

De novo gene birth, horizontal gene transfer and gene duplication as sources of new gene families associated with the origin of symbiosis in *Amanita*

Yen-Wen Wang^{1*}, Jaqueline Hess², Jason C. Slot³, Anne Pringle¹

¹Departments of Botany and Bacteriology, University of Wisconsin-Madison, Madison, WI

²Department of Soil Ecology, Helmholtz Centre for Environmental Research, Leipzig, Germany

³Department of Plant Pathology, The Ohio State University, Columbus, OH

*Author for Correspondence: Yen-Wen Wang, Departments of Botany and Bacteriology, University of Wisconsin-Madison, Madison, WI, USA. Tel: 608-338-2980; Email: ywang883@wisc.edu

Abstract

By introducing novel capacities and functions, new genes and gene families may play a crucial role in ecological transitions. Mechanisms generating new gene families include *de novo* gene birth, horizontal gene transfer and neofunctionalization following a duplication event. The ectomycorrhizal (ECM) symbiosis is a ubiquitous mutualism and the association has evolved repeatedly and independently many times among the fungi, but the evolutionary dynamics enabling its emergence remain elusive. We developed a phylogenetic workflow to first understand if gene families unique to ECM *Amanita* fungi and absent from closely related asymbiotic species are functionally relevant to the symbiosis, and then to systematically infer their origins. We identified 109 gene families unique to ECM *Amanita* species. Genes belonging to unique gene families are under strong purifying selection and are upregulated during symbiosis, compared to genes of conserved or orphan gene families. The origins of seven of the

unique gene families are strongly supported as either *de novo* gene birth (two gene families), horizontal gene transfer (four), and gene duplication (one). An additional 34 families appear new because of their selective retention within symbiotic species. Among the 109 unique gene families, the most upregulated gene in symbiotic cultures encodes an ACC deaminase, an enzyme capable of downregulating the synthesis of the plant hormone ethylene, a common negative regulator of plant-microbial mutualisms.

Keywords

convergent evolution, gene family evolution, mycorrhizal fungi, phylogenomics, mutualism

Significance statement

Mutualisms between fungi and plants appear complex but have evolved repeatedly and independently many times. This convergent evolution is typically explained by gene loss from fungi, but at the origin of a symbiosis new genes also appear in fungi: where do these new genes come from and what do they do? By systematically querying the origin of genes unique to symbiotic *Amanita* fungi and not found in close relatives we discover *de novo* gene birth, horizontal gene transfer, and gene duplication as the sources of a set of highly selected new genes upregulated during symbiosis.

Introduction

Evolutionary novelties are novel properties or features of organisms facilitating adaptation (Mayr 1963; Pigliucci 2008). The concept of an evolutionary novelty can connect dramatic

changes in morphologies or phenotypes with ecological transitions in niche. New gene families, without apparent homologies in ancestors, may be considered as genetic evolutionary novelties because they are heritable features potentially shaping adaptations and niche transitions (Villanueva-Cañas *et al.* 2017). New gene families are thought to have three principal sources (Long *et al.* 2013; Andersson *et al.* 2015): *de novo* gene birth, horizontal gene transfer (HGT) and gene duplication (Figure 1). However, genes as evolutionary novelties remain understudied and the functions of many young gene families are unknown.

De novo gene birth involves the formation of protein-coding exons from ancestral non-coding loci. Newly evolved exons are typically shorter and bear weaker signatures of purifying selection compared to existing genes (Carvunis *et al.* 2012; Ruiz-Orera *et al.* 2018; Vakirlis *et al.* 2018). Often, *de novo* genes are identified by the absence of homologous genes in protein databases; hypothetical *de novo* genes are confirmed by aligning their DNA sequences against putatively homologous, non-coding sequences found in closely related species (Cai *et al.* 2008; Knowles & McLysaght 2009). A robust example of a *de novo* gene birth is the *BSC4* gene of *Saccharomyces cerevisiae*. Gene expression data suggest *BSC4* is not a pseudogene, and the sequence homologies between *BSC4* and syntenic but non-coding regions in closely related species confirm the *de novo* nature of this gene (Cai *et al.* 2008).

As mechanisms mediating the emergence of evolutionary novelties, HGT and gene duplication are better understood compared to *de novo* gene birth. HGT genes are often identified when the topologies between a species phylogeny and the phylogeny of a putative HGT gene family are inconsistent (Keeling & Palmer 2008; Husnik & McCutcheon 2018). Because HGT typically involves the movement of genes into distantly related lineages, HGT genes may have distinct properties compared to surrounding genes and preserve a degree of the

donor genome's properties (Keeling & Palmer 2008). HGT is well documented in bacteria, while one of the most famous examples of HGT among eukaryotes is the HGT from a fungus to aphids (tribe Macrosiphini) (Moran & Jarvik 2010). The event enabled aphids to synthesize carotenoids (Moran & Jarvik 2010).

Gene duplication introduces paralogs of redundant sequence into a genome. Because they are copies, paralogs can escape the functional constraints of the original gene and undergo positive selection for new functions (Ohno 1970; Zhang 2003). A neofunctionalized gene copy will be more diverged from the originating gene compared to a copy which retains the same function (Assis & Bachtrog 2013). The relative timing of duplication events can be inferred by reconciling the gene tree with the species tree (Bansal *et al.* 2012; Jacox *et al.* 2016). A clear example of gene duplication and neofunctionalization involves the duplications of olfactory receptor genes in insects (Saad *et al.* 2018). Duplications created redundant paralogs and the paralogs evolved the ability to bind new ligands (Saad *et al.* 2018).

While *de novo* gene birth, HGT and gene duplication can each give rise to new gene families, their relative influence on genomes remains enigmatic. Nonetheless, these mechanisms have clearly shaped niche transitions. For example, the plant pathogen *Pyrenophora tritici-repentis* acquired its *ToxA* virulence gene through HGT from *Stagonospora nodorum*, enabling *P. tritici-repentis* to emerge as a devastating pathogen of wheat (Friesen *et al.* 2006). The transition of tetrapods from water to land was mediated by the duplication and neofunctionalization of *HOX* genes, resulting in the evolution of limbs from fins (Soshnikova *et al.* 2013). The evidence for *de novo* gene birth as a driver of niche transition is indirect. One potential example involves cnidarians (e.g. *Hydra*, jellyfish, coral, etc.); several genes involved in unique predatory behaviors are cnidarian-specific, suggesting their origin is *de novo* (Milde *et*

al. 2009). While associations between new genes and niche transitions have been explored in multiple systems, few have taken a whole genome approach. Genomics may enable the discovery of all genes, including previously unknown genes, associated with a transition event.

Ectomycorrhizal (ECM) symbioses stand among the most robust examples of niche transitions in nature, having evolved independently multiple times across the fungal kingdom (Matheny *et al.* 2006; Tedersoo *et al.* 2010; Bittleston *et al.* 2016). ECM symbioses are mutualistic associations between fungi and plants and enable the exchange of nutrients and photosynthetically derived carbon. The associations can be identified by a morphological feature termed the Hartig net, which appears as hyphal growth between plant cortical cells (Smith & Read 2008). Research on ECM niche transitions focuses on gene loss, and gene loss appears to characterize diverse origins of the symbiosis (Kohler *et al.* 2015; Peter *et al.* 2016; Hess *et al.* 2018; Murat *et al.* 2018). Although the dynamic of gene loss may explain the repeated emergence of ECM symbiosis across distinct lineages, it does not resolve the mechanisms underpinning the evolution of the association (e.g. how ECM fungi suppress or endure plant immune responses). Gene gain is more rarely the focus of ongoing work, but gene gain may enable the formation of symbiotic structures and exchange of resources. For example, small secreted proteins (SSPs) appear to play a crucial role in fungal-plant communication and SSPs have a larger repertoire in at least some ECM species compared to asymbiotic species (Plett, Daguerre *et al.* 2014; Kohler *et al.* 2015). Other studies have identified additional gene gains associated with the transition from a saprotrophic to ECM niche, for example expansions in cytochrome P450 and berberine bridge enzyme gene families (Hess *et al.* 2018). Although multiple lines of evidence suggest a role for new genes in transitions to the ECM niche, the origins of these genes remain unknown.

The fungal genus *Amanita* is an emerging evolutionary model and ideal system to test for connections between evolutionary novelty and gene gain. A single, well-resolved niche transition marks the origin of ECM *Amanita* from asymbiotic ancestral lineages (Wolfe *et al.* 2012; Hess & Pringle 2014; Hess *et al.* 2018). While genomic restructuring within ECM *Amanita* does involve the loss of plant cell wall degrading enzymes (Wolfe *et al.* 2012; Hess *et al.* 2018), the presence of gene families found only in ECM *Amanita* suggests they may also play a role in mediating the niche transition (Hess & Pringle 2014; Hess *et al.* 2018). By taking a closer look at these novel gene families, we aim to decipher the genetic underpinnings of the ECM symbiosis. We hypothesize novel genes enabled new functions within the emerged ECM lineage and seek to understand their sources.

Our aims are to (1) explore whether gene families unique to ECM *Amanita* function in the symbiosis and (2) identify the putative origins of these gene families. We developed a phylogenetic workflow to investigate the properties and origins of unique gene families, defined as genes only found in and shared by species of ECM *Amanita*. Analyses of transcriptomes and tests for selection support the hypothesis that unique gene families shaped the formation of the mutualism. Our workflow suggests all three gene acquisition processes were at play during the niche transition in *Amanita*, but HGT gave rise to the majority of new genes that retain enough signal for us to infer their origins.

Materials and Methods

Genome sequencing and annotation

The genomes of five *Amanita* species and one *Volvariella* species, including three ECM fungi (*A. muscaria* var. *guessowii*, *A. brunnescens* and *A. polypyramis*) and three asymbiotic

fungi (*A. inopinata*, *A. thiersii* and *V. volvacea*), were used to identify gene families unique to symbiotic species. Genome sequencing and annotation is fully described in Hess & Pringle (2014) and Hess *et al.* (2018). Data of four of the genomes are available through NCBI's GenBank ([Acc. JNHV02000000, JNHW02000000, JNHY02000000, JNHZ02000000](#)) and all data and developed bioinformatic pipelines are available at <https://doi.org/10.5061/dryad.g63c748>.

