icosanoid metabolic process
leukotriene metabolic process
modification of morphology or physiology of other organism
killing of cells of other organism
killing of cells in other organism involved in symbiotic interaction
alkene biosynthetic process
carboxylic acid transport
hemolysis in other organism involved in symbiotic interaction
phosphoenolpyruvate-dependent sugar phosphotransferase system
olefin biosynthetic process
organic acid transport
lipid metabolic process
leukotriene biosynthetic process
disruption of cells of other organism involved in symbiotic interaction
cell killing
proteolysis
organic substance transport
modification of morphology or physiology of other organism involved in symbiotic interaction
icosanoid biosynthetic process
DNA topological change
olefin metabolic process
hemolysis in other organism
disruption of cells of other organism
pore complex assembly
cytolysis in other organism
interspecies interaction between organisms
cellular alkene metabolic process
amino acid transport
unsaturated fatty acid biosynthetic process
0
1
2
3
4
5
6
7
-log(q-value)

G.


Group7 (average Ks value=3.82) over-represented GO terms
phosphorus metabolic process
metabolic process
cellular metabolic process
protein modification process
phosphate-containing compound metabolic process
macromolecule metabolic process
cellular protein modification process
macromolecule modification
cellular process
phosphorylation
protein metabolic process
primary metabolic process
cellular macromolecule metabolic process
protein phosphorylation
cellular protein metabolic process
0
50
100
150
200
250
300
-log(q-value)

H.

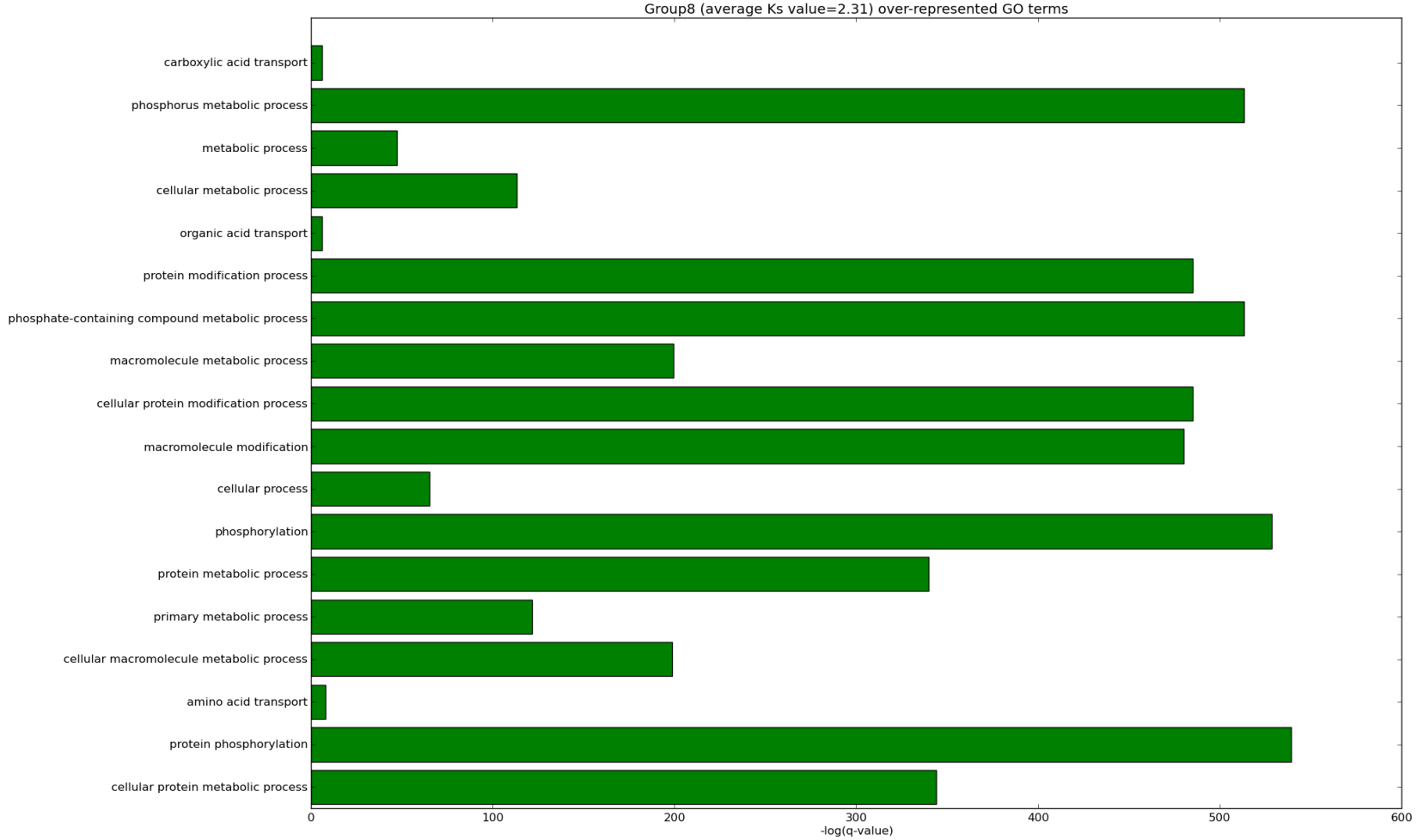

Group8 (average Ks value=2.31) over-represented GO terms
carboxylic acid transport
phosphorus metabolic process
metabolic process
cellular metabolic process
organic acid transport
protein modification process
phosphate-containing compound metabolic process
macromolecule metabolic process
cellular protein modification process
macromolecule modification
cellular process
phosphorylation
protein metabolic process
primary metabolic process
cellular macromolecule metabolic process
amino acid transport
protein phosphorylation
cellular protein metabolic process
0
100
200
300
400
500
600
-log(q-value)

Figure S10. Features of the *R. irregularis* protein-coding genes compared to other representative fungi. Intervals between quartile 1 and quartile 3 are represented by grey bars. Median and number of introns by gene are indicated in red for *Rhizophagus irregularis*. The intergenic region length is likely underestimated as the assembly is highly fragmented.

