bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Comparative genomics and divergence time estimation of the anaerobic fungi in herbivorous mammals

Yan Wang[1,2,*], Noha Youssef[3], M.B. Couger[4], Radwa Hanafy[3], Mostafa Elshahed[3], and Jason E. Stajich[1,2,*]

[1]Department of Microbiology and Plant Pathology, University of California, Riverside, Riverside, California, USA
[2]Institute for Integrative Genome Biology, University of California, Riverside, Riverside, California, USA
[3]Department of Microbiology and Molecular Genetics, Oklahoma State University, Stillwater, Oklahoma, USA
[4]High Performance Computing Center, Oklahoma State University, Stillwater, Oklahoma, USA

*Corresponding authors.

Yan Wang, yanxw.wang@gmail.com, and Jason E. Stajich, jason.stajich@ucr.edu.

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

The anaerobic gut fungi (AGF) or Neocallimastigomycota inhabit the rumen of herbivorous mammals where they support plant fiber degradation. These obligate anaerobes have large and AT-biased genomes, which have limited genome-wide studies and the phylogenomic analyses of these fungi. Using newly generated genomes and transcriptomes of 27 Neocallimastigomycota taxa, we have explored their evolutionary relationships, divergence time, and gene content. The most recent common ancestor of the AGF is estimated to have diverged 73.5±5 million years ago (Mya), which coincides with the estimated ages of grasses (Poaceae) and evolution of mammalian herbivory. Comparative genomics identified lineage-specific genes and protein family domains including three that may have been acquired through horizontal gene transfer from animal hosts, plants, and rumen gut bacteria. The concordance of independently estimated divergence times of Neocallimastigomycota fungi, grasses, and ruminate lineages suggest AGF were important in shaping the success of modern ruminants and enabling their efficient acquisition of energy from recalcitrant plant material.

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Diverse microbes inhabit the digestive tract of ruminant mammals and contribute to degradation of ingested plant fiber, which liberates nutrients for their hosts. Large scale rumen metagenomic sequencing and analyses have produced hundreds of novel bacterial genomes enabling discovery of plant biomass degrading enzymes and patterns of genomic evolution [1,2]. However, eukaryotic members of the rumen microbial community have been less intensely studied [3,4]. The phylum Neocallimastigomycota is composed of anaerobic gut fungi (AGF) associated with herbivorous mammals and reptiles and are important contributing members to the rumen environment [5]. Adaptation to living in the rumen requires ability to grow in the anoxic and prokaryotes-dominated environment. The AGF have acquired changes as part of the adaptation to ruminate hosts including loss of the mitochondria and gain of a hydrogenosome, the loss of respiratory capacities, and the substitution of ergosterol with tetrahymanol in the cell membrane [6]. Importantly, the AGF have a remarkably efficient plant biomass degradation machinery, which may be critical for competing with other microbes for resources and establishing growth in the herbivorous gut. Such capacity is reflected in the possession of an impressive arsenal of plant biomass degradation enzymes and the production of the cellulosomes—extracellular structures that harbor multiple extracellular enzymes bound to scaffoldins [4]. These metabolic and structural adaptations improve survivability, fitness, and competitiveness of the AGF in the herbivorous gut, but the genetic and evolutionary origins of these changes remain undescribed [3,7]. Genomic investigations of the AGF have identified a massive number of carbohydrate active enzymes coded that were likely acquired through Horizontal Gene Transfer (HGT) events from multiple bacterial lineages [3,4,7]. Multiple examples of HGT from bacteria to fungi have been documented [8–11]. However, HGT events in fungi that have an eukaryote origin are still rare with only a few described cases from animals [12], oomycetes [13], or plants [14]. The rumen is an intriguing context to explore patterns of HGT where degradative enzymes break down sorts of cells liberating DNA and RNA. Competing organisms can find advantage by evolving or acquiring enzymes that operate efficiently in an anaerobic environment to obtain nutrients from recalcitrant plant fibers.

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The Neocallimastigomycota were traditionally Chytridiomycota (chytrids) fungi, which have classified in part by their flagellated zoospore stage [15–18]. Efforts to resolve the phylogenetic relationship of AGF using ribosomal markers have yielded conflicting topologies [19,20]. Multilocus phylogeny or phylogenomics has not yet been applied to evaluate evolutionary relationships and divergence time of the AGF. Using a collection of genomes and transcriptomes from 27 different AGF taxa representing seven out of the ten recognized genera, we reconstruct a robust phylogenomic tree of the AGF and estimated their divergence time. We compared the genomes or transcriptomes of AGF and their non-rumen Chytridiomycota relatives to identify unique and shared gene content. This study examines the relatively recent divergence of the Neocallimastigomycota, their evolutionary relationships to other fungal relatives in the tree of life, and the concordance of the timing of the AGF evolution with the emergence of herbivory in mammals and the grasses they consume. As the AGF are also highly effective at plant biomass degradation, we also explored the potential contribution of HGT to the genomic evolution of these fungi and identified their possible donor lineages in animals, bacteria, and plants.

Results

Divergence time estimation and phylogenomic relationship of Neocallimastigomycota

Phylogenomic analysis placed the 27 AGF taxa into a single monophyletic clade with strong support of Bayesian posterior probability (1.0/1.0) and maximum likelihood bootstrap value (100%) (Fig. 1 and Supp. Fig. 1). All AGF genera (*Anaeromyces*, *Caecomyces*, *Feramyces,* *Neocallimastix*, *Orpinomyces*, *Pecoramyces*, and *Piromyces*) included in this study formed individual monophyletic clades that were also supported by both Bayesian (Fig. 1) and maximum likelihood analyses (Supp. Fig. 1). A conflict in the tree topology between the two phylogenetic reconstructions is the placement of the *Caecomyces* clade. This lineage is sister to the rest of the Neocallimastigomycota in the maximum likelihood tree (Supp. Fig. 1), while the *Caecomyces* position is swapped with *Piromyces* in the Bayesian phylogeny (Fig. 1). This is likely due to short internode distances, which suggest a rapid radiation of the ancestors of these two genera. The relative short bar of the highest-probability density (HPD) on the node of the AGF clade

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(Fig. 1) suggests the integrative natural history of this group of fungi and the outperforming resolving power of the genome-wide data in the molecular dating analysis.