ECM-specific orthologous gene family calling

We first identified homologous gene families among each of the six genomes using FastOrtho implementing MCL ver. 11.294 (van Dongen 2000) and BLASTp ver. 2.7.1 (Altschul *et al.* 1990) with default parameters. To investigate if the inflation value parameter affected results, we ran FastOrtho five additional times with different inflation values ranging from 1.2 to 6. We defined gene families unique to ECM *Amanita* as families for which homologs are present in all three ECM *Amanita* but not in any of the three asymbiotic fungi. A similar approach was used in Hess *et al.* (2018), but the resulting estimates are different from ours because different parameters were used.

Identifying selection pressures on gene families unique to ECM *Amanita*

We next sought to understand which gene families possess signals of purifying selection ($dN/dS < 1$). The putative protein sequences of all genes from each gene family were first aligned with MAFFT ver. 7.149b (Katoh *et al.* 2002) and then trimmed with trimAl ver. 1.4.rev15 (Capella-Gutiérrez *et al.* 2009) using default parameters. A phylogeny for each gene family was built with RAxML ver. 7.2.8 (Stamatakis 2006), using the trimmed protein alignment, and applying gamma rate heterogeneity and the best substitution model (either JTT, LG or WAG)

as determined by AICc values calculated by ProtTest ver. 3.4 (Darriba *et al.* 2011). The DNA sequences of coding regions (CDS) in each gene family were also aligned based on a codon substitution matrix, using PRANK ver. 140603 (Loytynoja & Goldman 2005). Protein phylogenies and CDS alignments were used to test the alternative hypothesis of dN/dS bias away from neutral selection using the codeml program implemented in PAML ver. 4.8 (model = 0, CodonFreq = 3, fix_kappa = 0, fix_omega = 0 vs 1) (Yang 2007).

The differential expression of gene families unique to ECM *Amanita*

To test if one of the ECM *Amanita*, *A. muscaria*, preferentially expresses genes unique to ECM species during symbiosis, we compared expression patterns of genes in each of three categories: (1) genes conserved across the six species (n=5264), (2) ECM-*Amanita*-unique genes (n=272) and (3) orphan genes (n=4989) (genes only found in *A. muscaria* var. *guessowii*). To assess if the proportion of genes upregulated in symbiosis (=the number of genes upregulated in symbiosis/total number of genes) is higher for ECM-*Amanita*-unique genes compared to conserved or orphan genes, we first identified all genes upregulated in symbiotic cultures. We retrieved the expression read count table generated from both symbiotic root tips and axenic cultures of *A. muscaria* var. *muscaria* from the JGI genome portal (project ID: 1025043) (Kohler *et al.* 2015). Because the expression data (used for training JGI's genome annotation and generating expression table) are taken from *A. muscaria* var. *muscaria* but the genome assembly was generated from *A. muscaria* var. *guessowii*, we mapped our genome annotation (trained with transcriptome from *A. muscaria* var. *guessowii*) to the expression data by finding the best hit of each gene in our genome annotation to the gene sequences used in the expression data with BLASTp (E-value = E-3). A Wald test was performed to screen for differentially expressed

genes using the R package DESeq2 (Love *et al.* 2014). Upregulated genes were defined using FDR adjusted P-values < 0.01 and an expression level fold-change > 2 , 4 or 8. The proportion of upregulated genes was compared across the three categories of gene families using Fisher's exact test and an FDR correction for the P-values. We also performed the same analyses to assess if the genes belonging to any of the three categories mentioned above are upregulated in axenic culture.

Identifying origins of gene families unique to ECM *Amanita*

Overview

Before describing our workflow in greater detail in the sections below, we outline our basic approach: In each family unique to ECM *Amanita*, we selected the longest gene from the *A. muscaria* var. *guessowii* genome as a query to find homologous genes in an in-house, curated proteome database designed to represent the diversity of the three domains of life (see below and Supplementary file 1, 2) (Staehlin *et al.* 2016), using BLAST. We considered genes with no hits as candidates for *de novo* gene birth. Next, to identify HGT events from the remaining genes, we compared gene trees to the species tree to look for incompatibilities. If a gene was found in a monophyletic clade without *Amanita* species, the gene was considered as a candidate for HGT. After excluding candidates for *de novo* gene birth and HGT, we identified potentially duplicated genes by looking for orthologs in asymbiotic species and ECM-paralogs in ECM *Amanita* (we define ECM-paralogs as the paralogs derived from duplications coinciding with the niche transition). If both orthologs and ECM-paralogs were found, the gene was considered as a candidate for duplication. Finally, we attributed genes whose homologs are only absent from

asymbiotic *Amanita* and *Volvariella* to the phenomenon of selective retention (multiple independent deletion events in the asymbiotic lineages but not in ECM *Amanita*).

De novo gene birth detection

To identify putative homologs of the gene families unique to ECM *Amanita*, we curated an in-house genome database with 354 fungal (Supplementary file 1; last accessed: May 2015), 1153 prokaryotic (Staehlin *et al.* 2016), and 88 plant genomes (Supplementary file 2; last accessed: October 2017). Then, we compared the protein sequence representing the longest *A. muscaria* var. *guessowii* gene of each ECM unique gene family to the database using uBLAST implemented in uSearch ver. 8.0.1517 (Edgar 2010). To maximize the probability of finding potentially homologous sequences, we screened using a conservative E-value of E-3 and minimum identity of 0.25. If matches did not return sequences of all three ECM *Amanita* for a given gene probe, results were considered inconclusive and these genes were discarded from the analyses entirely. When results consisted of only the three ECM species, and no other hits, gene families remained in consideration as possible *de novo* gene families. The probe sequences were aligned to the NCBI GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>) and UniProt (<http://www.uniprot.org/>) databases using BLASTp with the same E-value cutoff of E-3 and an additional filter for low complexity regions to check for homologs in either database. Sequences without matches were considered as candidates for *de novo* gene birth.

Next, to explore whether identified gene families are potentially derived from non-coding sequence, we identified the syntenic block of each target gene by matching five upstream and five downstream genes across the three ECM genomes. The same upstream and downstream genes were identified in the three asymbiotic genomes to locate syntenic blocks and putatively

homologous, non-coding sequences. The putatively homologous sequences were aligned to candidate *de novo* genes to explore synteny with MAFFT. We further used HISAT2 Galaxy Version 2.1.0 (Kim *et al.* 2015) with default parameters to map the Illumina transcriptomic raw reads sequenced from mycelia of asymbiotic species cultured in litter (*Amanita*) or potato dextrose broth (*V. volvacea*) (Acc. SRR089758, SRR619832, SRR7694628) (Bao *et al.* 2013; Hess *et al.* 2018) onto the three asymbiotic reference genomes to understand if the homologous regions of *de novo* genes in asymbiotic lineages are expressed.

Horizontal gene transfer detection

After excluding gene families categorized as stemming from *de novo* gene birth as well as gene families for which no strong conclusion could be made, we sought to identify gene families derived from HGT events using the putative homologs in our curated genome database. We first reconstructed a crude phylogeny for each gene to identify potential HGT events. We took uBLAST results and used OrthoMCL ver. 1.4 (Li *et al.* 2003) with an inflation value of 1.5 to cluster the uBLAST results returned for each gene and identify the genes in the same cluster with any target gene, eliminating highly diverged sequences. Refined uBLAST results were aligned with MAFFT and trimmed with trimAl using default parameters. (TrimAl failed to trim Family 12764, resulting an empty alignment, and so this family was not considered further). Preliminary phylogenies of the aligned and trimmed protein sequences were then constructed using FastTree ver. 2.1.7 (Price *et al.* 2010) and compared to the fungal taxonomy (Spatafora *et al.* 2017): genes of putative HGT families should nest within clades unrelated to *Amanita* (e.g. within the Ascomycota).

We next generated more accurate gene phylogenies for downstream analyses only with those gene families tentatively identified as resulting from HGT. Phylogenies were generated using a subset of sequences: for each putative HGT event, we identified branches with bootstrap support of > 90% housing between 100 to 350 sequences and including at least one gene from each of the three ECM *Amanita* species. These datasets were aligned and trimmed again and used to generate new trees using RAxML (Stamatakis 2006) with best evolutionary models identified by ProtTest based on AICc values.

To rigorously reject the null hypothesis of vertical inheritance of genes, we compared our unconstrained trees with vertical-inheritance-constrained trees by using AU tests to identify the best phylogenetic model for each putative HGT family (Shimodaira 2002). We manually constructed the constraint trees by enforcing the null hypotheses of vertical inheritance (either by enforcing monophyletic Agaricales including *Amanita*, or monophyletic putative donor group) in Mesquite (Maddison & Maddison 2017) (Supplementary file 3). Each constraint tree was used to reconstruct a RAxML phylogeny. To test if the unconstrained phylogenies (suggesting HGT) were strongly favored over constrained phylogenies (indicative of divergence in accordance with speciation), we used the per-site likelihood values of both the original unconstrained and constrained phylogenies to perform AU tests implemented using CONSEL ver. 1.2 (Shimodaira & Hasegawa 2001).

Finally, we explored the gene structure (the intron sites) of genes associated with putative HGT events. We hypothesize the structures of putative HGT genes will be more similar to the genes from putative donors than to genes from Agaricales. To compare putative HGT genes with homologous genes from putative donors and homologous genes from Agaricales, we used GenePainter (Hammesfahr *et al.* 2013) to visualize YAML-formatted gene structures generated

with WebScipio (Hatje *et al.* 2011) and protein alignments from MUSCLE ver. 3.8.31 (Edgar 2004).

Gene duplication

We next tested whether remaining gene families (gene families not resulting from either *de novo* gene birth or HGT) might result from duplication and the subsequent rapid evolution of paralogs. We took a two-step approach: first, we looked for orthologs of putative rapidly evolving genes in asymbiotic species; next, we identified ECM-paralogs (paralogs derived from gene duplications coinciding with the niche transition). The closest asymbiotic ortholog to the new gene family served as an outgroup, allowing us to distinguish new gene families from ECM-paralogs among ECM species.

To identify orthologs in asymbiotic species, we generated a more robust phylogeny using RAxML with the uBLAST results generated for the HGT analysis. If the MCL-reduced dataset housed more than 800 sequences, the dataset was further trimmed using the FastTree trimming method described above; if the MCL-reduced dataset contained 30 to 800 sequences, the MCL-reduced dataset was used; if the MCL-reduced dataset consisted of fewer than 30 sequences, the entire dataset was used to reconstruct the phylogeny. To generate phylogenies we used MAFFT, trimAl, ProtTest and RAxML as above. Next, to identify potential orthologs in asymbiotic species, we first rooted each phylogeny using its midpoint and then split each phylogeny (of each gene family) into single gene trees with TreeKO algorithm implemented in ete2 toolkit (Huerta-Cepas *et al.* 2010; Marcet-Houben & Gabaldón 2011). The TreeKO algorithm splits the phylogenies into multiple single gene trees by trimming off branches that represent duplication events until every species only has a single gene in any single gene tree, and

therefore genes in each single gene tree can be treated as orthologs (Marcet-Houben & Gabaldón 2011). We analyzed single gene trees to test if each tree houses (1) only sequences from ECM species or (2) sequences from both ECM and asymbiotic species. Scenario (1) would suggest no orthologous genes can be found in asymbiotic species so selective retention in ECM species (multiple deletions in asymbiotic species) is the more parsimonious explanation of why these families being identified as families unique to ECM *Amanita*. Scenario (2) suggests the presence of orthologs in asymbiotic species.