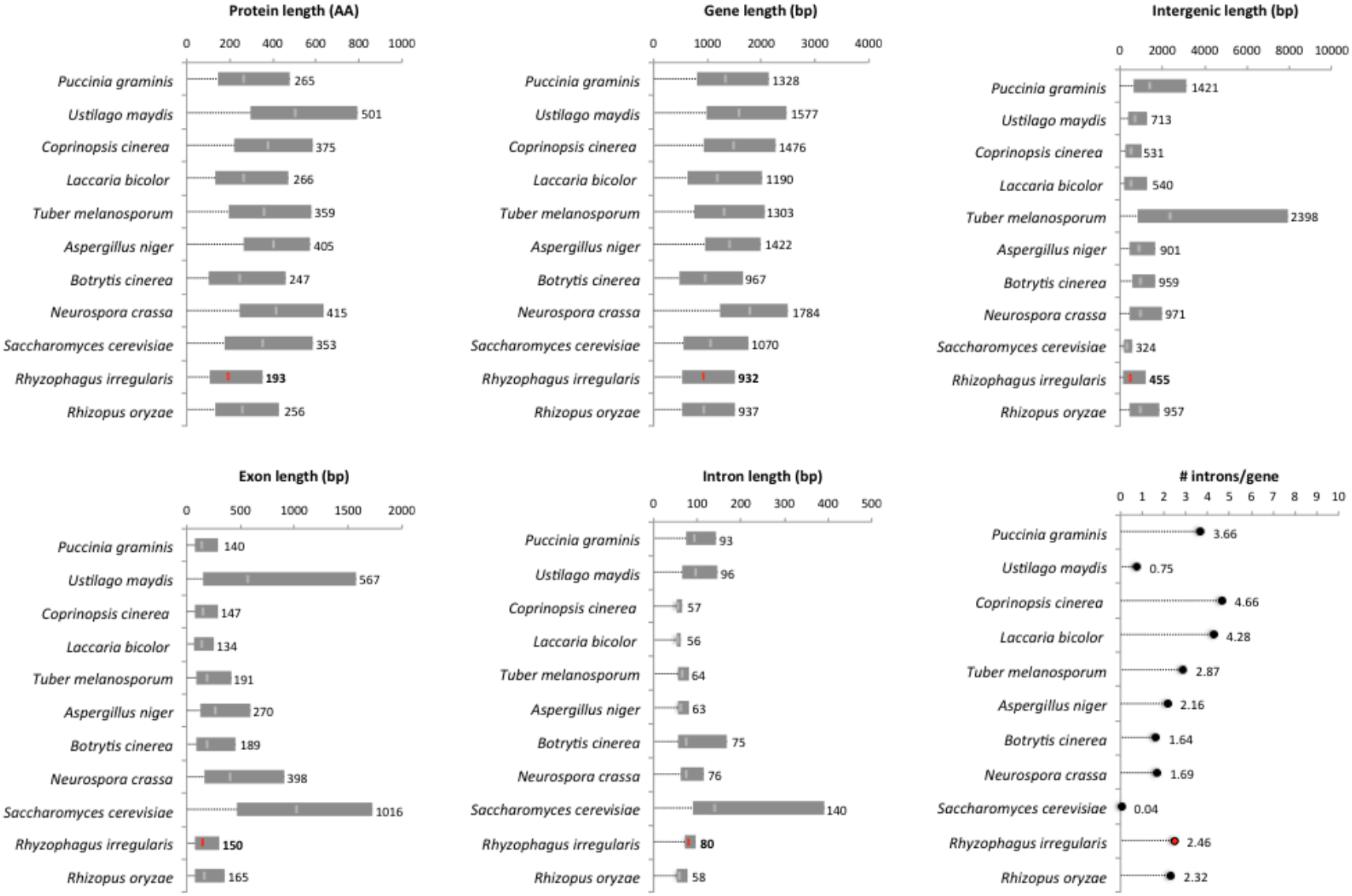


Figure S11. Analysis of molecular divergence between the *R. irregularis* proteome and selected organisms. The *Rhizophagus irregularis*–Mortierellomycotina pair displays the highest amino acid identity, in agreement with their phylogeny. In the figure, we represent the cumulative frequencies of amino acid identity across each set of potential orthologous pairs shown. (Basidiomycota: *Puccinia graminis, Ustilago maydis, Coprinopsis cinerea, Laccaria bicolor;* Ascomycota*: Tuber melanosporum, Aspergillus niger, Botrytis cinerea, Neurospora crassa, Saccharomyces cerevisiae;* Mucoromycotina*: Rhizopus oryzae;* Mortierellomycotina: *Mortierella elongata;* Amoebozoa/Choanoflagellida*: Dictyostelium discoideum, Monosiga brevicollis).*

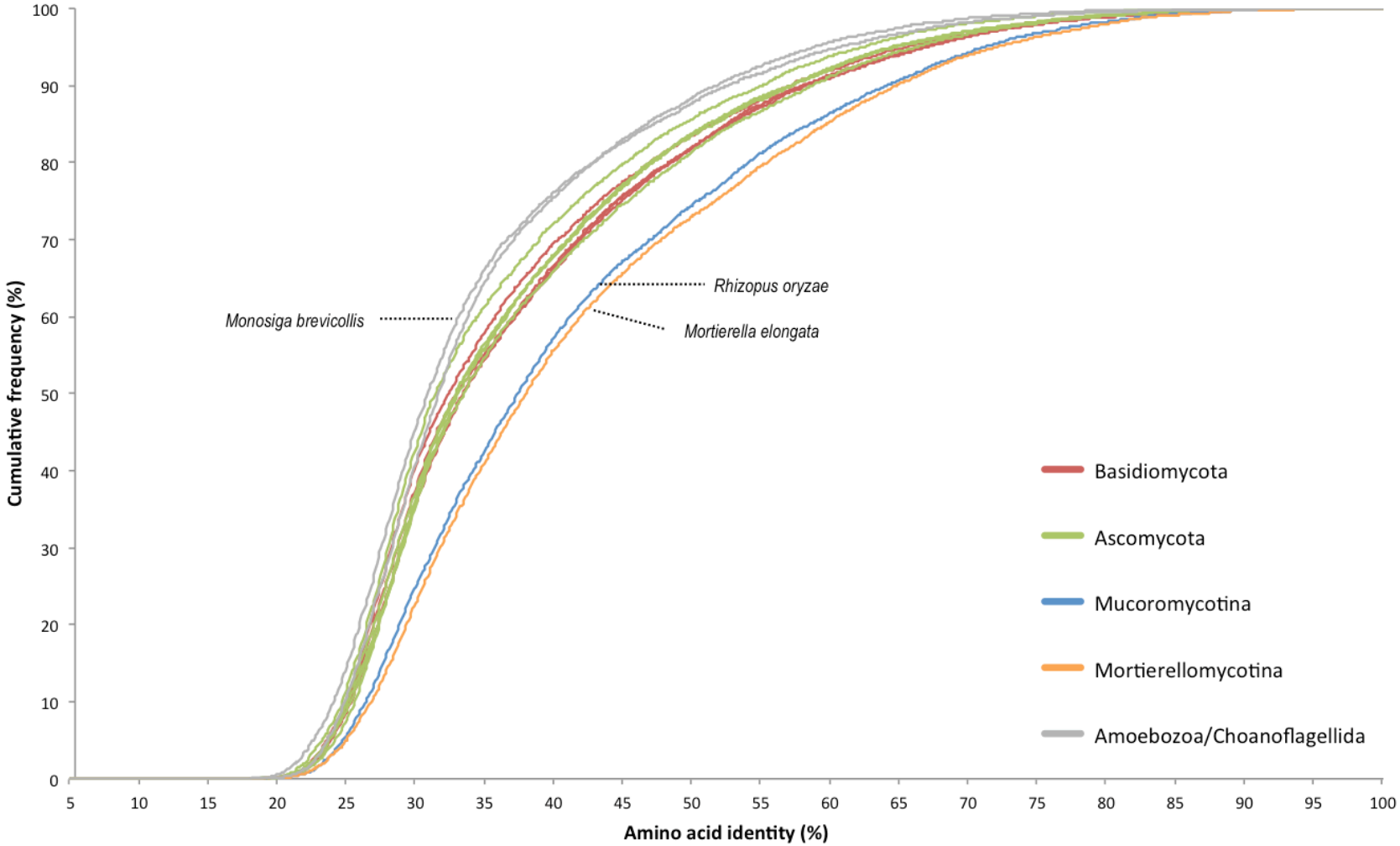


Figure S12. Gene orthology and evolution. Orthology assignment of 12 fungal genomes. Bars are subdivided to represent different types of orthology relationships. "*Core genes*" indicates universal core genes but absence in two genomes is tolerated. "*Basidio*" indicates Basidiomycota-specific genes with presence in at least two genomes. "*Asco*" indicates Ascomycota-specific genes with presence in at least two genomes; "*Basal*" indicates basal species-specific genes with presence in at least two genomes; "*Patchy*" indicates genes that are present in at least two genomes within different groups. "*Species specific*" indicates genes with no (detectable) homologs in other species, i.e. orphan genes. The phylogeny on the left was calculated using maximum likelihood analyses of a concatenated alignment of 85 single-copy proteins. The tree was rooted using *Dictyostelium discoideum* as outgroup.