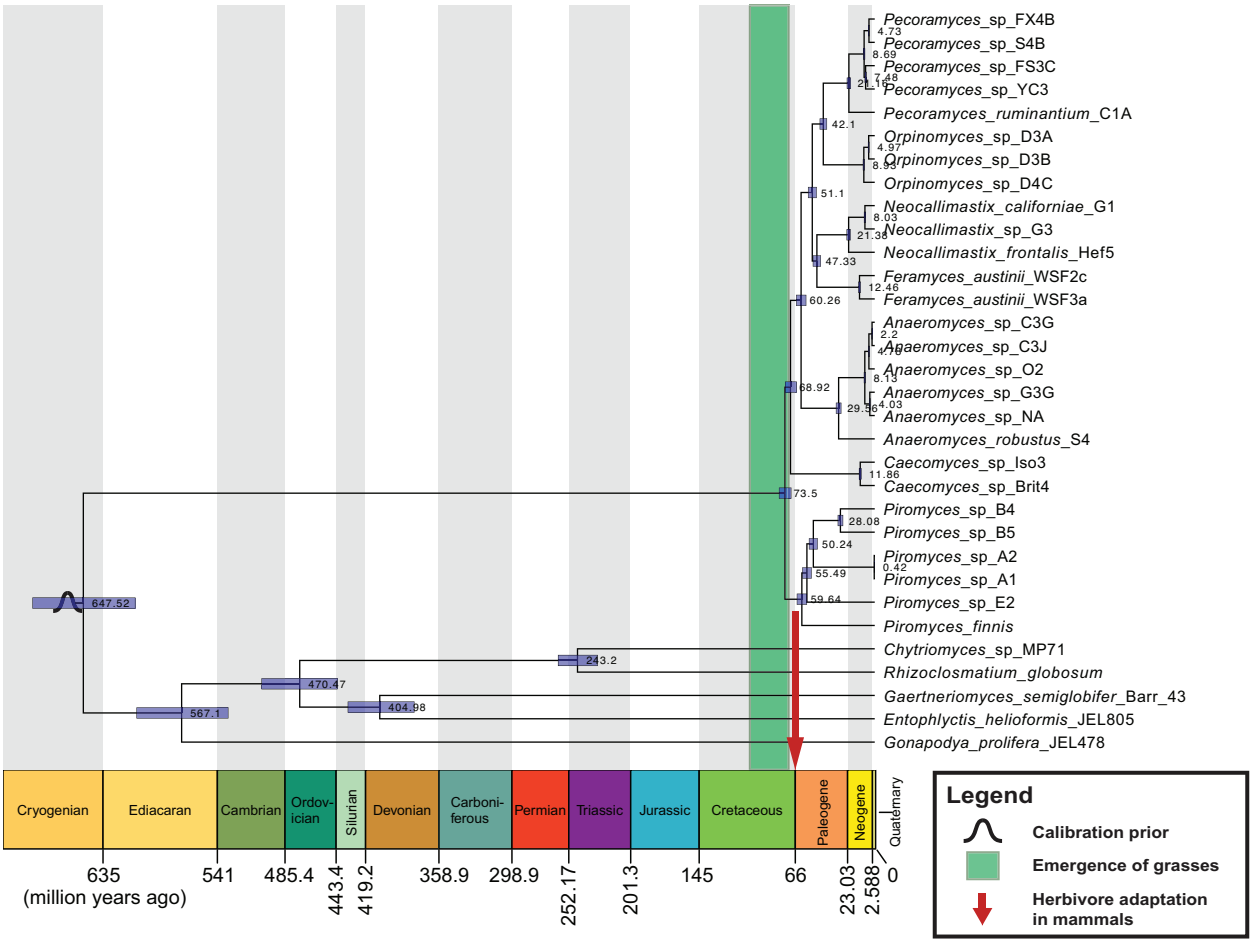

Fig. 1. Bayesian phylogenomic Maximum Clade Credibility tree of Neocallimastigomycota with divergence time estimation. All clades are fully supported by Bayesian posterior probabilities (BPP). Mean ages and 95% highest-probability density ranges (blue bars) are denoted on each node. The emergence time of the grasses is shaded in green box (within the Cretaceous). The red arrow indicates the dietary radiation time of placental mammals from insectivores to herbivores or omnivores (at the boundary of Cretaceous and Paleogene).

The divergence time of the Neocallimastigomycota clade is estimated at the end of Cretaceous 73.5 (±5) Mya (Fig. 1). This estimate corresponds to the age of the grasses (70-95 Mya), previously estimated using molecular (nuclear and chloroplast) markers with calibrations of fossil pollen, dinosaur coprolite, and the breakup time of the Gondwana [21–25]. The inferred divergence time of AGF is also consistent with the timing of a major diet change of placental mammals. Loss of the chitinase gene diversity in mammalian lineages was likely necessary as part of the diet transition from primarily insectivore to herbivorous and omnivorous mammalian

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

lineages that occurred around the boundary of Cretaceous–Paleogene, which also overlaps with the estimated divergence time of the AGF clade (Fig. 1) [26]. The chronogram displays a characteristic long branch leading to the AGF which bridges from boundary of Cryogenian–Ediacaran (~647 Mya) to the end Cretaceous (~73.5 Mya). This suggests the distinction of the AGF and that the modern lineages of AGF did not diverge from the most common ancestor until Cretaceous and further diversified in the Paleogene period (Fig. 1).

Genome-wide comparison of protein family domains and homologous genes

Pfam domains comparison identified distinct gain-and-loss patterns between AGF and their non-rumen associated chytrids relatives (Fig. 2). We identified 40 Pfam domains that are unique to the AGF, representing 0.67% of all (5,980) Pfams in the AGF pan-genome-transcriptome. The predicted functions of these domains include anaerobic ribonucleotide reductase ("NRDD"), metal transport and binding ("FeoA", "FeoB_C"), carbohydrate binding (e.g., "CBM_10", "CBM-like", "Cthe_2159"), atypical protein kinase ("CotH"), and glycoside hydrolase (e.g., "Glyco_hydro_6", "Glyco_hydro_11"). The AGF proteomes are also enriched for domains involved in plant material fermentation and polysaccharide degradation (Fig. 2). These include the "Chitin_binding_1", "CBM_1", "Cellulase", "Glyco_hydro_10", "Gly_radical", "RicinB_lectin_2", "Esterase", and "Polysacc_deac_1", which have higher copy number in AGF genomes than in the sister chytrid lineages. Three of the detected domains, "Cthe_2159", "Gal_lectin", and "Rhamnogal_lyase", have never been previously detected in fungal genomes. On the other hand, our analysis identified that 106 Pfam domains have been lost in AGF genomes and transcriptomes, most of which are related to oxidation reactions on cytochromes and mitochondria (Supp. Table 1). In addition, domains involved in the biosynthesis of nicotinic acid, uric acid, purine catabolism, photolyase, pathways of ureidoglycolate and kynurenine are also found to be missing in AGF species (Fig. 2 and Supp. Table 1). Similar patterns were revealed in the homologous gene comparisons (Supp. Fig. 2). A permissive criterion, allowing some missing copies, found a total of 2,728 gene families shared between AGF and chytrids. We discovered that 1,709 additional gene families are shared among AGF genomes (each gene

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

present in at least 22 out of the total 27 taxa) but missing from other chytrids, while another 367 families are missing in AGF lineages and present in the other chytrid lineages.

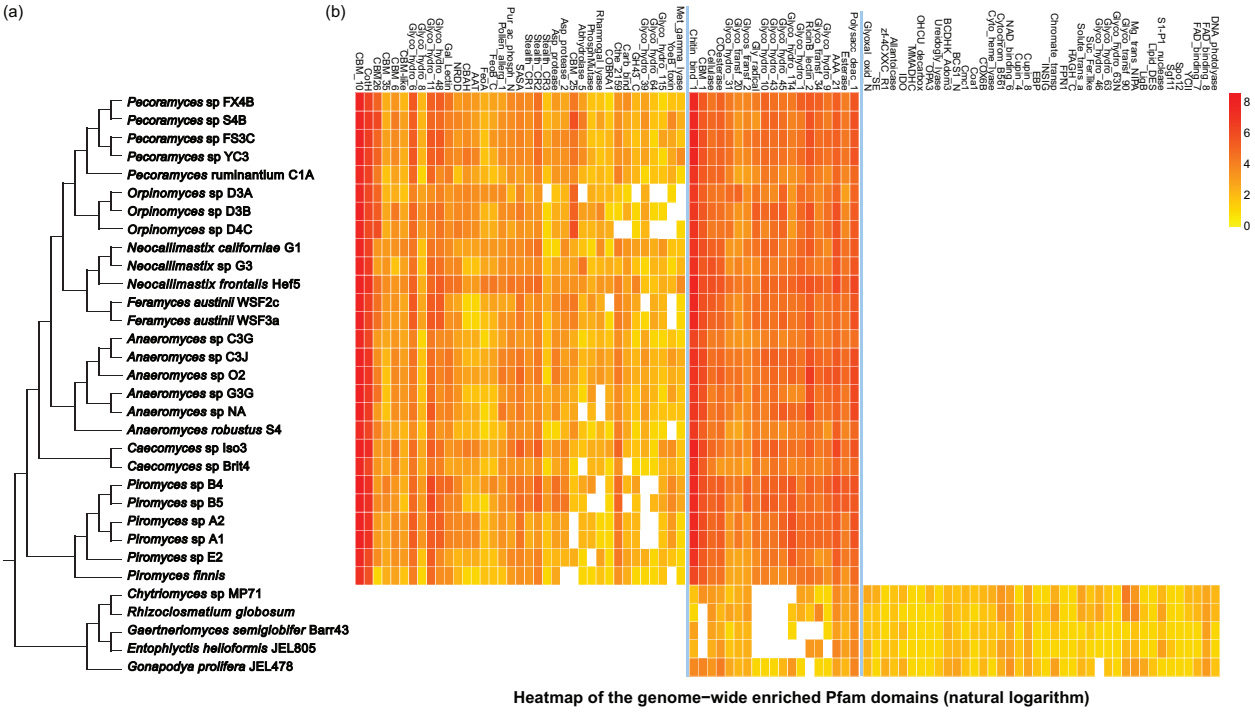

Fig. 2. Heatmap comparison of the genome-wide enriched Pfam domains between Neocallimastigomycota and Chytridiomycota. (a) Cladogram showing the phylogenetic relationship of the compared taxa (Neocallimastigomycota genomes are in bold, transcriptomes in standard type); (b) Heatmap plot of natural logarithm of the domain copy numbers (domains are ordered from Neocallimastigomycota in groups as uniquely gained, enriched, to lost; presented data are partial, for the full list please refer to Supp. Table 1).