Finally, to detect clear signals of gene duplications originating with the niche transition, we sought to identify ECM-paralogs associated with the transition to ECM niche. We first identified all nodes between (1) the most recent common ancestor (MRCA) of a given family (determined by FastOrtho), and (2) the MRCA of this family and the most phylogenetically proximate asymbiotic ortholog(s). We then identified ECM *Amanita* homologs diverging from the abovementioned intermediary nodes, and labeled these homologs as ECM-paralogs. If ECM-paralogs are present and no homolog from other species clusters with ECM-paralogs, we consider the origin of the family to be gene duplication and look for a bootstrap value ≥ 80 supporting the ECM-paralogs.

To test if the duplicated genes experienced novel selection pressure, we trimmed off the tips that are not from *Amanita* and *Volvariella* in the phylogeny of each gene family and labeled the branch of duplication as “foreground.” The phylogeny and their CDS alignment from PRANK was then analyzed by aBSREL (Smith *et al.* 2015), implemented in HYPHY ver. 2.5.1 (Kosakovsky Pond *et al.* 2005), to test if the duplication branch (foreground branch) has a proportion of codons that has experienced positive selection ($dN/dS > 1$).

Selective retention in asymbiotic Amanita and Volvariella

Gene families not fitting into any criterion described above were considered as the results of selective retention (multiple deletions in asymbiotic species). To be stringently considered as a gene family that has undergone selective retention, uBLAST results of genes within a given family must not contain any of the three asymbiotic species.

Results

Number and properties of unique gene families

The number of gene families found only in the three ECM *Amanita* genomes ranged from 89 to 120. Changing the parameter settings for FastOrtho greatly impacted the number of gene families identified as unique. The total number of gene clusters increased from 8,694 to 11,436 as the inflation value increased (Supplementary file 4). Because we have no prior knowledge of gene function we decided to use the default inflation value of 1.5, which balances sensitivity and selectivity and fits enzyme family nomenclature according to which reaction an enzyme catalyzes (EC annotation) (Li *et al.* 2003).

Using this inflation value we identified a total of 9,429 gene families and identified 109 gene families as unique to the three ECM *Amanita* genomes (Supplementary file 5). Among the 109 gene families, 107 are undergoing significant purifying selection ($dN/dS < 1$; LRT P-value < 0.05). Of the gene families experiencing purifying selection, values of dN/dS range from 0.00256 to 0.8504 (Figure 2). The dN/dS ratio provides evidence that genes from these gene families encode proteins and are not annotation artifacts. Two gene families had dN/dS ratios close to one suggesting either the genes do not code for functional proteins or the genes are

under neutral selection (in other words, natural selection does not influence the evolutionary trajectory of these genes).

When using a fold-change cutoff of four, a significantly higher proportion of genes in the 109 unique gene families were upregulated in ECM root tips, compared to orphan genes found only in *A. muscaria* and genes conserved across all six species. In addition, a significantly higher proportion of orphan genes were upregulated in ECM root tips, compared to genes in conserved gene families (Figure 3). However, the difference between unique and orphan genes was not significant using a fold-change cutoff of two or eight (Figure 3A and C). The difference between unique and conserved genes was significant regardless of the fold-change cutoff. In axenic cultures, unique and orphan were upregulated compared to conserved families when any fold-change cutoff was applied (Supplementary file 6).

***De novo* gene birth**

Based on uBLAST searches of our curated genome database, six families were identified as candidates for *de novo* gene birth. However, additional BLAST searches in the GenBank and UniProt databases detected putative homologs for four of these families, leaving only two gene families as candidates for *de novo* gene birth (families 1476 and 3446).

For each of the two gene families, each ECM *Amanita* species has only a single gene copy. The two putative *de novo* gene families have dN/dS ratios significantly lower than one (family 1476: 0.174 and family 3446: 0.198; codeml LRT P-value < 10^{-20}), suggesting the genes are experiencing strong natural selection. We returned to the transcriptomic data to probe expression patterns of the genes from these two gene families. The gene representing gene family 1476 is expressed constitutively in both symbiotic and axenic cultures (14 to 26 RPKM in each treatment;

RPKM=reads per kilobase of exon model per million mapped reads). Transcripts of the gene representing the gene family 3446 are detected but are not present at levels greater than 1 RPKM in any treatment. While the evidence suggests these are real genes, there is no evidence for the upregulation of either of the two gene families in ECM root tips.

Neither of the two genes has a known function. The lengths of the proteins are 178 (family 1476) and 297 (family 3446) amino acids. Genes from the two gene families have a GC content of 49.9% and 52.0%, respectively, and these GC contents more closely resemble the CDS of conserved genes (49.4%) compared to intergenic regions (46.0%) although each is presumably derived from an intergenic region (Supplementary file 7). In addition, each gene possesses at least one intron (gene 1476 from *A. muscaria* var. *guessowii* has two introns, but genes in the same family from *A. brunnescens* and *A. polypyramis* have only one intron). Introns are less commonly reported in genes derived from *de novo* gene birth, but there may be a bias because most research focuses on recently birthed genes. Our data suggest the evolutionary history of ECM *Amanita* is long enough that these two gene families acquired introns.

Gene family 1476 is located within a non-syntenic region. We hypothesize that this gene family is located in a relatively variable region. In contrast, genes from family 3446 are located within a conserved region across the three species (Figure 4). We attempted to find the homologous non-coding sequence of these genes by aligning homologous regions from ECM and asymbiotic species. However, the pairwise identities (the proportion of aligned nucleotides) of these multiple sequence alignments were low and ranged from 42.0 to 53.3%. When we searched for evidence of expression of the homologous sequences in asymbiotic species, we found a few raw reads of this region in the transcriptome of all three asymbiotic species: *A. inopinata* (1 read), *A. thiersii* (3 reads) and *V. volvacea* (9 reads) (Supplementary file 8). The low

abundance of transcripts from this region leaves open the question of whether this homologous, presumably non-coding region is actually transcribed by the fungus in nature. While no gene is present in these regions in *A. inopinata* and *A. thiersii* according to their annotations, three genes are annotated in the homologous region from *V. volvacea*, and one gene is responsible for its transcriptomic reads. However, these three genes show no homology with gene family 3446.

Horizontal gene transfer

Comparing the species tree and crude gene phylogenies generated by FastTree, six gene families emerged as candidates for HGT. After building the RAxML gene phylogenies, one gene family was no longer placed with the putative donor lineage and therefore we did not consider it further. Of the remaining five families, AU tests failed to reject vertical inheritance as a possibility for one of the gene families (7854). In families 11987 and 12806, AU tests rejected the two hypotheses which would suggest vertical inheritance: (1) all genes from Agaricales (including *Amanita*) forming a monophyletic group (P-values < 0.05) and (2) the genes from the putative donor forming a monophyletic group (P-values < 0.01); these tests strongly suggest this family derived from HGT. In two other families, only one hypothesis suggesting vertical inheritance was rejected. In family 10418, only the first hypothesis was rejected (P-value = 0.001), whereas in family 2813, the second hypothesis was rejected (P-value = 0.01). However we were unable to test for the first hypothesis in family 2813 (no other Agaricales homologs were found) (Figure 5; Supplementary file 9). In summary, we consider each of these four families (2813, 10418, 11987 and 12806) to be the result of HGT. Gene structures provide additional evidence for the HGT of gene family 12806. In this family the ECM *Amanita* genes

share more intron sites with homologs from putative donors than with homologs from other Agaricales species (Figure 6).

ECM *Amanita* genes in three of the four HGT families form monophyletic groups in the donor lineage. We hypothesize the genes in family 2813 do not form a monophyletic group because an insufficient phylogenetic signal leads to poorly resolved branches: bootstrap values supporting polyphyly range from 3 to 36. Each of the four families clusters inside Eurotiomycetes, or inside Ascomycota but with Eurotiomycetes as the sister group to the HGT genes. However, there is no evidence that the HGT genes are linked in either the donor lineage or *Amanita* genomes so there is no support for a single transfer of the four genes (e.g. as a gene cluster), which would be more parsimonious.

GO terms were assigned to three HGT families, including GO:0008660 (1-aminocyclopropane-1-carboxylate (ACC) deaminase activity), GO:0009310 (amine catabolic process), GO:0030170 (pyridoxal phosphate binding), GO:0003993 (acid monophosphatase activity), GO:0046872 (metal ion binding), GO:0016787 (hydrolase activity), GO:0016788 (esterase activity) (Supplementary file 5). All HGT gene families have a dN/dS < 1 (0.066 to 0.178) and are expressed. Only the gene from family 10418 (an ACC deaminase) is upregulated in symbiotic cultures, and it is 52-fold overexpressed.

Gene duplications

Orthologs of 16 remaining gene families were found in asymbiotic species, and clear evidence of ECM-paralogs is found for four of these 16 families. However, common ancestry with ECM-paralogs is supported by a bootstrap value ≥ 80 for only one family, family 1119 (Figure 7). But only two species (*A. brunnescens* and *A. polypyramis*) possess these

ECM-paralogs, in *A. muscaria* var. *guessowii* there are no family 1119 ECM-paralogs. The closest asymbiotic ortholog of this family is clustered with some but not all of the ECM-paralogs by FastOrtho, which suggests a divergence after duplication. We hypothesized that this family experienced positive selection shortly after gene duplication. However, we failed to reject the null hypothesis of an absence of positive selection (aBSREL LRT P-value = 0.063). In addition, these genes are not located in syntenic regions and so we are unable to test if family 1119 is the more recent ECM-paralog.