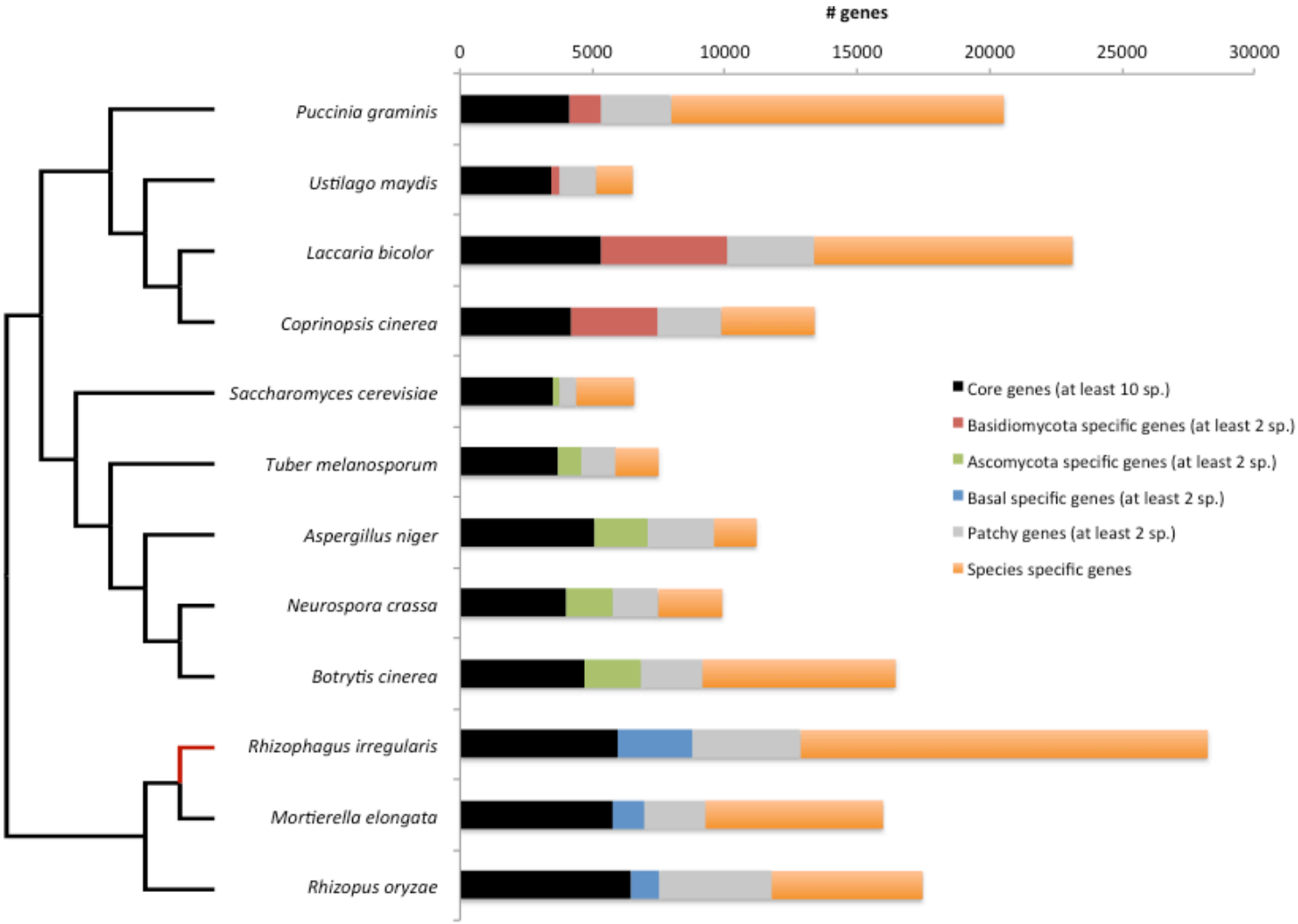


Figure S13. Genome redundancy in the *R. irregularis* genome. The figure represents the total number of gene families in each species or node. The numerals on branches and pie charts at each branch terminus show the proportion of expanded (red), unchanged (black/grey) or contracted (blue) gene families along lineages by comparison to the putative pan-genome. CAFE analysis, *p*-value <0.001. MRCA, most recent common ancestor.