Genomic interactions within the gut system of mammalian herbivores

We focused on three unique Neocallimastigomycota Pfam domains ("Rhamnogal_lyase", "Gal_lectin", and "Cthe_2159") that have never been encountered in the fungal kingdom. Phylogenetic analyses strongly suggest that these three domains were horizontally acquired from plants, animals, and bacteria respectively.

a) *A plant-like rhamnogalacturonate lyase ("Rhamnogal_lyase")*

In plants, the rhamnogalacturonate lyases are involved in the fruit ripening-related process, cell-wall modification, and lateral-root and root-hair formation [27,28]. The Pfam database classifies two types of domains for rhamnogalactoside degrading activity: "Rhamnogal_lyase" and "RhgB_N". They are both N-terminal catalytic domains on the rhamnogalacturonan lyase protein (polysaccharide lyase family 4, PL4) and flanked persistently by the group of "fn3_3" and

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

"CBM-like" domains, with a particular function to degrade the rhamnogalacturonan I (RG-I) backbone of pectin. The "Rhamnogal_lyase" domain exists in members of plants and plant-pathogenic bacteria (e.g. *Erwinia chrysanthemi*), whereas the "RhgB_N" domain has a wider distribution and can be found in bacteria, fungi, and oomycetes [29]. Searches against current databases (e.g., EnsEMBL, MycoSosm, Pfam) suggest that the "Rhamnogal_lyase" domain found in AGF represents the first documented observation of this domain in Fungi. The phylogenetic tree including available PL4 proteins shows that AGF "Rhamnogal_lyase" domains are more similar to the plant homologs and distantly related to the clades of fungi and oomycetes, which implies the novelty and potential foreign origin of these AGF domains (Fig. 3). The presence of "Rhamnogal_lyase" domain suggests that the AGF may possess a plant-like ability to soften, modify, and degrade the plant pectin within the rumen. In addition, the phylogenetic tree groups the "Rhamnogalacturonate lyase A" with the "RhgB_N"-containing proteins, which indicates that they are probably synonymous names.

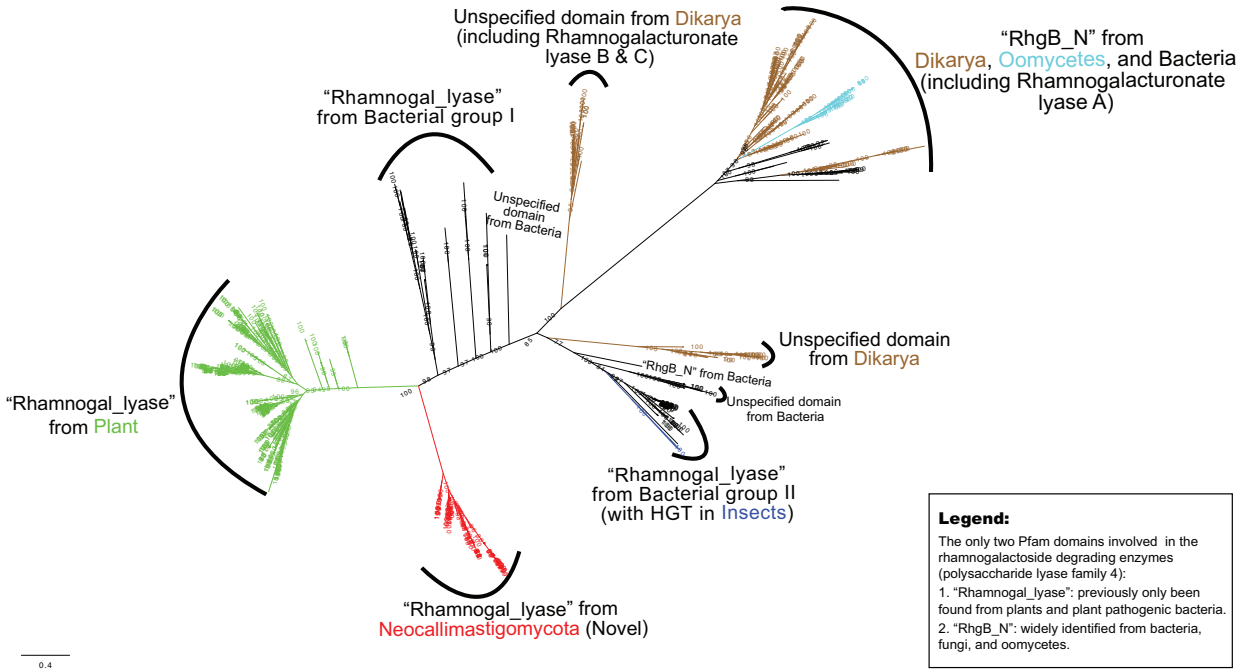

Fig. 3. Radial phylogenetic tree of the plant-like "Rhamnogal_lyase" domain encoded by the Neocallimastigomycota (red). Plant copies are colored in green and other homologous fungal genes are colored in brown. Oomycetes are in cyan and animal copies only known in the mountain pine beetles *Dendroctonus ponderosae* are in blue. Bacterial branches remain in black. The tree also included homologs of "RhgB_N" and "Rhamnogalacturonan lyase A, B, and C". Domain names are suggested using NCBI's conserved domain search tool (cutoff $1E^{-5}$) with unaligned FASTA sequences (refer to Supp. Fig. 3 for a detailed mid-point rooting tree with all leaf labels).

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

b) *An animal-like galactose binding lectin domain ("Gal_Lectin")*

Among the Fungi, the "Gal_Lectin" domain is uniquely found in AGF but is absent in chytrids and all other fungal genomes. Phylogenetic analysis recovered a monophyletic clade of AGF "Gal_Lectin" domains embedded with animal homologs (Fig. 4a). The three distinct animal subclades represent "Gal_Lectin"-containing proteins with different annotated functions (Fig. 4b). The homologs of the ruminant hosts (e.g., horse, sheep) of AGF are also included in each of the animal subclades. The proteins in the animal subclade 1 were annotated as the "polycystic kidney disease" (*PKD1*) gene product polycystin-1 (PC-1). The members of the animal subclade 2 matched with the "adhesion G-protein coupled receptor L1/3" (ADGRL1). The animal subclade 3 contains homologs of the "EVA-1 protein", which includes two adjacent