This family is transcribed and annotated as a membrane protein (GO:0016020) (Supplementary file 5). In the transcribed product, no signal peptide was predicted by SignalP ver. 5.0 (Almagro Armenteros *et al.* 2019), 10 transmembrane domains were found by TMHMM ver. 2.0 (Krogh *et al.* 2001), and the product was predicted to be located on Golgi apparatus membranes by DeepLoc ver. 1 (Almagro Armenteros *et al.* 2017). HMMER 3.3 (Potter *et al.* 2018) predicted the family belongs to the DUF6 (domain of unknown function 6 or EamA) family. The limited evidence of positive selection and the ability to differentiate the functions of the newly duplicated genes and their paralogs prevents any inference of neofunctionalization following duplication for this family.

Selective retention

Thirty-four gene families unique to ECM *Amanita* appear to be the result of selective retention (equivalent to multiple deletions in asymbiotic lineages). Based on uBLAST and an E-value < E-3, these gene families lack homologs in the three asymbiotic species. False positive signals of selective retention may result from either our choice of E-value cutoff, HGT from unconsidered lineages, or false negatives in annotation. However, our estimate remains

conservative because some selectively retained gene families may house paralogs in asymbiotic species that diverged before the origin of the ECM symbiosis, and these families are not considered here.

Discussion

The origins of gene families found only in symbiotic *Amanita*

New gene families shape genome evolution and can drive adaptation to novel niches (Friesen *et al.* 2006; Milde *et al.* 2009). Among ECM *Amanita*, the genes of unique gene families (gene families found in the three ECM species but not found in closely related asymbiotic fungi) are upregulated in ECM root tips compared to the genes of conserved or orphan gene families, suggesting the new gene families acquired during the niche transition function during symbiosis. Most of the genes of unique gene families have also undergone strong natural selection.

We discovered evidence for the precise origins of 41 families. Thirty-four families are inferred as unique because of selective retention, and the other seven appear as truly new gene families, derived from either *de novo* gene birth (two), HGT (four) or divergence after gene duplication (one). The low number of *de novo* gene families is not surprising. The high turnover rate of *de novo* genes in any genome results in a low preservation of *de novo* families (Palmieri *et al.* 2014). Although the mechanism(s) driving HGT in fungi remain elusive, accumulating evidence suggests HGT is a key to evolutionary innovation in the fungal kingdom (Soanes & Richards 2014). Finding four gene families derived from HGT suggests HGT was also critical to the changes in *Amanita* ecology, but this discovery might also reflect the better preservation of the phylogenetic signal of HGT over time, compared to signals from *de novo* gene birth or gene duplication. Unexpectedly, using our workflow we only identified one new gene family derived

from gene duplication, contrary to expectations that gene duplication is a common source of new genes in fungi and other eukaryotes (Ohno 1970; Zhang 2003; Wisecaver *et al.* 2014). The discrepancy may result from the use of a Markov clustering algorithm (MCL) to identify new families, as the algorithm can consolidate paralogs into a single gene family if they have not undergone rapid divergence (Li *et al.* 2003), but the discrepancy may also result from a lack of phylogenetic support for the monophyly of paralogs. Changing the parameters used with MCL may influence how clusters are identified (Li *et al.* 2003). For example, while we identified 109 families as unique to ECM genomes Hess *et al.* (2018) identified 171 unique gene families. We used no match cutoff or identity cutoff and used an inflation value of 1.5, while Hess *et al.* (2018) used a match cutoff of 60%, an identity cutoff of 30%, and an inflation value of 3. Parameter choice involves a balance: choosing more stringent clustering results in greater numbers of clusters (Hess *et al.* 2018) and may lead to the identification of greater numbers of duplicated gene families. However, whether these gene families are functionally diverged enough to be judged as novel is an open question, and choosing higher cluster tightness can break orthologs into different families (Li *et al.* 2003). In summary, the number of gene origins inferred reflects the number of gene family birth events, rates of gene turnover, the decay of phylogenetic signal and the choice of clustering algorithm parameters.

Research on the origins of new genes has focused on orphan genes found in single species and not gene families found across closely related species. Most of these orphan genes are more recent than the niche transition in *Amanita* (the origin of symbiosis among *Amanita* fungi dates to 80 million years ago (Varga *et al.* 2019)). By focusing on younger genes, these studies take advantage of opportunities to trace homologies and syntenies to elucidate the molecular

mechanisms of the emergence of new genes (Donoghue *et al.* 2011; Arendsee *et al.* 2019).

Moreover, HGT and duplication events are generally left out of their foci.

Other studies on gene family origins have used two major approaches to identify different evolutionary events. The first approach estimates the gene turnover rate with the gene counts of a family in different species, typically using gain-and-death (GD) or birth-death-and-innovation (BDI) models (Librado *et al.* 2012). The second approach uses a species tree/gene tree reconciliation method to identify duplication, gene loss and HGT (the DTL scenario) (Bansal *et al.* 2012; Jacox *et al.* 2016); other algorithms also incorporate incomplete lineage sorting (DTLI) (Stolzer *et al.* 2012). Because innovation events in the BDI model account for both *de novo* gene birth and HGT (Karev *et al.* 2002), this first approach does not provide information on the different mechanisms mediating the origins of new gene families. On the other hand, the second approach does not account for *de novo* gene birth (Tofigh *et al.* 2011). In addition, methods using the DTL scenario require a well-curated species tree for not only the species of interest but also the putative donors for HGT (Bansal *et al.* 2012; Jacox *et al.* 2016) and these are not always available. To account for *de novo* gene birth, HGT and gene duplication simultaneously, and to avoid a reliance on a well-curated species tree, our workflow includes (1) gene clustering to identify new gene families, (2) the search for homologs in a curated genome database, (3) phylogenetic analyses designed for different origin hypotheses, followed by (4) integration of additional support from analyses of e.g. gene structure, composition, and synteny (Gluck-Thaler & Slot 2015).

***De novo* gene families have lost some properties of *de novo* genes**

Interest in *de novo* gene birth is growing. However, misidentification of *de novo* genes can result from either the rapid evolution of an extant gene or a poorly resolved comparison between a putative *de novo* gene and an incomplete genome database (Moyers & Zhang 2015). Casola (2018) suggests investigating four gene features to avoid misidentification and enable recognition of real *de novo* gene families: the absence of homologs in other taxa, a lack of conserved domains, conserved synteny, and substitutions enabling genes encoding proteins (e.g. generating a start codon). We successfully identified the first three features for at least one family (3446). We were not able to detect the last feature (substitutions enabling genes encoding proteins) because of the dissimilarity between the putative *de novo* genes and their homologous non-coding sequences. Because *de novo* gene families can evolve from sequences encoding long non-coding RNAs (Schlötterer 2015), we searched for transcripts of non-coding sequences orthologous to putative *de novo* gene families in asymbiotic species. However, the low number of raw reads we discovered prevents firm conclusions as to whether or not these regions are transcribed by the fungi in nature.

The genes of *de novo* gene families are reported to possess several distinctive characteristics, including short gene lengths (around 300 to 400 bp) and few to no introns (Wu & Knudson 2018). Sequences may be GC poor or have GC content similar to conserved genes, depending on the species (Palmieri *et al.* 2014; Wu & Knudson 2018). However, the two *de novo* gene families we identified have a structure and composition consistent with the conserved gene families in *Amanita* genomes. Combined with the absence of similarities between *de novo* gene families and their homologous non-coding sequences, these observations lead us to conclude that these gene families have had sufficient time to become ameliorated and now resemble other coding genes (Marri & Golding 2008).

De novo gene families of animals and plants have low gene expression levels and higher tissue-specific expression compared to older genes (Schlötterer 2015), especially genes expressed in animal testes (Begun *et al.* 2007). One of the two *de novo* families we identified has a higher expression rate in axenic mycelium compared to ECM root tips whereas the other is not differentially expressed so these genes are unusual compared to the genes of the other 109 gene families. We hypothesize that these gene families are not directly involved in the plant-fungal interaction, but with other processes shared among ECM *Amanita* species.

Genes horizontally transferred to ECM *Amanitas* are from Ascomycota, and we hypothesize ACC deaminase (family 10418) was directly involved in the transition to symbiosis in *Amanita*

As with identification of *de novo* gene birth, strategies to identify HGT using only similarity as a criterion are also problematic (Guindon & Perrière 2001). Keeling and Palmer (2008) have suggested gene-species phylogenetic incongruence as the gold standard for HGT detection. Based on that gold standard, we identified four unique gene families originating from HGT. In two families, we successfully rejected the null hypotheses that (1) genes from Agaricales and *Amanita* are monophyletic and (2) the genes from donor group itself are monophyletic, but we failed to reject one of the null hypotheses for the other two families. The failure to reject all null hypotheses may be caused by accumulated substitutions in transferred genes, or too few taxa in our curated genome database. In addition, one gene family showed a high similarity of exon/intron structure to its putative donor, providing further evidence for HGT. These four gene families are not the first record of HGT to ECM *Amanita*. We previously reported that genes of carbohydrate esterases family 1 (CE1s) were transferred to ECM *Amanita* from bacteria (Chaib

De Mares *et al.* 2015), but these genes are not found in *A. polypyraxis* and thus do not fit our current definition of gene families unique to ECM *Amanita*.

Interestingly, all four HGT families are inferred to have been transferred from Ascomycota and specifically from Eurotiomycetes. Multiple HGTs to a single lineage from the same donor have been reported before and are usually enabled by their transfer as a gene cluster or syntenic block (Slot 2017), but "highways" of HGT between lineages have also been inferred (Qiu *et al.* 2016). In the case of HGT to ECM *Amanita*, no evidence of physical linkage in either *Amanita* or the putative donors was found. The lack of synteny suggests either these genes were transferred independently or the genes migrated into different genomic locations after HGT.

HGT genes may facilitate adaptation to new niches among the fungi (Soanes & Richards 2014). Of the four gene families transferred to ECM *Amanita*, we consider the family encoding ACC deaminase as the best candidate for driving niche transition because of its expression patterns and putative function. The gene from this family is upregulated in ECM root tips more than 52-fold, which is the highest fold difference among all genes from the 109 gene families. ACC deaminase can remove the amine group from an ACC and produce α -ketobutyrate (Honma & Smmomura 1978). ACC is the immediate precursor of ethylene, and ACC deaminase can therefore inhibit the ethylene signaling pathway, a negative regulatory pathway of ECM symbioses (Plett, Khachane *et al.* 2014). In fact, the ACC deaminase knockouts of bacteria involved in a similar mutualistic system, nitrogen fixing rhizobia, are less capable of nodulation compared to the wild types (Nascimento *et al.* 2016). In addition, in arbuscular mycorrhizal fungi, the *SP7* gene also inhibits the ethylene signaling pathway (Kloppholz *et al.* 2011). We hypothesize the ACC deaminases of ECM *Amanita* reduce the concentration of ACC and ethylene in ECM roots during symbiosis, and the reduction of ethylene in ECM root tips enables

the lateral branching of ECM roots and formation of the Hartig net. Moreover, our finding may provide new evidence of a molecular convergence among mycorrhizal and root-nodulating associations. Lastly, we note ACC deaminase genes are also found in other ECM *Amanita* genomes not included in our analyses, for example *A. bisporigera*, *A. phalloides* and *A. jacksonii* etc. (van der Nest *et al.* 2014; Pulman *et al.* 2016).