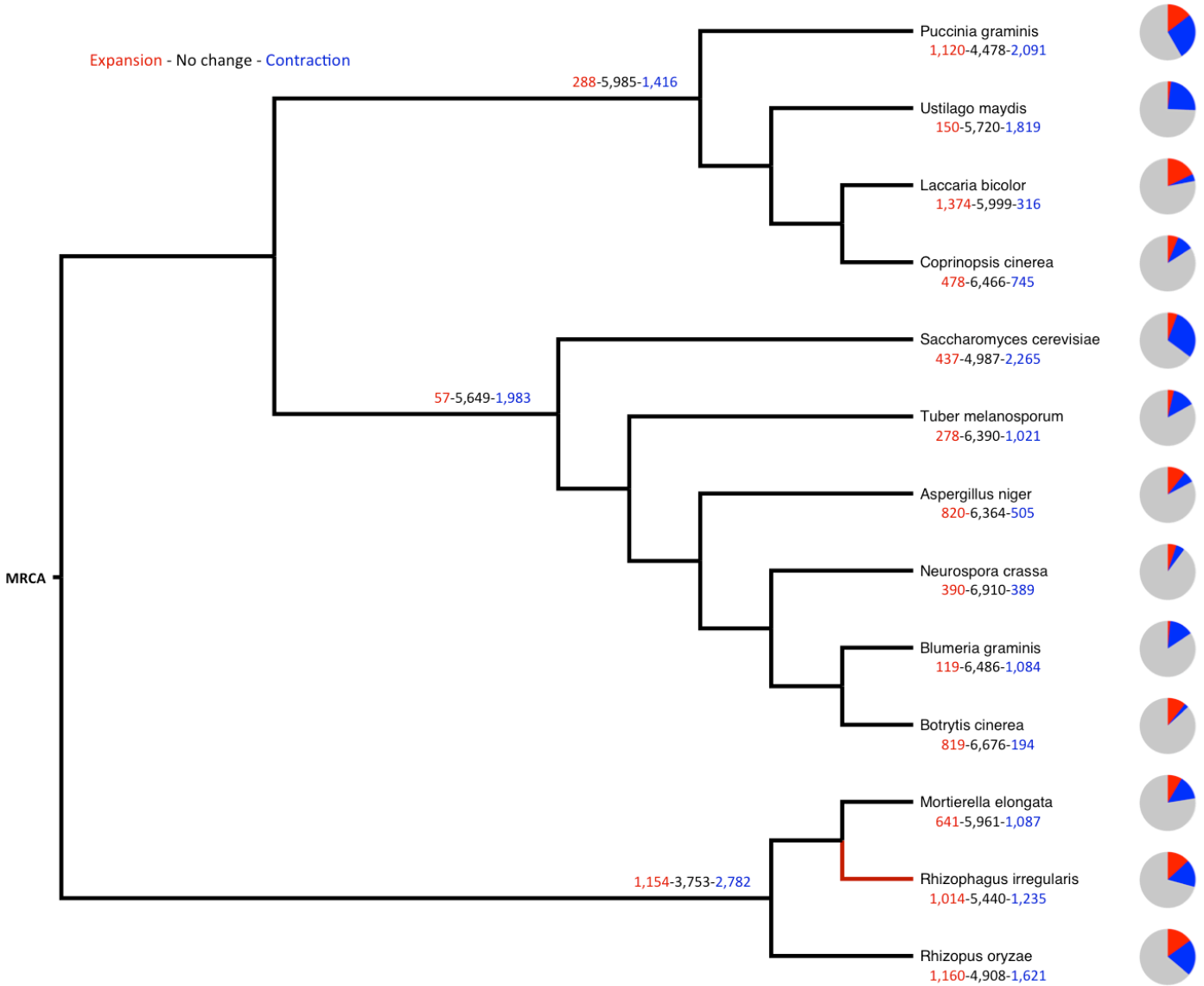


Figure S14. Phylogenetic reconstruction, expression profiling and protein domains in full-length *Rhizophagus*-specific tyrosine kinase-like proteins. **A**, Phylogenetic tree of 88 DAOM-197198-specific whole catalytic domain tyrosine kinase-like proteins obtained via neighbor-joining phylogenetic analysis. **B**, Expression levels measured by mapping RNA-Seq reads from DAOM-197198 germinated spores (DAOM), C2 germinated spores (C2), *R. diaphanum* germinated spores (*G. diaph*), DAOM-197198 spores germinated for 2 days (Spores 2d), DAOM-197198 spores germinated for 9 days (Spores 2d) and in *Rhizophagus-Medicago* symbiotic roots (*in Planta*). Expression levels is represented by a colour scale from 0 (white) to 100 (red). **C**, Location of tyrosine kinase-like and other PFAM domains (LRR_1: Leucine-rich repeat; Sel1: Sel1 repeat; TMH: transmembrane helices; TKL: tyrosine kinase-like; HMG_box: high mobility group).

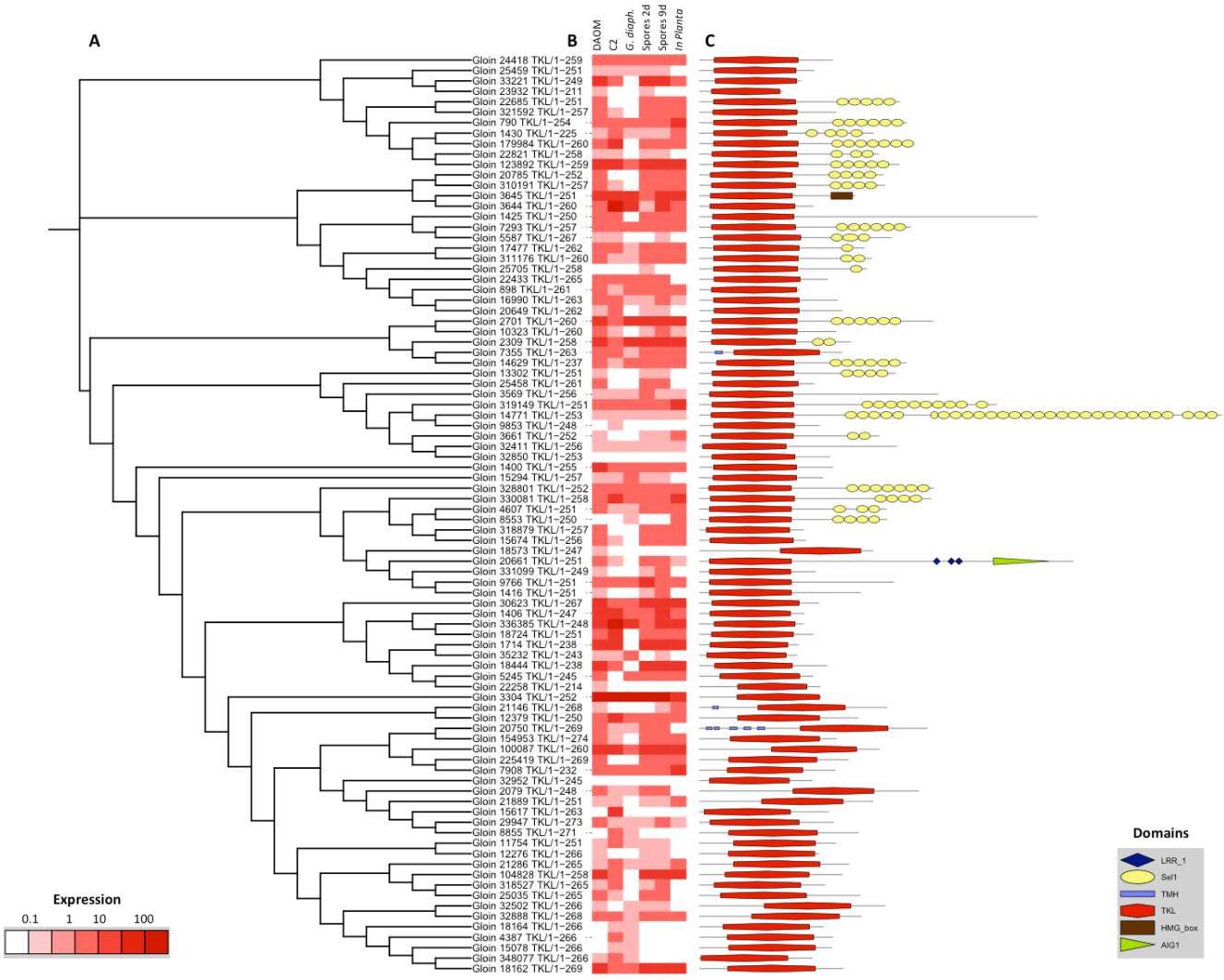


Figure S15. Double clustering of the CAZyme families acting on plant cell wall polysaccharides and lignin-related oxidoreductases from *Rhizophagus irregularis* (red arrow and line) and representative fungal genomes. Top tree: fungal species. Left tree: the enzyme families are represented by their class (GH, glycoside hydrolase; PL, polysaccharide lyase; AA1: laccase; AA2: C2 peroxidase) and family number according to the Carbohydrate-Active enZYmes Database database (http://www.cazy.org/). Abundance of the different enzymes within a family is represented by a color scale from 0 (blue) to 44 occurrences (red) per species.