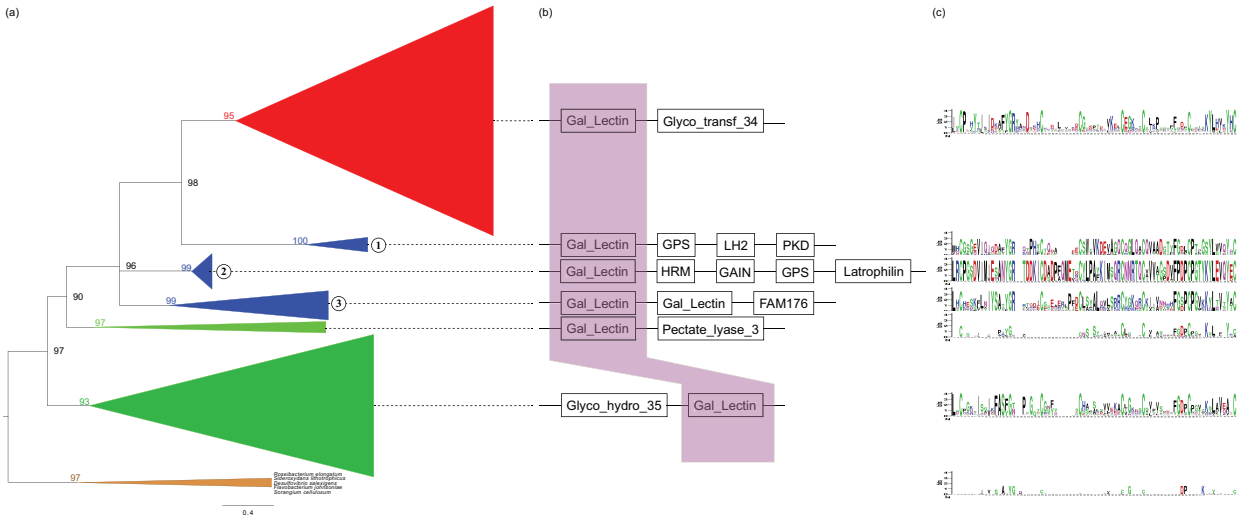

Fig. 4. Phylogenetic tree of the animal-like "Gal-Lectin" found in Neocallimastigomycota. (a) Collapsed phylogenetic tree based on aligned Gal_Lectin amino acid sequences, including clades of Neocallimastigomycota (red), Animals (blue; three clades are labeled as 1-3), green algae (light green), plant (dark green), and bacteria (brown); (b) Conserved flanking domains of the "Gal_Lectin" region across different clades (shaded region highlights the aligned sequences); (c) protein logos of the "Gal_Lectin" domain for each clade.

"Gal_Lectin" domains. The distinguishing features of the 3 types of animal homologs also reflect in the flanking domains (Fig. 4b). The AGF "Gal_Lectin" domain likely has an animal origin based on the embedded position in the phylogenetic tree, which demonstrates a close relationship with the animal PC-1 protein. In addition, two amino acid sites are conserved among animal and AGF copies of the protein and substituted in all other clades, for example, the Arginine at position 18 and Tyrosine at the position 74 (Fig. 4c).

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

c) *A bacteria-like biomass-binding and putatively polysaccharide lyase domain ("Cthe_2159")*

Multiple biomass-binding (and degrading) domains were identified uniquely in AGF but missing in the other chytrid lineages (Fig. 2). One domain,"Cthe_2159", was described as a novel polysaccharide lyase-like protein and originally identified from the thermophilic and biomass-degrading bacterium *Clostridium thermocellum* [30]. This family is characterized as a cellulose and acid-sugar (polygalacturonic acid, a major component of the pectin) binding beta-helix protein and is primarily found in Archaea and Bacteria. Remarkably, there are at least 126 copies of the "Cthe_2159" domain found among the 27 genomes or transcriptomes of AGF, however, this domain is absent in all other Eukaryote genomes examined. A phylogenetic tree of the "Cthe_2159", including homologs revealed from Archaea, Bacteria, and AGF, strongly supports the hypothesis of the HGT origin of the fungal "Cthe_2159" (Fig. 5). It implies that the donor of the fungal gene is an obligate mammalian gut bacterium, *Clostridiales bacterium* (ML bootstrap value 98%) and this gene is also closely related to the ones in a clade of anaerobic bacteria (comprising *Oribacterium sinus*, *Oribacterium* sp., and *Hungatella hathewayi*) (Fig. 5).

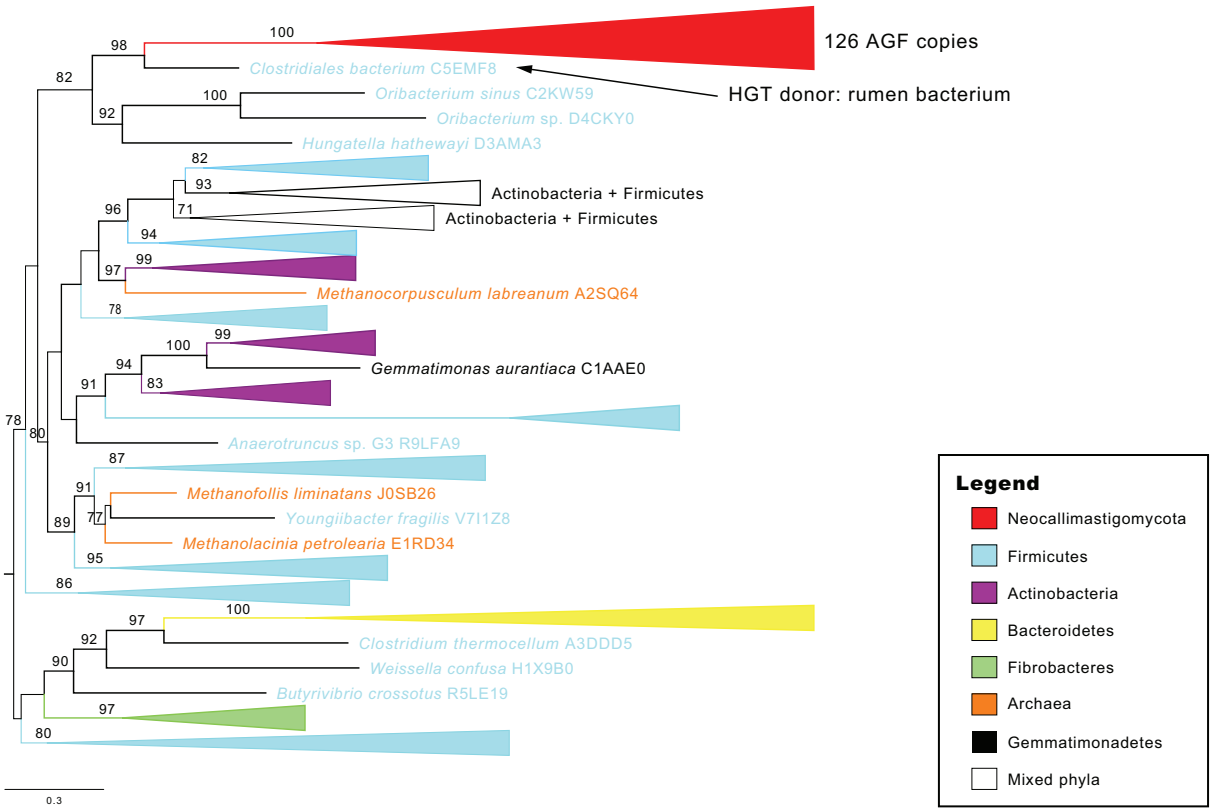

Fig. 5. Mid-point rooted phylogenetic tree of the "Cthe_2159" domain. All 126 Neocallimastigomycota (AGF) copies form a single clade (red) indicating the HGT donor, *Clostridiales bacterium* C5EMF8 (an obligate rumen bacterium), with strong support of maximum likelihood bootstrap (98/100). Included bacterial lineages were assigned different colors according to their phylogenetic classification (see legend).