No significant evidence for positive selection on newly duplicated genes

Many established algorithms are available to detect gene duplication. However, because we are specifically interested in potential duplication events coinciding with niche transition, we chose to detect the orthologs and ECM-paralogs of each of our identified unique gene families. Using this strategy, the only family duplicated during niche transition has a closest ortholog in *A. thiersii* and ECM-paralogs in *A. brunnescens* and *A. polypyramis*. Multiple evolutionary scenarios can explain the phylogenetic topology (e.g. duplication before the MRCA of *A. inopinata* and ECM *Amanita*, followed by deletion of both homologs in *A. inopinata*), but a single deletion in *A. inopinata* and duplication along the branch leading to niche transition is the most parsimonious explanation. We are also able to detect homologs in *A. inopinata* and *V. volvacea*, but these homologs are from diverged phylogenetic clusters, suggesting the family possesses a dynamic evolutionary background.

Gene duplication provides functional redundancy and paralogs often experience novel selection pressure, undergoing neofunctionalization (Saad *et al.* 2018). An extreme case suggests asymmetric evolutionary rates between two paralogs (or ohnologs to be precise) in yeast (Byrne & Wolfe 2007). However, in the new gene family derived from duplication in ECM *Amanita*, we failed to detect signals of positive selection along the branch leading to the new gene family

cluster. We hypothesize synonymous substitutions have reached saturation at selected sites and the substitutions have removed the trace of positive selection (Gharib & Robinson-Rechavi 2013). However, it is also possible the family emerged as a result of non-selective events.

The gene family stemming from a recent duplication encodes two DUF6 domains in the form of a 5 + 5 transmembrane protein (each “5” is one DUF6 domain). Many proteins with this configuration are transporters, e.g., O-acetylserine/cysteine export proteins and nucleotide sugar transporters (Västermark *et al.* 2011), but at least one gene, *PecM*, is involved in the degradation of pectin and cellulose (Rouanet & Nasser 2001). Although this gene family was identified as a new family, conserved in ECM *Amanita*, there is no evidence for its differential expression. We hypothesize that if this family is not solely the result of stochastic evolution, this gene family could be involved in transmembrane transport, substrate degradation by absorptive hyphae or controlled post-transcriptionally.

Conclusion

We developed a workflow for identifying the origins of new gene families unique to symbiotic fungi and not found in closely related free-living fungi. Among the 109 new gene families present in ECM *Amanita*, two, four and one families appear derived from *de novo* gene birth, HGT and gene duplication, respectively, but 34 families only appear new due to selective retention in symbiotic species. The genes of gene families unique to ECM *Amanita* are upregulated during symbiosis and are likely functionally relevant to the symbiosis. The horizontally transferred gene encoding ACC deaminase is potentially crucial to the mutualistic relationship, possibly regulating the immune response in plant symbionts. Our findings suggest a

new possibility for ECM evolution: the transition to ECM niche in fungi is not only driven by gene loss but by coincident new gene acquisition as well.

Data availability: Scripts and additional supplementary data are deposited in Dryad under accession doi:10.5061/dryad.s4mw6m94f.

Acknowledgments

This work was supported by the National Science Foundation [grant number DEB-1638999]; and the University of Wisconsin-Madison. Computational work was performed at UW-Madison's Center for High Throughput Computing and at the Ohio Supercomputer Center. We thank Jacob Golan for assistance with writing.

References

- Almagro Armenteros JJ *et al.* 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* 37:420–423. doi: 10.1038/s41587-019-0036-z.
- Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. 2017. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics.* 33:3387–3395. doi: 10.1093/bioinformatics/btx431.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410. doi: 10.1016/S0022-2836(05)80360-2.
- Andersson DI, Jerlström-Hultqvist J, Nasväll J. 2015. Evolution of new functions *de novo* and from preexisting genes. *Cold Spring Harb. Perspect. Biol.* 7:a017996. doi: 10.1101/cshperspect.a017996.

660 Arendsee Z *et al.* 2019. Fagin: synteny-based phylostratigraphy and finer classification of young
 661 genes. BMC Bioinformatics. 20:440. doi: 10.1186/s12859-019-3023-y.
 662 Assis R, Bachtrog D. 2013. Neofunctionalization of young duplicate genes in *Drosophila*. Proc.
 663 Natl. Acad. Sci. 110:17409–17414. doi: 10.1073/pnas.1313759110.
 664 Bansal MS, Alm EJ, Kellis M. 2012. Efficient algorithms for the reconciliation problem with
 665 gene duplication, horizontal transfer and loss. Bioinformatics. 28:i283–i291. doi:
 666 10.1093/bioinformatics/bts225.
 667 Bao D *et al.* 2013. Sequencing and comparative analysis of the straw mushroom (*Volvariella*
 668 *volvacea*) genome. PLoS One. 8: e58294. doi: 10.1371/journal.pone.0058294.
 669 Begun DJ, Lindfors HA, Kern AD, Jones CD. 2007. Evidence for *de novo* evolution of
 670 testis-expressed genes in the *Drosophila yakuba*/*Drosophila erecta* clade. Genetics.
 671 176:1131–1137. doi: 10.1534/genetics.106.069245.
 672 Bittleston LS, Pierce NE, Ellison AM, Pringle A. 2016. Convergence in multispecies interactions.
 673 Trends Ecol. Evol. 31:269–280. doi: 10.1016/j.tree.2016.01.006.
 674 Byrne KP, Wolfe KH. 2007. Consistent patterns of rate asymmetry and gene loss indicate
 675 widespread neofunctionalization of yeast genes after whole-genome duplication. Genetics.
 676 175:1341–1350. doi: 10.1534/genetics.106.066951.
 677 Cai J, Zhao R, Jiang H, Wang W. 2008. *De novo* origination of a new protein-coding gene in
 678 *Saccharomyces cerevisiae*. Genetics. 179:487–496. doi: 10.1534/genetics.107.084491.
 679 Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. TrimAl: a tool for automated
 680 alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 25:1972–1973. doi:
 681 10.1093/bioinformatics/btp348.

682 Carvunis AR *et al.* 2012. Proto-genes and *de novo* gene birth. *Nature*. 487:370–374. doi:
683 10.1038/nature11184.

684 Casola C. 2018. From *de novo* to “*de novo*”: the majority of novel protein-coding genes
685 identified with phylostratigraphy are old genes or recent duplicates. *Genome Biol. Evol.*
686 10:2906–2918. doi: 10.1093/gbe/evy231.

687 Chaib de Mares M *et al.* 2015. Horizontal transfer of carbohydrate metabolism genes into
688 ectomycorrhizal *Amanita*. *New Phytol.* 205:1552–1564. doi: 10.1111/nph.13140.

689 Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models
690 of protein evolution. *Bioinformatics*. 27:1164–1165.

691 van Dongen S. 2000. Graph clustering by flow simulation. doi: 10.1016/j.cosrev.2007.05.001.

692 Donoghue MT, Keshavaiah C, Swamidatta SH, Spillane C. 2011. Evolutionary origins of
693 Brassicaceae specific genes in *Arabidopsis thaliana*. *BMC Evol. Biol.* 11:47. doi:
694 10.1186/1471-2148-11-47.

695 Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high
696 throughput. *Nucleic Acids Res.* 32:1792–1797. doi: 10.1093/nar/gkh340.

697 Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*.
698 26:2460–2461. doi: 10.1093/bioinformatics/btq461.

699 Friesen TL *et al.* 2006. Emergence of a new disease as a result of interspecific virulence gene
700 transfer. *Nat. Genet.* 38:953–956. doi: 10.1038/ng1839.

701 Gharib WH, Robinson-Rechavi M. 2013. The branch-site test of positive selection is surprisingly
702 robust but lacks power under synonymous substitution saturation and variation in GC. *Mol. Biol.*
703 *Evol.* 30:1675–1686. doi: 10.1093/molbev/mst062.

704 Gluck-Thaler E, Slot JC. 2015. Dimensions of horizontal gene transfer in eukaryotic microbial
705 pathogens. PLoS Pathog. 11:e1005156. doi: 10.1371/journal.ppat.1005156.

706 Guindon S, Perrière G. 2001. Intragenomic base content variation is a potential source of biases
707 when searching for horizontally transferred genes. Mol. Biol. Evol. 18:1838–1840. doi:
708 10.1093/oxfordjournals.molbev.a003972.

709 Hammesfahr B, Odrionitz F, Mühlhausen S, Waack S, Kollmar M. 2013. GenePainter: a fast tool
710 for aligning gene structures of eukaryotic protein families, visualizing the alignments and
711 mapping gene structures onto protein structures. BMC Bioinformatics. 14:77. doi:
712 10.1186/1471-2105-14-77.

713 Hatje K *et al.* 2011. Cross-species protein sequence and gene structure prediction with fine-tuned
714 Webscipio 2.0 and Scipio. BMC Res. Notes. 4:265. doi: 10.1186/1756-0500-4-265.

715 Hess J *et al.* 2018. Rapid divergence of genome architectures following the origin of an
716 ectomycorrhizal symbiosis in the genus *Amanita*. Mol. Biol. Evol. 35:2786–2804. doi:
717 10.1093/molbev/msy179.

718 Hess J, Pringle A. 2014. The natural histories of species and their genomes: asymbiotic and
719 ectomycorrhizal *Amanita* fungi. Adv. Bot. Res. 70:235–257. doi:
720 10.1016/B978-0-12-397940-7.00008-2.

721 Honma M, Smmomura T. 1978. Metabolism of 1-aminocyclopropane-1-carboxylic acid. Agric.
722 Biol. Chem. 42:1825–1831. doi: 10.1080/00021369.1978.10863261.

723 Huerta-Cepas J, Dopazo J, Gabaldón T. 2010. ETE: a python environment for tree exploration.
724 BMC Bioinformatics. 11:24. doi: 10.1186/1471-2105-11-24.

725 Husnik F, McCutcheon JP. 2018. Functional horizontal gene transfer from bacteria to eukaryotes.
726 Nat. Rev. Microbiol. 16:67–79. doi: 10.1038/nrmicro.2017.137.