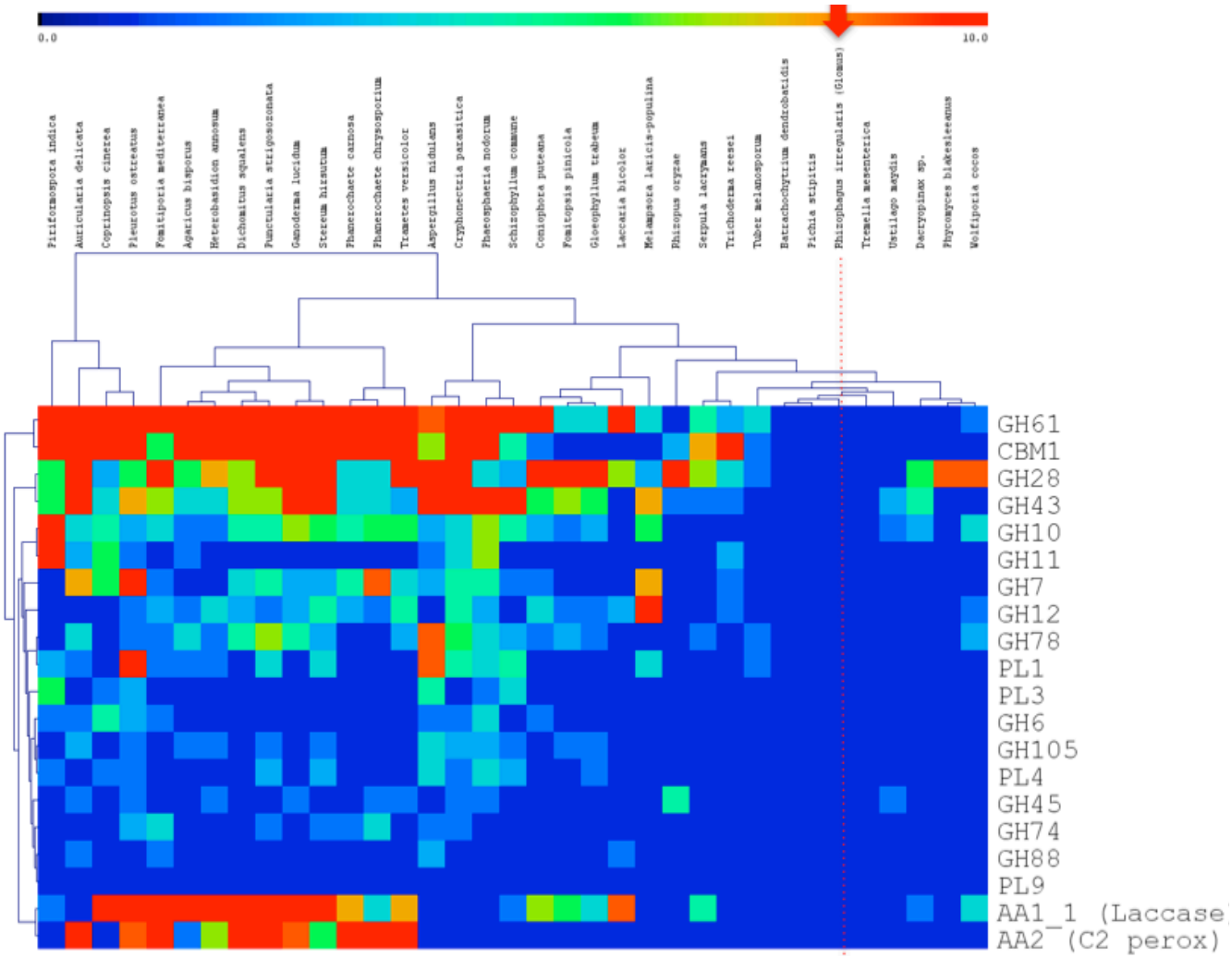


Figure S16. Correlation of KEGG pathway profiles between 12 fungal species from the early diverging Fungi (including *R. irregularis*, black arrow) and Dikarya. Pearson correlation distance matrix was calculated based on presence/absence profile of protein-coding genes assign to each KEGG orthology for each species. Colors are coded from dark red representing high correlation to light red representing low correlation. **A**, All KEGG categories; **B**, Metabolism; **C**, Genetic Information Processing; **D**, Environmental Information Processing and **E**, Cellular Processes. (Pucgr: *Puccinia graminis*; Ustma: *Ustilago maydis*; Copci: *Coprinopsis cinerea*; Lacbi: *Laccaria bicolor*; Tubme: *Tuber melanosporum*; Aspni: *Aspergillus niger*; Botci: *Botrytis cinerea*; Neucr: *Neurospora crassa*; Rhiir: *Rhizophagus irregularis*; Morel: *Mortierella elongata*; Rhior: *Rhizopus oryzae*).

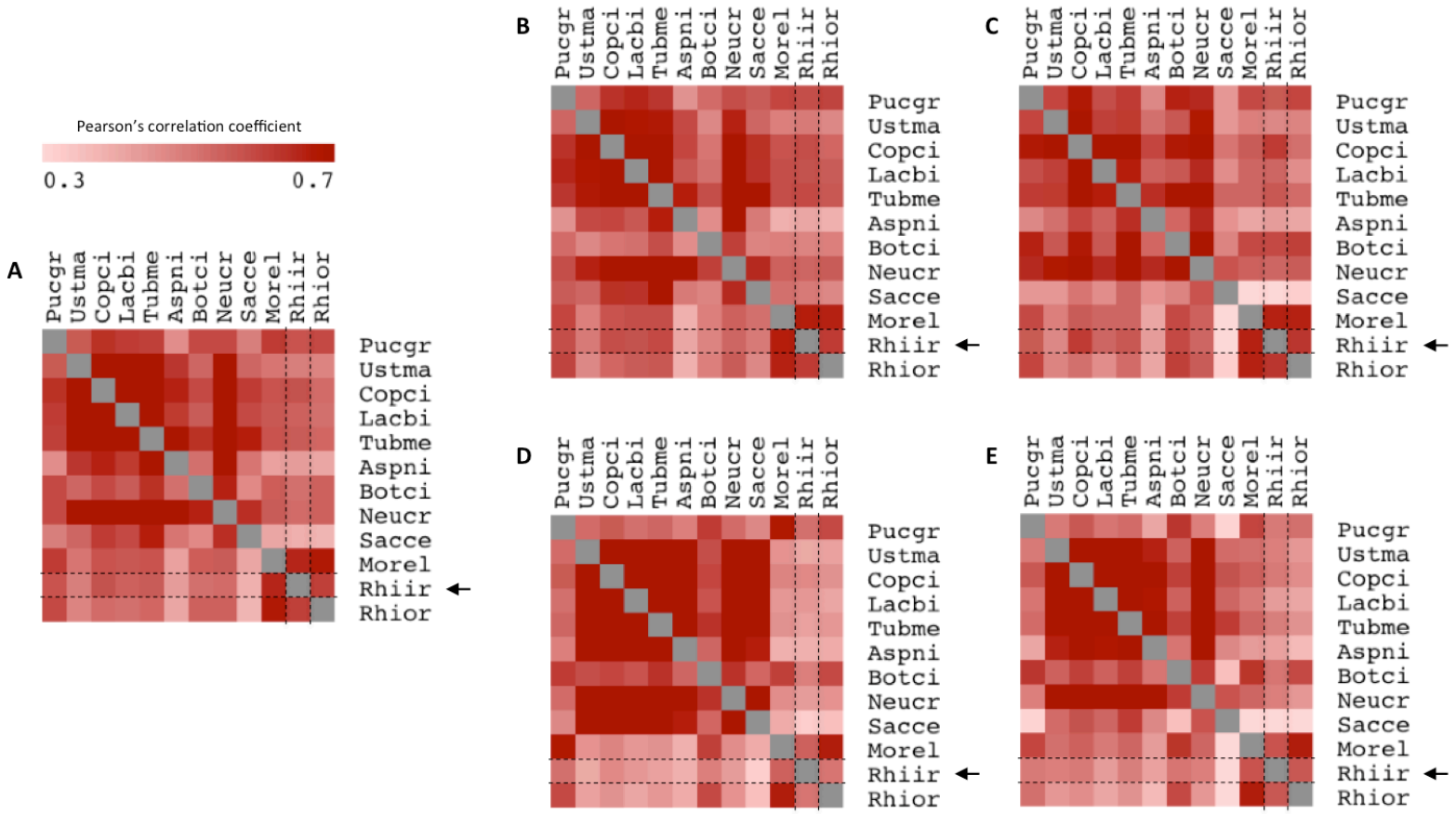


Figure S17. A hierarchical cluster tree showing the relationships between the Glomeromycota *R. irregularis*, and representative Mucoromycotina, Ascomyta and Basidiomycota based on the *Saccharomyces cerevisiae* core genes missing from their genomes (based on presence/absence profile of protein-coding genes assign to *Saccharomyces cerevisiae* core genes for each species) (see Table at http://mycor.nancy.inra.fr/IMGC/GlomusGenome/download3.php?select=anno). Tree was generared using hierarchical clustering (average linkage clustering) method with MeV software (http://www.tm4.org/mev.html). BLASTP e-value, 1^{e}-5. Blugr, *Blumeria graminis*; Pucgr, *Puccinia graminis*; Ustma, *Ustilago maydis*; Copci, *Coprinopsis cinerea*; Lacbi, *Laccaria bicolor*; Mella, *Melampsora larici-populina*; Tubme; *Tuber melanosporum*; Aspni, *Aspergillus niger*; Botci, *Botrytis cinerea*; Neucr, *Neurospora crassa*; Mucci, *Mucor circinelloides*; Phybl, *Phycomyces blakesleeanus*; Rhior: *Rhizopus oryzae*; Batde, *Batrachochytrium dendrobatidis;* Rhiir, *Rhizophagus irregularis*.