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

Our results provide new insights to ruminant gut fungi evolution. The analyses find that AGF have diverged relatively recently (73.5±5 Mya, Fig. 1), an age that is in remarkable concordance with salient events in plant (evolution of the Poaceae) and mammalian (transition from insectivore to herbivore and omnivore) evolution. Grass evolution enabled the herbivory transition, and the adaptations drove an increase in developmental and morphological complexity of the digestive tract, compartmentalization, and the development of dedicated anaerobic fermentation chambers (e.g., rumen and caecum) in the herbivorous alimentary tract to improve biomass degradation efficiency [31]. This transition to plant-based (or plant-exclusive) diets required additional partnership with microbes since mammals lack cellulolytic and hemi-cellulolytic enzymes necessary to liberate sugars for absorption [5]. The divergence time of the AGF lineages is coincident with the transitionary time of the mammal diets (66 Mya) and explains their primary colonizer roles for the forage in the rumen of herbivorous animals [32]. In addition, the genome content comparisons help illustrate and predict new biological roles AGF play in the mammalian herbivore guts. Our analyses identified multiple instances of Pfam domain gains (n=40) and losses (n=106) within the Neocallimastigomycota clade (Fig. 2 and Supp. Table 1). Interestingly, three of these incidents that resulted in the acquisition of Pfam domains are missing from the rest of the examined genomes of the fungal kingdom. These Pfam domains were horizontally acquired from animals, bacteria, and plants separately (Figure 3-5) and highlighted how HGT has contributed to broaden the lignocellulolytic capacities and potentially increase the fungal abilities in cell recognition with immune advantages in the rumen.

The “Rhamnogal_lyase” (PL4 family) and “Cthe_2159” function in pectin binding or degradation activities, and the possession of both suggests that AGF may have specialized ability to deconstruct pectin distinctly from other fungi. Pectin is abundant in primary cell walls and the middle lamella in both dicotyledonous plants (making up 20-35% dry weight) and grasses (2-10%) serving as a protection of plant cells from degrading enzymes produced by animals [33–36]. Efforts have shown that removal of pectin can effectively increase the exposed cell wall surface, and improve the accessibility of other polysaccharides (cellulose and hemicellulose) masked by

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

pectin [37]. "Rhamnogal_lyase" has been suggested to play important roles for cell wall modification in plants [27], while in the view of pathogenic bacteria, it can be used to disorganize the plant tissue for invasion purposes [38]. The gain of "Rhamnogal_lyase" indicates a key synapomorphy of the modern AGF taxa, which could be an important contributor to the AGF ability to utilize the polysaccharide in plant cell walls. The "Cthe_2159" was recently identified with the resolved crystal structure showing the ability to bind cellulosic and pectic substrates [30]. Although the right-handed parallel -helix structure is conserved across bacteria and some archaea (a common fold to CE, GH, and PL enzyme families), the "Cthe_2159" does not align with any characterized enzyme. Three Calcium ions binding sites of the "Cthe_2159" exhibit similarity with pectate lyases in the PL9 family. Utilization of "Cthe_2159" and "Rhamnogal_lyase" may have contributed to the AGF efficient machinery by degrading the pectin that bind plant cells together so that more surface of the ingested plant materials can be exposed to diverse polysaccharide enzymes in the rumen. These Pfam domains could account for the superior performance that AGF have to weaken forage fiber and degrade polysaccharides [39,40]. The AGF may benefit or depend on these acquired domains in its capacity in the primary degradation of ingested forage, which leaves other microbes to process partially digest remains [41].

The third AGF-unique Pfam domain, "Gal_Lectin", bears the phylogenetic hallmark of being acquired from an animal donor. Animals use galactose-binding lectins to recognize foreign entities [42] and participate in anti-microbial defenses [43,44]. Interestingly, the animal origin of the AGF "Gal_Lectin" domain is similar to the extracellular motif of the PC-1 (Fig. 4), an integral membrane protein functioning in cell recognition [45,46]. In vitro, PC-1 shows binding ability to carbohydrate matrices and collagens type I, II, and IV [47]. As such, we postulate that the acquisition of the animal-like "Gal_Lectin" domain contributes to the AGF abilities of cell-cell recognition and interaction with other microbes in the rumen. Syntenic relationship of the gene copies shows that the AGF copies are flanked by the Glyco_transf_34 domain, which is not found in any other animal homologs (Fig. 4b). The AGF-equipped Glyco_transf_34 belongs to the galactosyl transferase GMA12/MNN10 family and may help catalyze the transfer of the sugar moieties in cooperating with the adjacent "Gal_Lectin" domain. Our investigation found that HGT has contributed to the genome evolution of AGF with donors from both bacteria and

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

eukaryotic lineages of animals and plants. HGT may have played important roles in the ability of these fungi to acquire new functions and thrive in the anaerobic gut as a key member of the microbial community degrading plant material in animal hosts.

Materials and Methods

Transcriptome and genome datasets

We generated the transcriptomes of 22 strains of Neocallimastigomycota fungi from cow, sheep, horse, and goat feces, and rumen fluid of fistulated cows in the Stillwater, OK area (Murphy and Youssef et al. in review; Table 1). These strains were maintained under anaerobic conditions using the modified Hungate method as described previously [48–51] . Total volume of RNA was harvested from the growing fungal strains and processed for transcriptomics sequencing, which was performed using an external commercial service provided by Novogene (Beijing, China). The RNAseq data were assembled into *de novo* transcript assemblies using Trinity (v2.6.6), followed by TransDecoder (v5.0.2) to predict ORFs [52]. The generated proteomes and corresponding coding sequences were used as input to phylogenomic and comparative genomic analyses. The five published Neocallimastigomycota genome sequences were obtained from JGI MycoCosm database [53,54]. These are *Anaeromyces robustus* S4, *Neocallimastix californiae* G1, *Pecoramyces ruminantium* C1A (synonym *Orpinomyces* sp.), *Piromyces finnis (v3.0)*, and *Piromyces* sp. E2 [3,4]. Five outgroup Chytridiomycota taxa with sequenced genomes were chosen. These are *Chytriomyces* sp. MP 71, *Entophlyctis helioformis* JEL805, *Gaertneriomyces semiglobifer* Barr 43, *Gonapodya prolifera* JEL478, and *Rhizoclosmatium globosum* JEL800 [55,56].

Table 1. Information for the AGF strains included in this study.

Genus	Species	Strain	Source	Type	Host
Anaeromyces	*Anaeromyces contortous*	Na	OkState	Transcriptome	Cow feces
Anaeromyces	*Anaeromyces contortous*	C3J	OkState	Transcriptome	Cow feces
Anaeromyces	*Anaeromyces contortous*	G3G	OkState	Transcriptome	Goat feces
Anaeromyces	*Anaeromyces contortous*	O2	OkState	Transcriptome	Cow feces (frozen)

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genus	Species	Strain	Source	Type	Host
Anaeromyces	*Anaeromyces contortous*	C3G	OkState	Transcriptome	Cow feces
Anaeromyces	*Anaeromyces robustus*	S4	JGI	Annot Genome	Sheep feces
Caecomyces	*Caecomyces* sp.	Brit4	OkState	Transcriptome	Rumen-cow
Caecomyces	*Caecomyces* sp.	Iso3	OkState	Transcriptome	Cow feces
Feramyces	*Feramyces austinii*	WSF3a	OkState	Transcriptome	Feces- aoudad sheep (wild)
Feramyces	*Feramyces austinii*	WSF2c	OkState	Transcriptome	Feces- aoudad sheep (wild)
Orpinomyces	*Orpinomyces sp.*	D3A	OkState	Transcriptome	Cow digesta
Orpinomyces	*Orpinomyces sp.*	D3B	OkState	Transcriptome	Cow digesta
Orpinomyces	*Orpinomyces sp.*	D4C	OkState	Transcriptome	Cow digesta
Pecoramyces	*Pecoramyces ruminantium*	C1A	JGI	Annot Genome	Cow - Feces
Pecoramyces	*Piromyces* sp.	S4B	OkState	Transcriptome	sheep feces
Pecoramyces	*Piromyces* sp.	FX4B	OkState	Transcriptome	Rumen-cow
Pecoramyces	*Piromyces* sp.	FS3c	OkState	Transcriptome	Rumen-cow
Pecoramyces	*Piromyces* sp.	YC3	OkState	Transcriptome	Rumen-cow
Piromyces	*Piromyces finnis*	Pirfi3	JGI	Annot Genome	Horse
Piromyces	*Piromyces* sp.	E2	JGI	Annot Genome	Elephant - Feces
Piromyces	*Piromyces* sp.	A1	OkState	Transcriptome	Sheep feces (frozen)
Piromyces	*Piromyces* sp.	A2	OkState	Transcriptome	Sheep feces (frozen)
Piromyces	*Piromyces* sp.	B4	OkState	Transcriptome	Cow feces (frozen)
Piromyces	*Piromyces* sp.	B5	OkState	Transcriptome	Cow feces (frozen)
Neocallimastix	*Neocallimastix californiae*	G1	JGI	Annot Genome	Horse
Neocallimastix	*Neocallimastix frontalis*	Hef5	OkState	Transcriptome	Cow - Feces
Neocallimastix	*Neocallimastix* sp.	G3	OkState	Transcriptome	Sheep feces

Phylogenomics and divergence time estimation

A set of 434 highly conserved and generally single-copy protein coding genes in fungi and animal and plant outgroups (https://github.com/1KFG/Phylogenomics_HMMs) were used for phylogenomic analyses in the PHYling pipeline (https://github.com/stajichlab/PHYling_unified).