727 Jacox E, Chauve C, Szöllosi GJ, Ponty Y, Scornavacca C. 2016. EcceTERA: comprehensive
 728 gene tree-species tree reconciliation using parsimony. *Bioinformatics*. 32:2056–2058. doi:
 729 10.1093/bioinformatics/btw105.

730 Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS, Koonin E V. 2002. Birth and death of
 731 protein domains: a simple model of evolution explains power law behavior. *BMC Evol. Biol.*
 732 2:18. doi: 10.1186/1471-2148-2-18.

733 Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple
 734 sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30:3059–3066. doi:
 735 10.1093/nar/gkf436.

736 Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.*
 737 9:605–618. doi: 10.1038/nrg2386.

738 Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory
 739 requirements. *Nat. Methods*. 12:357–360. doi: 10.1038/nmeth.3317.

740 Kloppeholz S, Kuhn H, Requena N. 2011. A secreted fungal effector of *Glomus intraradices*
 741 promotes symbiotic biotrophy. *Curr. Biol.* 21:1204–1209. doi: 10.1016/j.cub.2011.06.044.

742 Knowles DG, McLysaght A. 2009. Recent *de novo* origin of human protein-coding genes.
 743 *Genome Res.* 19:1752–1759. doi: 10.1101/gr.095026.109.

744 Kohler A *et al.* 2015. Convergent losses of decay mechanisms and rapid turnover of symbiosis
 745 genes in mycorrhizal mutualists. *Nat. Genet.* 47:410–415. doi: 10.1038/ng.3223.

746 Kosakovsky Pond SL, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies.
 747 *Bioinformatics*. 5:676–679. doi: 10.1093/bioinformatics/bti079.

748 Krogh A, Larsson B, Von Heijne G, Sonnhammer ELL. 2001. Predicting transmembrane protein
 749 topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*
 750 305:567–580. doi: 10.1006/jmbi.2000.4315.

751 Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic
 752 genomes. *Genome Res.* 13:2178–2189. doi: 10.1101/gr.1224503.

753 Librado P, Vieira FG, Rozas J. 2012. BadiRate: estimating family turnover rates by
 754 likelihood-based methods. *Bioinformatics.* 28:279–281. doi: 10.1093/bioinformatics/btr623.

755 Long M, VanKuren NW, Chen S, Vibranovski MD. 2013. New gene evolution: little did we
 756 know. *Annu. Rev. Genet.* 47:307–333. doi: 10.1146/annurev-genet-111212-133301.

757 Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for
 758 RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8.

759 Loytynoja A, Goldman N. 2005. From the cover: an algorithm for progressive multiple
 760 alignment of sequences with insertions. *Proc. Natl. Acad. Sci.* 102:10557–10562. doi:
 761 10.1073/pnas.0409137102.

762 Maddison WP, Maddison DR. 2017. Mesquite: a modular system for evolutionary analysis.
 763 Version 1.0. doi: <http://mesquiteproject.org>.

764 Marcet-Houben M, Gabaldón T. 2011. TreeKO: a duplication-aware algorithm for the
 765 comparison of phylogenetic trees. *Nucleic Acids Res.* 39:e66. doi: 10.1093/nar/gkr087.

766 Marri PR, Golding GB. 2008. Gene amelioration demonstrated: the journey of nascent genes in
 767 bacteria. *Genome.* 51:164–168. doi: 10.1139/G07-105.

768 Matheny PB *et al.* 2006. Major clades of Agaricales: a multilocus phylogenetic overview.
 769 *Mycologia.* 98:982–995. doi: 10.3852/mycologia.98.6.982.

770 Mayr E. 1963. *Animal species and evolution*. Cambridge (MA): Harvard University Press. doi:
 771 10.2307/1440881.

772 Milde S *et al.* 2009. Characterization of taxonomically restricted genes in a phylum-restricted
 773 cell type. *Genome Biol.* 10:R8. doi: 10.1186/gb-2009-10-1-r8.

774 Moran NA, Jarvik T. 2010. Lateral transfer of genes from fungi underlies carotenoid production
 775 in aphids. *Science.* 328:624–627. doi: 10.1126/science.1187113.

776 Moyers BA, Zhang J. 2015. Phylostratigraphic bias creates spurious patterns of genome
 777 evolution. *Mol. Biol. Evol.* 32:258–267. doi: 10.1093/molbev/msu286.

778 Murat C *et al.* 2018. Pezizomycetes genomes reveal the molecular basis of ectomycorrhizal
 779 truffle lifestyle. *Nat. Ecol. Evol.* 2:1956–1965. doi: 10.1038/s41559-018-0710-4.

780 Nascimento FX, Brígido C, Glick BR, Rossi MJ. 2016. The role of rhizobial ACC deaminase in
 781 the nodulation process of leguminous plants. *Int. J. Agron.* 2016. doi: 10.1155/2016/1369472.

782 van der Nest MA *et al.* 2014. Draft genomes of *Amanita jacksonii*, *Ceratocystis albifundus*,
 783 *Fusarium circinatum*, *Huntia omanensis*, *Leptographium procerum*, *Rutstroemia sydowniana*,
 784 and *Sclerotinia echinophila*. *IMA Fungus* 5:472-486. doi: 10.5598/imafungus.2014.05.02.11.

785 Ohno S. 1970. *Evolution by gene duplication*. New York: Springer-Verlag. doi:
 786 10.1007/978-3-642-86659-3.

787 Palmieri N, Kosiol C, Schlötterer C. 2014. The life cycle of *Drosophila* orphan genes. *Elife.*
 788 3:e01311. doi: 10.7554/eLife.01311.001.

789 Peter M *et al.* 2016. Ectomycorrhizal ecology is imprinted in the genome of the dominant
 790 symbiotic fungus *Cenococcum geophilum*. *Nat. Commun.* 7:12662. doi: 10.1038/ncomms12662.

791 Pigliucci M. 2008. What, if anything, is an evolutionary novelty? *Philos. Sci.* 75:887–898. doi:
 792 10.1086/594532.

793 Plett JM, Daguerre Y, *et al.* 2014. Effector MiSSP7 of the mutualistic fungus *Laccaria bicolor*
794 stabilizes the *Populus* JAZ6 protein and represses jasmonic acid (JA) responsive genes. Proc.
795 Natl. Acad. Sci. 111:8299–8304. doi: 10.1073/pnas.1322671111.

796 Plett JM, Khachane A, *et al.* 2014. Ethylene and jasmonic acid act as negative modulators during
797 mutualistic symbiosis between *Laccaria bicolor* and *Populus* roots. New Phytol. 202:270–286.
798 doi: 10.1111/nph.12655.

799 Potter SC *et al.* 2018. HMMER web server: 2018 update. Nucleic Acids Res. 46:W200–W204.
800 doi: 10.1093/nar/gky448.

801 Pulman JA, Childs KL, Sgambelluri RM, Walton JD. 2016 Expansion and diversification of the
802 MSDIN family of cyclic peptide genes in the poisonous agarics *Amanita phalloides* and *A.*
803 *bisporigera*. BMC Genomics 17:1038. doi: 10.1186/s12864-016-3378-7

804 Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 - approximately maximum-likelihood trees for
805 large alignments. PLoS One. 5:e9490. doi: 10.1371/journal.pone.0009490.

806 Qiu H, Cai G, Luo J, Bhattacharya D, Zhang N. 2016. Extensive horizontal gene transfers
807 between plant pathogenic fungi. BMC Biol. 14:41. doi: 10.1186/s12915-016-0264-3.

808 Rouanet C, Nasser W. 2001. The PecM protein of the phytopathogenic bacterium *Erwinia*
809 *chrysanthemi*, membrane topology and possible involvement in the efflux of the blue pigment
810 indigoidine. J. Mol. Microbiol. Biotechnol. 3:309–318.

811 Ruiz-Orera J, Verdaguer-Grau P, Villanueva-Cañas JL, Messeguer X, Albà MM. 2018.
812 Translation of neutrally evolving peptides provides a basis for *de novo* gene evolution. Nat. Ecol.
813 Evol. 2:890–896. doi: 10.1038/s41559-018-0506-6.

814 Saad R, Cohan AB, Kosloff M, Privman E. 2018. Neofunctionalization in ligand binding sites
815 of ant olfactory receptors. Genome Biol. Evol. 10:2490–2500. doi: 10.1093/gbe/evy131.

816 Schlötterer C. 2015. Genes from scratch - the evolutionary fate of *de novo* genes. Trends Genet.
817 31:215–219. doi: 10.1016/j.tig.2015.02.007.

818 Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol.
819 51:492–508. doi: 10.1080/10635150290069913.

820 Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree
821 selection. Bioinformatics. 17:1246–1247. doi: 10.1093/bioinformatics/17.12.1246.

822 Slot JC. 2017. Fungal gene cluster diversity and evolution. Adv. Genet. 100:141–178. doi:
823 10.1016/bs.adgen.2017.09.005.

824 Smith MD *et al.* 2015. Less is more: an adaptive branch-site random effects model for efficient
825 detection of episodic diversifying selection. Mol. Biol. Evol. 32:1342–1353. doi:
826 10.1093/molbev/msv022.

827 Smith S, Read D. 2008. *Mycorrhizal symbiosis*. 3rd ed. Cambridge (MA): Academic Press. doi:
828 10.1016/B978-0-12-370526-6.X5001-6.

829 Soanes D, Richards TA. 2014. Horizontal gene transfer in eukaryotic plant pathogens. Annu.
830 Rev. Phytopathol. 52:583–614. doi: 10.1146/annurev-phyto-102313-050127.

831 Soshnikova N, Dewaele R, Janvier P, Krumlauf R, Duboule D. 2013. Duplications of Hox gene
832 clusters and the emergence of vertebrates. Dev. Biol. 378:194–199. doi:
833 10.1016/j.ydbio.2013.03.004.

834 Spatafora JW *et al.* 2017. The fungal tree of life: from molecular systematics to genome-scale
835 phylogenies. Microbiol. Spectr. 5:5. doi: 10.1128/microbiolspec.FUNK-0053-2016.

836 Staehlin BM, Gibbons JG, Rokas A, O'Halloran TV, Slot JC. 2016. Evolution of a heavy metal
837 homeostasis/resistance island reflects increasing copper stress in enterobacteria. Genome Biol.
838 Evol. 8:811–826. doi: 10.1093/gbe/evw031.

839 Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with
840 thousands of taxa and mixed models. *Bioinformatics*. 22:2688–2690. doi:
841 10.1093/bioinformatics/btl446.