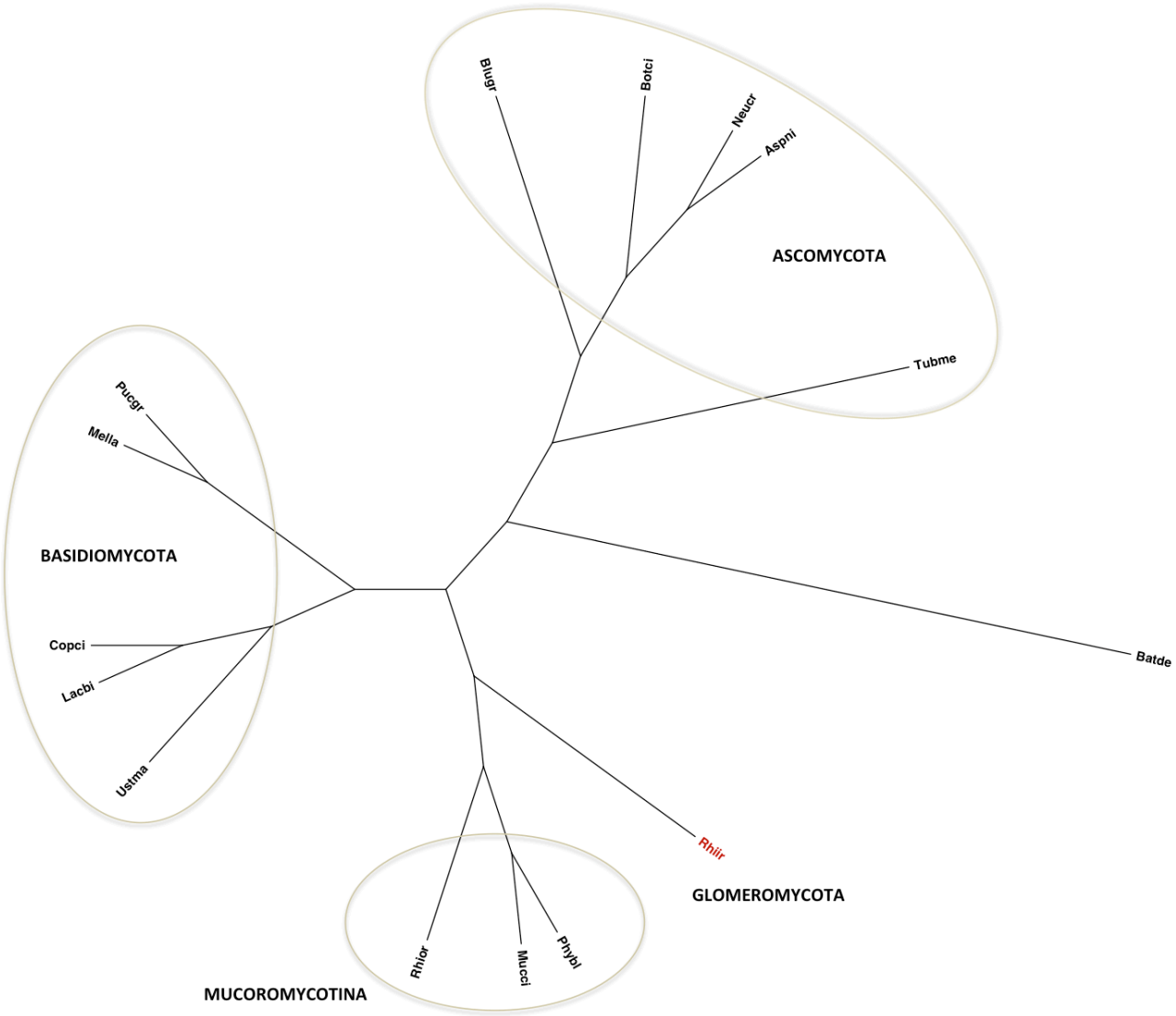


Figure S18. Symbiosis-upregulated genes in *Rhizophagus-Medicago* symbiotic roots. Distribution of genes (%) into functional categories according to the KOG classification. Red bars indicate the distribution of genes induced *in planta* (fold change > 2; FDR < 0.05), whereas grey bars indicate to total gene distribution per KOG category. Asterisks indicate overrepresented KOG categories in induced gene distribution relative to global gene distribution (Fischer's exact test, P < 0.05).