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Profile-Hidden-Markov-Models of these markers were searched in the chytrid predicted protein sequences using HMMER3 (v3.1b2). A total of 426 (out of 434) conserved orthologous markers were identified with hmmsearch (cutoff=1E^{-10}) in the 27 Neocallimastigomycota and 5 Chytridiomycota. The identified protein sequence homologs in each species, for each phylogenetic marker, were aligned with hmmalign to the marker profile-HMM. The protein alignments were also back translation into codon alignments guided by the protein alignment using the tool bp_mrtrans.pl [57]. The protein and coding sequences of the markers were concatenated into a super-alignment with 426 partitions defined by each gene marker. The 426 gene partitions were further collapsed into 33 partitions by PartitionFinder v.2.1.1 with a greedy search to find partitions with consistent phylogenetic signals [58]. Phylogenetic trees were constructed from this super-alignment and partition scheme with two methods—maximum likelihood implemented in IQ-TREE (v.1.5.3) and Bayesian inference implemented in BEAST (v.1.8.4) [59,60]. Configuration files for divergence time estimation analysis were coded in BEAUti v.1.8.4 using the 33 partitions and one calibration prior to account for the emergence of Chytridiomycota (672 ± 30 Ma) [55,61]. The Birth-Death incomplete sampling tree model was employed for inter-species relationships analyses [62]. Unlinked strict clock models were used for each partition. Three independent runs were performed separately for 50 million generations each with random starting seeds. Sufficient ESS (>200) values were obtained after the default burn-in (10%) for the final sampled trees. The Maximum Clade Credibility (MCC) tree was compiled using TreeAnnotator v.1.8.4.

Identification of AGF-specific genes and Pfam domains

Orthologous genes across the 32 genomes or transcriptomes were identified using a comparative genomic pipeline that utilized all-vs-all BLASTp (cutoff=1E^{-5}) to obtain the similarity pairs, Orthagogue to identify putative orthologous relationships, and the Markov-Clustering Algorithm (MCL using the inflation value of 1.5) to generate disjoint clusters and deployed in a pipeline (https://github.com/stajichlab/Comparative_pipeline) [63–65]. Comparisons of shared gene content of the Orthologous clusters was performed among the Chytridiomycota lineages using a permissive strategy of counting a gene family as shared if it is missing in up to 5 of the 27

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Neocallimastigomycota taxa and 1 of the 5 chytrids genomes. In this scenario, genes absent in all chytrids genomes and maintained by more than 22 out of the 27 Neocallimastigomycota genomes/transcriptomes are defined as AGF unique genes; on the other hand, genes missing from all Neocallimastigomycota and present in at least 4 out of the 5 chytrids genomes are treated as AGF lost genes.

Protein domains were identified by searching the proteomes predicted from each genome assembly or transcriptome assembly against the Protein Family (Pfam) database (v31.0, last accessed at March 20th, 2018). The enrichment heatmap of the Pfam domains across the included taxa was produced using the "aheatmap" function in the R package "NMF" based on the total copy number count in each assembly [66]. To identify potentially unique and important Pfam domains in the AGF, we firstly distinguished the ones uniquely present in the AGF genomes and missing from all of the included free-living chytrids relatives.

To identify genes in AGF that are likely important for interactions with mammalian hosts and plant material breakdown, we further compared the five available AGF genomes to the genomes of their animal hosts (e.g., sheep, horse, elephant, yak) [67–70], the diet plant (e.g., moss, rice, palm, maize, sorghum) [71–80] (Table S2), and the 1,165 available fungal genomes from the ongoing 1KFG project [17,18,53,54]. To prioritize AGF genes that may have been laterally acquired from these hosts, a Python script [12] and similarity search tool BLAT [81] was applied to filter out genetic elements in AGF with higher similarity to animal or plant homologs than any fungal ones, excluding the AGF themselves. Candidate genes for lateral transfer were ranked by the combination of the two strategies. The candidate genes with an assigned functional or biological process annotation were analyzed with priority using phylogenetic reconstruction to infer their potential origin.

Identification of homologous sequences and potential origin of HGT candidate loci

Three Pfam domains "Rhamnogal_lyase", "Gal_Lectin", and "Cthe_2159" were identified to be unique to the AGF genomes as compared to the Chytridiomycota fungi or all other fungal members (identified using following approaches). To predict the donor lineages for these putative HGT events, we searched more broadly for homologues in genome databases of Plant,

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Metazoa, Fungi, Bacteria, and Protists in EnsEMBL (v37) [82] via the web-implemented HMMER tool (https://www.ebi.ac.uk/Tools/hmmer/) (cutoff=1E^{-3}). Additional fungal homologues were found by searching the DOE JGI's MycoCosm database [53,54]. The profile Hidden Markov Model tool phmmer in the HMMer package [83] was used to search for similar sequences in the 1,165 fungal genomes using the query of edge-trimmed domain sequences from *An. robustus* (cutoff=1E^{-3}).

The "RhgB_N" domain is distantly related to "Rhamnogal_lyase" (24% similarity between copies from the *Aspergillus nidulans* and *An. robustus*), and both are involved in the degradation of the pectin rhamnogalacturonan I region. Members of the "RhgB_N" sequences were obtained from the Pfam database classified in the "*RhgB_N*" (PF09284) family [29] along with previously annotated members of the rhamnogalacturonate lyase families A, B, and C from GenBank [84–86]. A single dataset of "RhgB_N" and "Rhamnogal_lyase" family members from animals, fungi, plants, and bacteria was constructed from these searches. Domain names were confirmed using NCBI's conserved domain search tool (cutoff=1E^{-5}) with unaligned FASTA sequences [87]. Similarly, homologs of the "Gal_Lectin" and "Cthe_2159" were obtained by searching for similar sequences in the previously described genome databases and the categorized Pfam database (families of "*Gal_Lectin* (PF02140)" and "*Cthe_2159* (PF14262)"). Homologous sequences containing the "Cthe_2159" domain were only identified in Archaea and Bacteria, while the AGF copies are the first eukaryotic representatives identified with this domain. Highly similar sequences (>90%) were filtered using CD-HIT v4.6.4 followed by multiple sequence alignment with MUSCLE v3.8.31 [88,89]. Sequence and phylogenetic analyses were performed on the High-Performance Computing Center (HPCC) at the University of California, Riverside.

Phylogenetic trees of candidate HGT and homologs

In total, 747 sequences of the rhamnogalacturanate degradation proteins (including both "Rhamnogal_lyase" and "RhgB_N") were included in the alignment. For the other two domains, "Gal_Lectin" and "Cthe_2159", the alignments include 297 and 234 unique variants respectively. Both the upstream and downstream flanking regions of the studied Pfam domain

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sequences were trimmed using the Mesquite software [90]. Selection of appropriate substitutional model, maximum-likelihood phylogenetic tree reconstruction, and ultrafast bootstrapping (1000 replicates) were conducted using the IQ-TREE v1.5.5 package [59,91,92]. Sequence logos of the "Gal_Lectin" domains were generated using the WebLogo tool for each clade [93]. The protein homology of the animal subclades of the "Gal_Lectin" phylogenetic tree was suggested by similarity searches using BLASTp against the available non-redundant (nr) database [64].