842 Stolzer M *et al.* 2012. Inferring duplications, losses, transfers and incomplete lineage sorting
843 with nonbinary species trees. *Bioinformatics*. 28:i409–i415. doi: 10.1093/bioinformatics/bts386.

844 Tedersoo L, May TW, Smith ME. 2010. Ectomycorrhizal lifestyle in fungi: global diversity,
845 distribution, and evolution of phylogenetic lineages. *Mycorrhiza*. 20:217–263. doi:
846 10.1007/s00572-009-0274-x.

847 Tofigh A, Hallett M, Lagergren J. 2011. Simultaneous identification of duplications and lateral
848 gene transfers. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 8:517–535. doi:
849 10.1109/TCBB.2010.14.

850 Vakirlis N *et al.* 2018. A molecular portrait of *de novo* genes in yeasts. *Mol. Biol. Evol.*
851 35:631–645. doi: 10.1093/molbev/msx315.

852 Varga T *et al.* 2019. Megaphylogeny resolves global patterns of mushroom evolution. *Nat. Ecol.*
853 *Evol.* 3:668–678. doi: 10.1038/s41559-019-0834-1.

854 Västermark Å, Almén MS, Simmen MW, Fredriksson R, Schiöth HB. 2011. Functional
855 specialization in nucleotide sugar transporters occurred through differentiation of the gene
856 cluster EamA (DUF6) before the radiation of Viridiplantae. *BMC Evol. Biol.* 11:123. doi:
857 10.1186/1471-2148-11-123.

858 Villanueva-Cañas JL *et al.* 2017. New genes and functional innovation in mammals. *Genome*
859 *Biol. Evol.* 9:1886–1900. doi: 10.1093/gbe/evx136.

860 Wisecaver JH, Slot JC, Rokas A. 2014. The evolution of fungal metabolic pathways. *PLoS*
861 *Genet.* 10:e1004816. doi: 10.1371/journal.pgen.1004816.

862 Wolfe BE, Tulloss RE, Pringle A. 2012. The irreversible loss of a decomposition pathway marks
863 the single origin of an ectomycorrhizal symbiosis. PLoS One. 7:e39597. doi:
864 10.1371/journal.pone.0039597.

865 Wu B, Knudson A. 2018. Tracing the *de novo* origin of protein-coding genes in yeast. MBio. 9:4.
866 doi: 10.1128/mBio.01024-18.

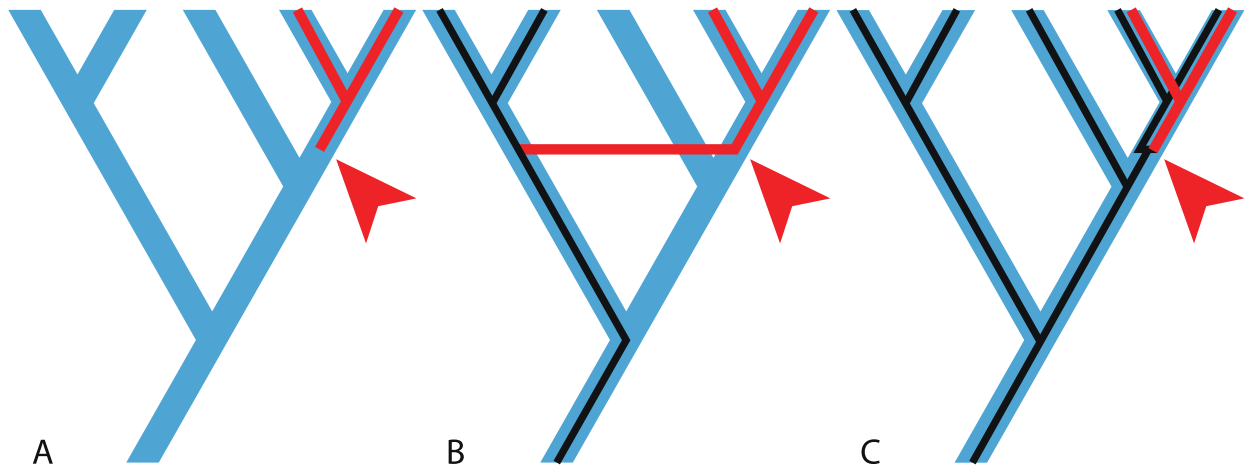
867 Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol.
868 24:1586–1591. doi: 10.1093/molbev/msm088.

869 Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol. Evol. 18:292–298. doi:
870 10.1016/S0169-5347(03)00033-8.

871

872

873 Figures



874

875 Fig. 1. Three mechanisms of gene family emergence. A: *de novo* gene birth; B: HGT; C: gene

876 duplication. Blue: species tree; Red: new gene families (arrow indicates origin); Black:

877 homologous genes.

878

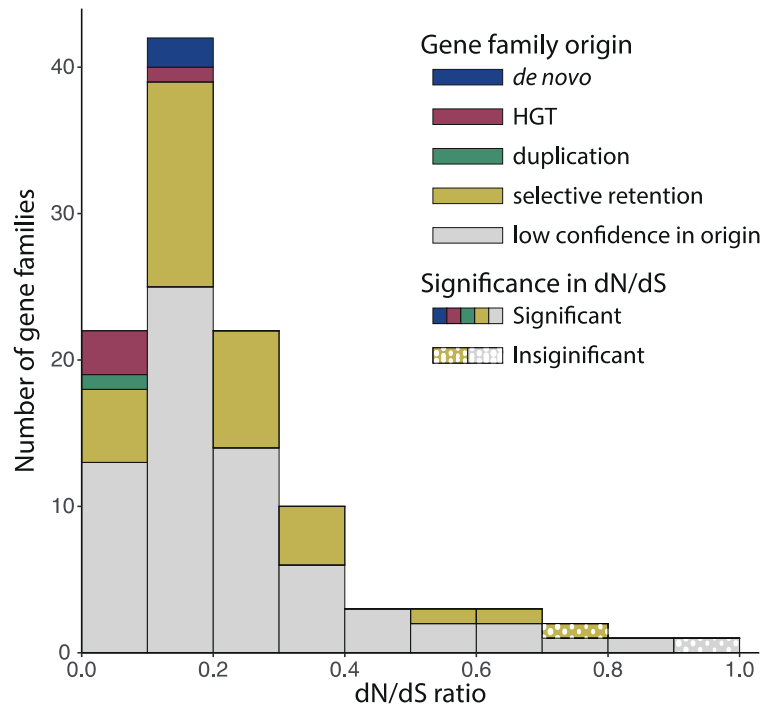


Fig. 2. Frequency distribution of the dN/dS ratios of the 109 gene families unique to ECM *Amanita*. Most gene families (98.2%) have experienced purifying selection with dN/dS ratios significantly lower than one (P-value < 0.05).

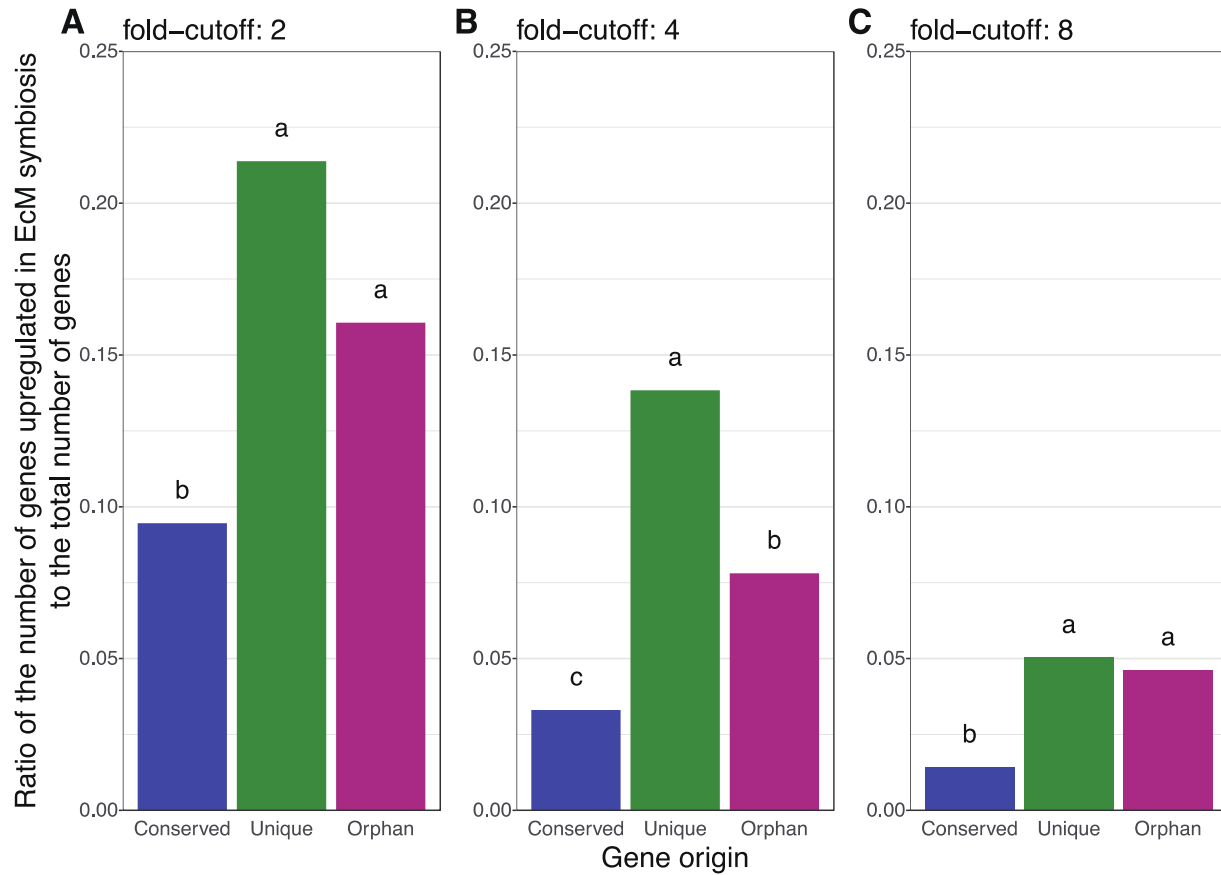


Fig. 3. The ratio of the number of genes upregulated in ECM symbiosis to the total number of genes for conserved (n=5264), unique (n=272) or orphan families (n=4989). Different fold cutoffs (two, four and eight) are used to define upregulation in subplots A, B and C. Fisher exact test P-value < 10^{-4} for all cutoffs. Letters mark significant (adjusted P-value < 0.05).