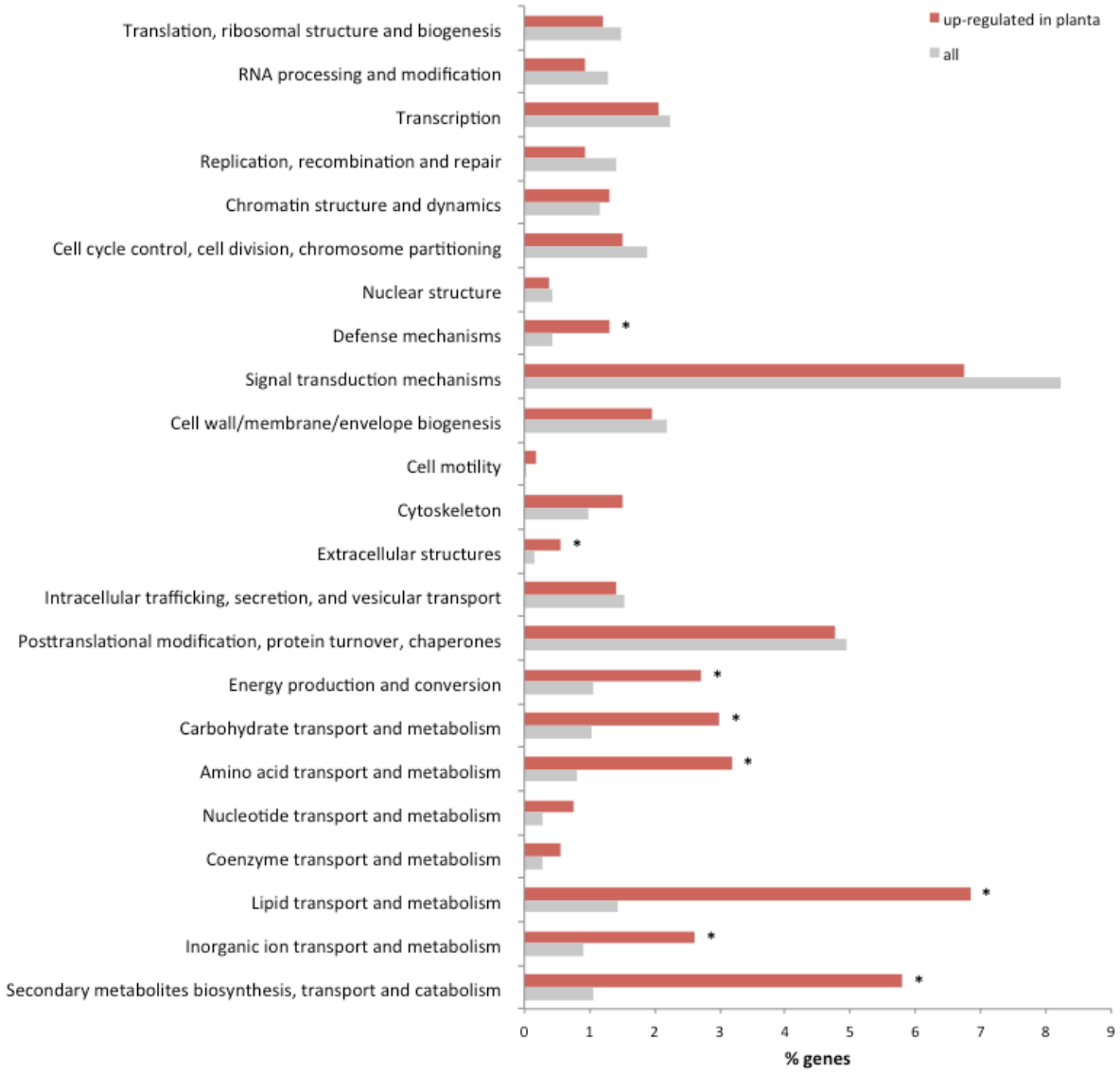


Figure S19. Distribution of genes with MATA-related HMG domains in *R. irregularis* DAOM-197198 and representative species of Ascomycota, Basidiomycota, Zygomycota and Chytridiomycota. Left tree: phylogenetic tree based on ribosomal RNA genes. The location of the MATA-related HMG in the *R. irregularis* assembly and their best reciprocal hit against the GenBank nr database are shown in Table S17.

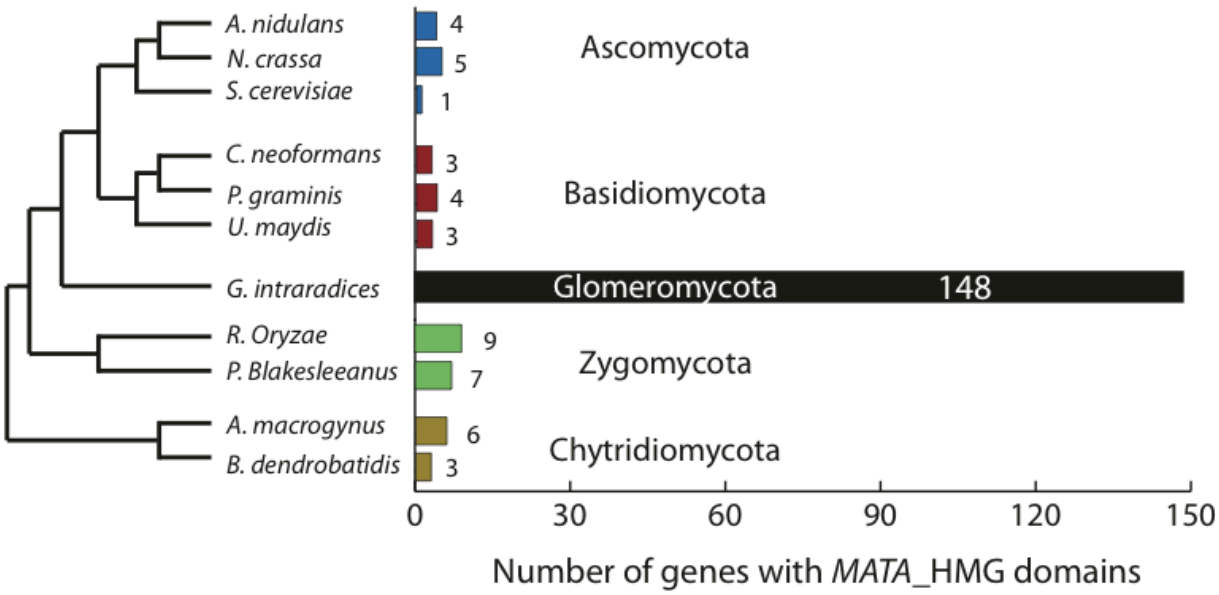
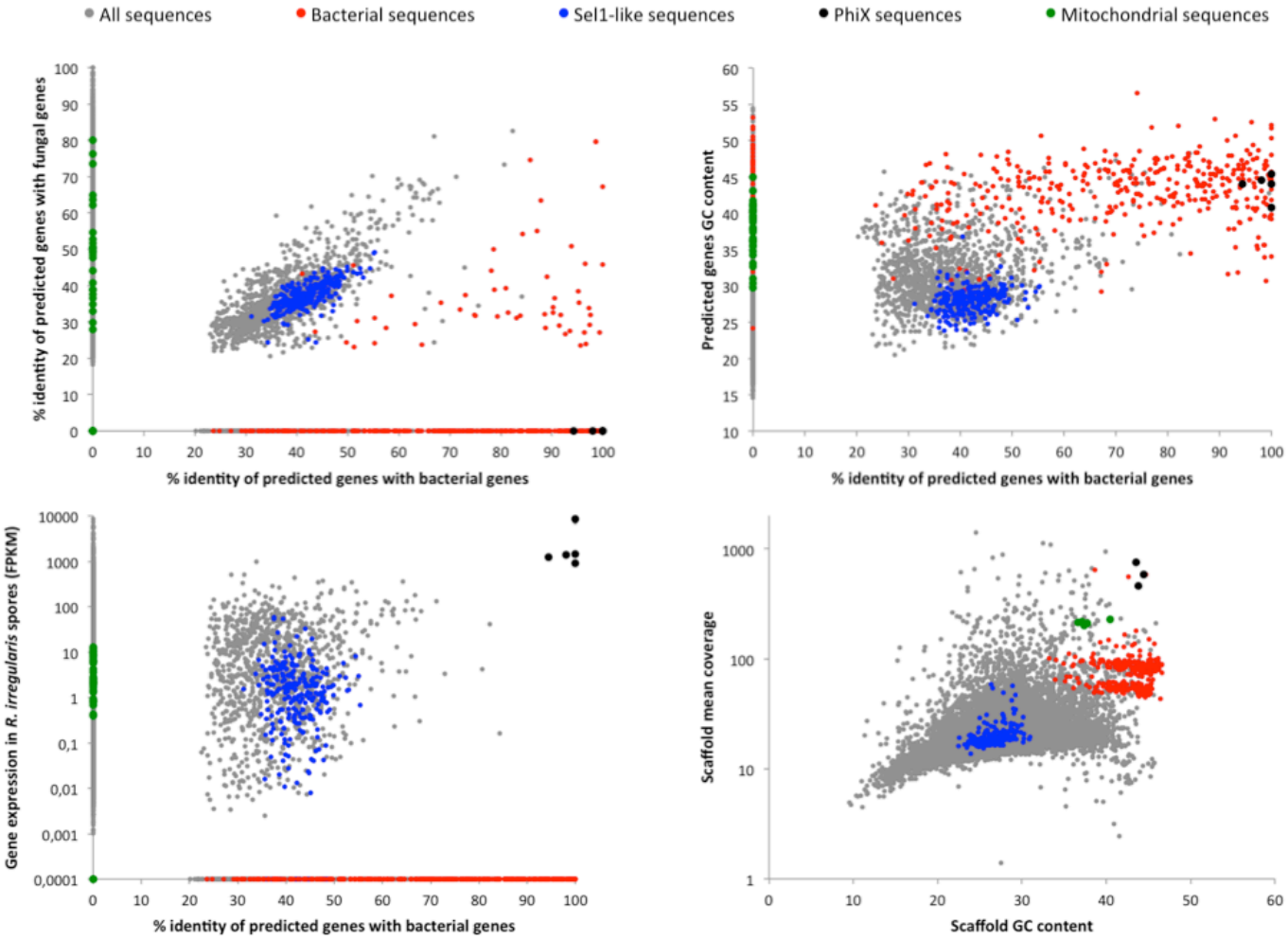


Figure S20. Features of *R. irregularis* genome scaffolds. Top left panel, % identity to fungal and bacterial sequences; top right panel, % identity with bacterial genes *vs.* GC content of genes; bottom left panel, genes expression in germinated spores vs. % identity with bacterial genes; bottom right, scaffold GC content vs. scaffold sequence coverage. Bacterial contaminant (red), Illumina Phix sequence contaminant (black), mitochondrial (green) and Sel1-like (blue) sequences are indicated. All others sequences are represented by grey color.