Acknowledgements

This work was supported by National Science Foundation Grants (DEB-1557110 to J.E.S. and DEB-1557102 to N.Y. and M.E.). Data analyses were performed on the University of California Riverside High-Performance Computational Cluster supported by NSF DBI-1429826 and NIH S10-OD016290.

References

1. Stewart, R. D. *et al.* metagenomic sequencing of the cow rumen. *Nat. Commun.* **9,** 870 (2018).

2. Seshadri, R. *et al.* Cultivation and sequencing of rumen microbiome members from the Hungate1000 Collection. *Nat. Biotechnol.* **36,** 359–367 (2018).

3. Youssef, N. H. *et al.* The genome of the anaerobic fungus orpinomyces sp. strain c1a reveals the unique evolutionary history of a remarkable plant biomass degrader. *Appl. Environ. Microbiol.* **79,** 4620–4634 (2013).

4. Haitjema, C. H. *et al.* A parts list for fungal cellulosomes revealed by comparative genomics. *Nat Microbiol* **2,** 1–8 (2017).

5. Gruninger, R. J. *et al.* Anaerobic fungi (phylum Neocallimastigomycota): Advances in understanding their taxonomy, life cycle, ecology, role and biotechnological potential. *FEMS Microbiol. Ecol.* **90,** 1–17 (2014).

6. Yarlett, N., Orpin, C. G., Munn, E. A., Yarlett, N. C. & Greenwood, C. A.

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Hydrogenosomes in the rumen fungus *Neocallimastix patriciarum*. *Biochem. J.* **236,** 729–739 (1986).

7. Solomon, K. V. *et al.* Early-branching gut fungi possess a large, comprehensive array of biomass-degrading enzymes. *Science (80-.).* **351,** 1192–1195 (2016).

8. Dhillon, B. *et al.* Horizontal gene transfer and gene dosage drives adaptation to wood colonization in a tree pathogen. *Proc. Natl. Acad. Sci.* **112,** 3451–3456 (2015).

9. Gardiner, D. M. *et al.* Comparative Pathogenomics Reveals Horizontally Acquired Novel Virulence Genes in Fungi Infecting Cereal Hosts. *PLoS Pathog.* **8,** (2012).

10. Pombert, J.-F. *et al.* Gain and loss of multiple functionally related, horizontally transferred genes in the reduced genomes of two microsporidian parasites. *Proc. Natl. Acad. Sci.* **109,** 12638–12643 (2012).

11. Chaib De Mares, M. *et al.* Horizontal transfer of carbohydrate metabolism genes into ectomycorrhizal Amanita. *New Phytol.* **205,** 1552–1564 (2014).

12. Wang, Y., White, M. M., Kvist, S. & Moncalvo, J.-M. Genome-wide survey of gut fungi (Harpellales) reveals the first horizontally transferred ubiquitin gene from a mosquito host. *Mol. Biol. Evol.* **33,** 2544–2554 (2016).

13. Sun, G., Yang, Z., Kosch, T., Summers, K. & Huang, J. Evidence for acquisition of virulence effectors in pathogenic chytrids. *BMC Evol. Biol.* **11,** (2011).

14. Richards, T. A. *et al.* Phylogenomic analysis demonstrates a pattern of rare and ancient horizontal gene transfer between plants and fungi. *Plant Cell* **21,** 1897–1911 (2009).

15. James, T. Y. *et al.* Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature* **443,** 818–822 (2006).

16. James, T. Y. *et al.* A molecular phylogeny of the flagellated fungi (Chytridiomycota) and description of a new phylum (Blastocladiomycota). *Mycologia* **98,** 860–871 (2006).

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

17. Spatafora, J. W. *et al.* The fungal tree of life: from molecular systematics to genome-scale phylogenies. *Microbiol. Spectr.* **5,** 3–34 (2017).

18. Stajich, J. E. Fungal genomes and insights into the evolution of the Kingdom. *Microbiol. Spectr.* **5,** 1–15 (2017).

19. Wang, X., Liu, X. & Groenewald, J. Z. Phylogeny of anaerobic fungi (phylum Neocallimastigomycota), with contributions from yak in China. *Antonie Van Leeuwenhoek* **110,** 87–103 (2017).

20. Liggenstoffer, A. S., Youssef, N. H., Couger, M. B. & Elshahed, M. S. Phylogenetic diversity and community structure of anaerobic gut fungi (phylum Neocallimastigomycota) in ruminant and non-ruminant herbivores. *ISME J.* **4,** 1225–1235 (2010).

21. Vicentini, A., Barber, J. C., Aliscioni, S. S., Giussani, L. M. & Kellogg, E. A. The age of the grasses and clusters of origins of C4 photosynthesis. *Glob. Chang. Biol.* **14,** 2963–2977 (2008).

22. Prasad, V., Strömberg, C. A. E., Alimohammadian, H. & Sahni, A. Dinosaur coproliites and the early evolution of grasses and grazers. *Science (80-.).* **310,** 1177–1180 (2005).

23. Gaut, B. S. Evolutionary dynamics of grass genomes. *New Phytol.* **154,** 15–28 (2002).

24. Bremer, K. Gondwanan evolution of the grass alliance of families (Poales). *Evolution (N. Y).* **56,** 1374–1387 (2002).

25. Christin, P. A. *et al.* Molecular dating, evolutionary rates, and the age of the grasses. *Syst. Biol.* **63,** 153–165 (2014).

26. Emerling, C. A., Delsuc, F. & Nachman, M. W. Chitinase genes (CHIAs) provide genomic footprints of a post-Cretaceous dietary radiation in placental mammals. *Sci. Adv.* **4,** eaar6478 (2018).

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

27. Molina-Hidalgo, F. J. *et al.* The strawberry (Fragaria×ananassa) fruit-specific rhamnogalacturonate lyase 1 (FaRGLyase1) gene encodes an enzyme involved in the degradation of cell-wall middle lamellae. *J. Exp. Bot.* **64,** 1471–1483 (2013).

28. Ponniah, S. K., Thimmapuram, J., Bhide, K., Kalavacharla, V. K. & Manoharan, M. Comparative analysis of the root transcriptomes of cultivated sweetpotato (Ipomoea batatas [L.] Lam) and its wild ancestor (Ipomoea trifida [Kunth] G. Don). *BMC Plant Biol.* **17,** 1–14 (2017).

29. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44,** D279–D285 (2016).

30. Close, D. W., D'angelo, S. & Bradbury, A. R. M. A new family of β-helix proteins with similarities to the polysaccharide lyases. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **70,** 2583–2592 (2014).

31. Hackmann, T. J. & Spain, J. N. Ruminant ecology and evolution: perspectives useful to ruminant livestock research and production. *J. Dairy Sci.* **93,** 1320–1334 (2010).

32. Nicholson, M. J., Theodorou, M. K. & Brookman, J. L. Molecular analysis of the anaerobic rumen fungus Orpinomyces - Insights into an AT-rich genome. *Microbiology* **151,** 121–133 (2005).

33. Vogel, J. Unique aspects of the grass cell wall. *Curr. Opin. Plant Biol.* **11,** 301–307 (2008).

34. Voragen, A. G. J., Coenen, G. J., Verhoef, R. P. & Schols, H. A. Pectin, a versatile polysaccharide present in plant cell walls. *Struct. Chem.* **20,** 263–275 (2009).

35. Xiao, C. & Anderson, C. T. Roles of pectin in biomass yield and processing for biofuels. *Front. Plant Sci.* **4,** 1–7 (2013).

36. Salem, H. *et al.* Drastic genome reduction in an herbivore's pectinolytic symbiont. *Cell*

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

171, 1520–1525 (2017).