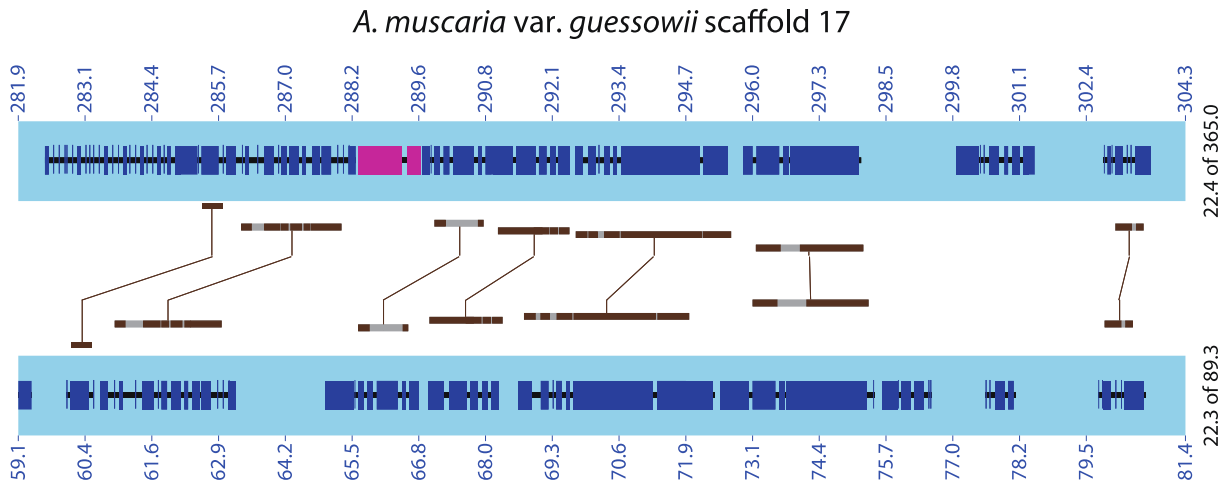
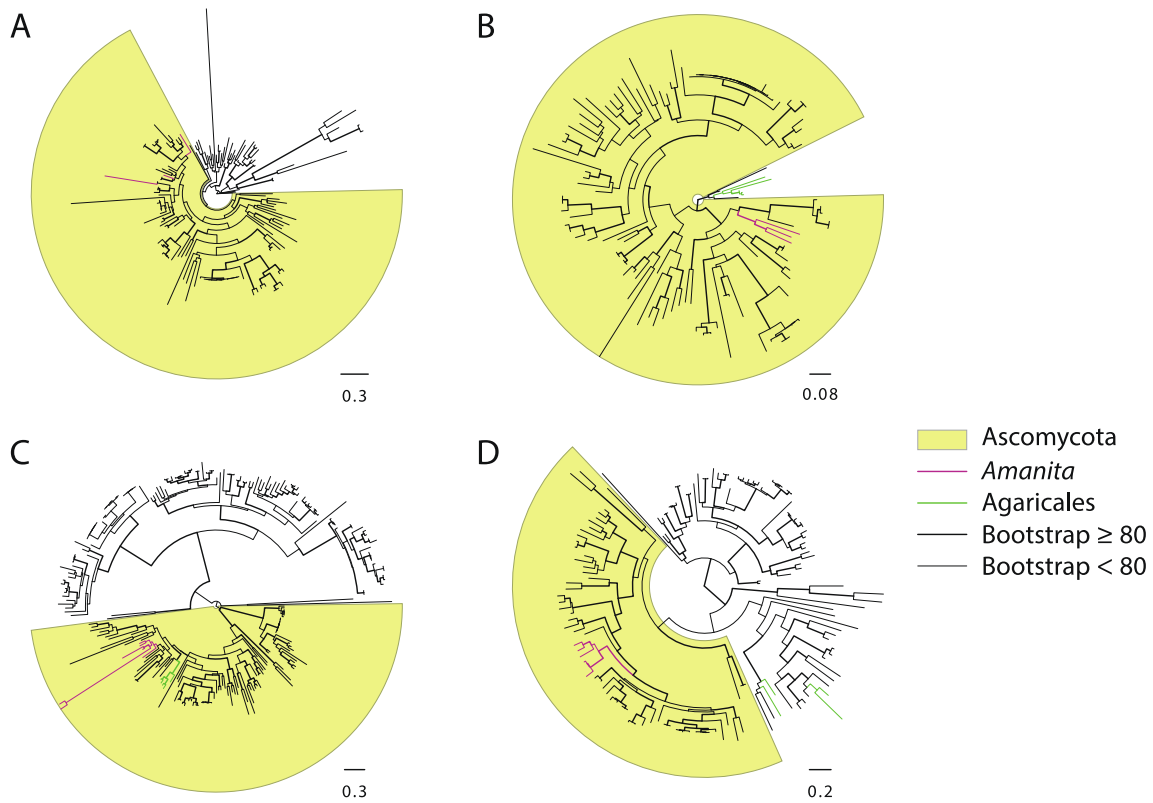


Fig. 4. Syntenic structure between scaffold 17 of *A. muscaria* var. *guessowii* (ECM) and the scaffold 77 of *A. thiersii* (asymbiotic) highlighting the location of one gene (AmumRNA.3446.1; purple) putatively identified as deriving from *de novo* gene birth. Light blue: scaffolds; Dark blue: exons; Black: introns; Brown: homologous genes (determined by BLASTx); Grey: non-alignable regions. Units: kbp.



897

898 Fig. 5. Protein trees of the four putative HGT gene families. A: family 2813 (with unknown
 899 function); B: family 10418 (ACC deaminase); C: family 11987 (metal binding acid phosphatase);
 900 D: family 12806 (with unknown function).

901

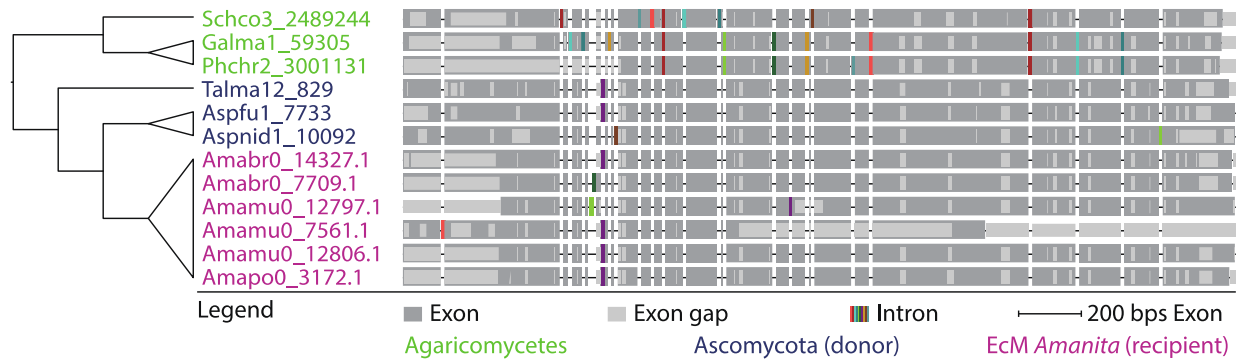


Fig. 6. Gene structures of the HGT family 12806 (with unknown function). The intron structures of ECM *Amanita* genes deriving from HGT are more similar to their putative donors' homologs than to homologs belonging to other Agaricales members.

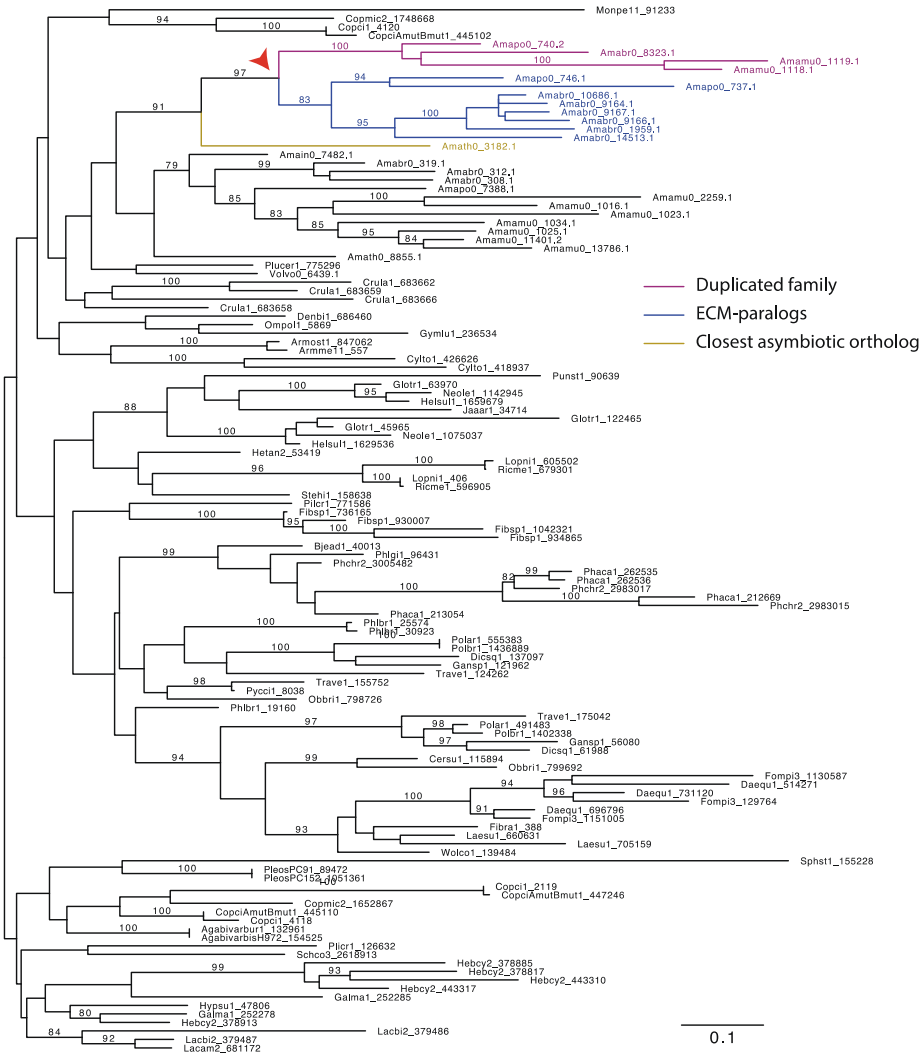


Fig. 7. Phylogeny of gene duplication in family 1119 (DUF6). Branches with bootstrap values ≥ 80 are labeled. Arrow head: duplication generating family 1119.