References

1. Stockinger H, Walker C, Schüssler A. (2009) Glomus intraradices DAOM197198, a model fungus in arbuscular mycorrhiza research, is not Glomus intraradices. *New Phytol.* 183:1176–87
2. Chabot S, Bécard G, Piché Y. (1992) The life cycle of Glomus intraradices in root organ culture. *Mycologia* 84: 315-321
3. Bécard G and Fortin JA (1988) Early events of vesicular-arbuscularmycorrhiza formation on Ri T-DNA transformed roots. *New Phytol.* 108:211-218
4. Koch et al. (2004) High genetic variability and low local diversity in arbuscular mycorrhizal fungal population. *Proc Natl Acad Sci USA* 101:2369–2374
5. Giovanetti M and Mosse B (1980) An evaluation of techniques for measuring vesicular arbuscular mycorrhizal infection in roots. *New Phytol.* 84:489-500
6. Hewitt EJ (1966) Sand and water culture methods used in the study of plant nutrition. *Commonwealth Bureau of Horticulture and Plantation Crops, East Malling, Technical Communication* 22:547
7. Saghai-Maroof MA, Soliman KM, Jorgensen RA, Allard RW (1984) Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics. *Proc Natl Acad Sci USA* 81:8014–8018
8. Lassman T, Hayashizaki Y, Daib CO (2009) TagDust—a program to eliminate artifacts from next generation sequencing data. *Bioinformatics* 25:2839-40
9. Li H and Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics* 26:589-95
10. Li H et al. (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 25:2078-9.
11. Yu J, et al. (2005) The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol* 3:e38.
12. Price AL, Jones NC and Pevzner PA (2005) De novo identification of repeat families in large genomes. *Bioinformatics* 21:i351-8
13. UniProt Consortium (2007) The universal protein resource (UniProt). *Nucleic Acids Res.* 36:D190-5
14. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. *J. Mol. Biol.* 215:403–410
15. Jurka J et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. *Genome Res.* 110:462–467
16. Mc Carthy E and Mc Donald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* 19:362-7
17. Smit AFA, Hubley R and Green P (1996-2010) RepeatMasker Open-3.0. http://www.repeatmasker.org
18. Murat C et al. (2011) Distribution and localization of microsatellites in the Perigord black truffle genome and identification of new molecular markers. *Fungal Genet Biol.* 48:592–601
19. Murat C, Payen T, Petitpierre D, Labbé J (2013) in *The Ecological Genomics of Fungi*, ed. Martin F (Wiley-Blackwell)
20. Riley DE, Krieger JN (2009) UTR dinucleotide simple sequence repeat evolution exhibits recurring patterns including regulatory sequence motif replacements. *Gene* 429:80–86
21. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27:573-580
22. Salamov AA and Solovyev VV (2000) Ab initio Gene Finding in Drosophila Genomic DNA. *Genome Res*. 10:516-522
23. Birney EM Clamp and Durbin R (2004) GeneWise and Genomewise. *Genome Res.* 14:988-995
24. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, and Borodovsky M (2008) Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* 18:1979-1990
25. Zdobnov EM, and Apweiler R (2001) InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17:847-848
26. Bairoch A et al. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.* 33:154-159
27. Ogata H et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 27:29-34
28. Bateman A (2004) The Pfam protein families database. *Nucleic Acids Res*. 32:138D-141

29. Bairoch A (2000) The ENZYME database in 2000. *Nucleic Acids Res*. 28:304-305
30. Ashburner M et al. (2000) Gene Ontology: tool for the unification of biology. *Nat Genet.* 25:25-29
31. Koonin EV et al. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol* 5:R7-R7
32. Nielsen H, Brunak S, and von Heijne G (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.* 12:3-9
33. Krogh A, Larsson B, von Heijne G, and Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol.* 305:567-580
34. Grigoriev IV, et al. (2011) The Genome Portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Res.* 40:D26-32
35. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30:1575-1584
36. De Bie T, Cristianini N, Demuth JP and Hahn MW (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22:1269-1271
37. Eddy SR (2009) A New Generation of Homology Search Tools Based on Probabilistic Inference. *Genome Inform.* 23:205-211
38. Finn RD et al. (2010) The Pfam protein families database. *Nucleic Acids Res.* 38:D211-222
39. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* 8:785-786
40. Horton P et al. (2007) WoLF PSORT: Protein Localization Predictor. *Nucleic Acids Res.* 35:W585-7
41. Emanuelsson O et al. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.* 2:953–971
42. Nielsen H, Engelbrecht J,Brunak S and von Heijne G (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 10:1-6
43. Rawlings ND, Barrett AJ, Bateman A (2012) MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* 40:D343-D350.
44. Khaldi N et al. (2010) SMURF: genomic mapping of fungal secondary metabolite clusters. *Fungal Genet. Biol.* 47:736 –741
45. Miranda-Saavedra D and Barton GJ (2007) Classification and functional annotation of eukaryotic protein kinases. *Proteins* 68:893-914.
46. Cantarel BL et al. (2009) The Carbohydrate-Active EnZymes database(CAZy): an expert resource for glycogenomics. *Nucleic Acids Res.* 37:D233–D238
47. De Caceres M, Oliva F, Font X, Vives S (2007) Ginkgo, a program for non-standard multivariate fuzzy analysis. *Advancesin Fuzzy Sets and Systems* 2:41-56
48. Lang D, Weiche B, Timmerhaus G, Richardt S, Riano-Pachon DM, Correa LGG, Reski R, Mueller-Roeber B, Rensing SA (2010). Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion and correlation with complexity. *Genome Biol Evol* 2:488-503.
49. Krzywinski M et al. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.* 19:1639–1645
50. Robinson JT et al. (2011) Integrative Genomics Viewer. *Nat Biotechnol.* 29:24–26
51. Letunic I and Bork P (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* 39:W475-8
52. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25:1105–1111
53. Trapnell C et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol.* 28:511–515
54. Trapnell C et al. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol.* 31:46–53
55. Rensing SA et al. (2007). An ancient genome duplication contributed to the abundance of metabolic genes in the moss Physcomitrella patens. *BMC Evol Biol* 7:130
56. Lynch M and Conery JS (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290:1151-1155
57. McLachlan G et al. (1999) The EMMIX Algorithm for the Fitting of Normal and t-Components. *J Stat Softw* 4:2
58. Jiao Y et al. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature* 473:97-100

59. Falcon S, Gentleman R (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics* 23:257-258.
60. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Series B Stat Methodol.* 57:289-300.
61. Spanu P, et al. (2010) Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. Science 330:1543-1546.
62. Hogekamp C, Arndt D, Pereira PA, Becker JD, Hohnjec N, Küster H (2011) Laser microdissection unravels cell-type-specific transcription in arbuscular mycorrhizal roots, including CAAT-Box transcription factor gene expression correlating with fungal contact and spread. *Plant Physiology* 157:2023-2043.