37. Pakarinen, A., Zhang, J., Brock, T., Maijala, P. & Viikari, L. Enzymatic accessibility of fiber hemp is enhanced by enzymatic or chemical removal of pectin. *Bioresour. Technol.* **107,** 275–281 (2012).

38. Laatu, M. & Condemine, G. Rhamnogalacturonate lyase rhiE is secreted by the out system in *Erwinia chrysanthemi*. *J. Bacteriol.* **185,** 1642–1649 (2003).

39. Borneman, W. S., Akin, D. E. & Ljungdahl, L. G. Fermentation products and plant cell wall-degrading enzymes produced by monocentric and polycentric anaerobic ruminal fungi. *Appl. Environ. Microbiol.* **55,** 1066–1073 (1989).

40. Nagpal, R. *et al.* Anaerobic rumen fungi: potential and applications. in *Agriculturally Important Microorganisms* (eds. Khachatourians, G. G., Arora, D. K., Rajendran, T. P. & Srivastava, A. K.) **I,** 375–393 (Academic World International, 2009).

41. Haitjema, C. H., Solomon, K. V., Henske, J. K., Theodorou, M. K. & O'Malley, M. A. Anaerobic gut fungi: advances in isolation, culture, and cellulolytic enzyme discovery for biofuel production. *Biotechnol. Bioeng.* **111,** 1471–1482 (2014).

42. García-Maldonado, E., Cano-Sanchez, P. & Hernandez-Santoyo, A. Molecular and functional characterization of a glycosylated galactose-binding lectin from *Mytilus californianus*. *Fish Shellfish Immunol.* **66,** 564–574 (2017).

43. Low, D. H. P. *et al.* Molecular interfaces of the galactose-binding protein tectonin domains in host-pathogen interaction. *J. Biol. Chem.* **285,** 9898–9907 (2010).

44. Uhlenbruck, G. & Steinhausen, G. Tridacnins: symbiosis-profit or defense-purpose? *Dev. Comp. Immunol.* **1,** 183–192 (1977).

45. Hughes, J. *et al.* The polycystic kidney disease 1 (PKD1) gene encodes a novel protein with multiple cell recognition domains. *Nat. Genet.* **10,** 151–160 (1995).

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

46. Weston, B. S., Malhas, A. N. & Price, R. G. Structure-function relationships of the extracellular domain of the autosomal dominant polycystic kidney disease-associated protein, polycystin-1. *FEBS Lett.* **538,** 8–13 (2003).

47. Weston, B. S., Bagnéris, C., Price, R. G. & Stirling, J. L. The polycystin-1 C-type lectin domain binds carbohydrate in a calcium-dependent manner, and interacts with extracellular matrix proteins in vitro. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1536,** 161–176 (2001).

48. Hanafy, R. A., Elshahed, M. S., Liggenstoffer, A. S., Griffith, G. W. & Youssef, N. H. Pecoramyces ruminantium, gen. nov, sp. nov., an anaerobic gut fungus from the feces of cattle and sheep. *Mycologia* **5514,** 37–41 (2017).

49. Bryant, M. P. Commentary of anaerobic on the Hungate technique for culture. *The American Journal of Clinical Nutrition* **25,** 1324–1328 (1972).

50. Balch, W. E. & Wolfe, R. S. New approach to the cultivation of methanogenic bacteria: 2-mercaptoethanesulfonic acid (HS-CoM)-dependent growth of Methanobacterium ruminantium in a pressurized atmosphere. *Appl. Environ. Microbiol.* **32,** 781–791 (1976).

51. Hungate, R. E. & Macy, J. The roll-tube method for cultivation of strict anaerobes. *Bull. from Ecol. Res. Comm.* **17,** 123–126 (1973).

52. Haas, B. J. *et al. De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8,** 1494–1512 (2013).

53. Grigoriev, I. V. *et al.* MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.* **42,** D699–D704 (2014).

54. Spatafora, J. W. 1000 fungal genomes to be sequenced. *IMA Fungus* **2,** 41 (2011).

55. Chang, Y. *et al.* Phylogenomic analyses indicate that early fungi evolved digesting cell walls of algal ancestors of land plants. *Genome Biol. Evol.* **7,** 1590–1601 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

56. Mondo, S. J. *et al.* Widespread adenine N6-methylation of active genes in fungi. *Nat. Genet.* **49,** 964–968 (2017).

57. Stajich, J. E. *et al.* The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* **12,** 1611–1618 (2002).

58. Lanfear, R., Calcott, B., Ho, S. Y. W. & Guindon, S. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.* **29,** 1695–1701 (2012).

59. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32,** 268–274 (2015).

60. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7,** 214 (2007).

61. Wang, Y., White, M. & Moncalvo, J. Divergence time estimation of the gut fungi *Smittium* (Harpellales) suggests a co-emergence with the complete metamorphosis of their lower Diptera (fly) hosts. *Rev.*

62. Stadler, T. On incomplete sampling under birth–death models and connections to the sampling-based coalescent. *J. Theor. Biol.* **261,** 58–66 (2009).

63. Ekseth, O. K., Kuiper, M. & Mironov, V. OrthAgogue: an agile tool for the rapid prediction of orthology relations. *Bioinformatics* **30,** 734–736 (2014).

64. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215,** 403–410 (1990).

65. Van Dongen, S. Graph clustering by flow simulation. **May,** (University of Utrecht, 2000).

66. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11,** 367 (2010).

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

67. Broad Institute. Elephant Genome Project. (2018). Available at: http://www.broadinstitute.org/scientific-community/science/projects/mammals-models/elephant/elephant-genome-project.

68. Wade, C. M. *et al.* Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science (80-.).* **326,** 865–867 (2009).

69. The International Sheep Genomics Consortium *et al.* The sheep genome reference sequence: a work in progress. *Anim. Genet.* **41,** 449–453 (2010).

70. Qiu, Q. *et al.* The yak genome and adaptation to life at high altitude. *Nat. Genet.* **44,** 946–949 (2012).

71. Martin, G. *et al.* Improvement of the banana '*Musa acuminata*' reference sequence using NGS data and semi-automated bioinformatics methods. *BMC Genomics* **17,** 1–12 (2016).

72. Singh, R. *et al.* Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. *Nature* **500,** 335–339 (2013).

73. Peng, Z. *et al.* The draft genome of the fast-growing non-timber forest species moso bamboo (*Phyllostachys heterocycla*). *Nat. Genet.* **45,** 456–461 (2013).

74. Zimin, A. V. *et al.* Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a grogenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *bioRxiv* **27,** 787–792 (2017).

75. Jiao, Y. *et al.* Improved maize reference genome with single-molecule technologies. *Nature* **546,** 524–527 (2017).

76. The Rice Annotation Project. Curated genome annotation of *Oryza sativa* ssp. japonica and comparative genome analysis with *Arabidopsis thaliana*. *Genome Res.* **17,** 175–183 (2007).

77. The International Brachypodium Initiative *et al.* Genome sequencing and analysis of the

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

model grass Brachypodium distachyon. *Nature* **463,** 763–768 (2010).

78. Paterson, A. H. *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457,** 551–556 (2009).

79. Swarbreck, D. *et al.* The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* **36,** D1009–D1014 (2008).

80. Rensing, S. A. *et al.* The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science (80-.).* **319,** 64–69 (2008).

81. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).

82. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46,** D754–D761 (2018).

83. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7,** (2011).

84. Yoshino-Yasuda, S., Karita, S., Kato, M. & Kitamoto, N. Sequence analysis and heterologous expression of rhamnogalacturonan lyase A gene (AsrglA) from Shoyu Koji Mold, Aspergillus sojae KBN1340. *Food Sci. Technol. Res.* **18,** 901–909 (2012).

85. Gomez-Cortecero, A., Harrison, R. J. & Armitage, A. D. Draft Genome Sequence of a European Isolate of the Apple Canker Pathogen Neonectria ditissima. *Genome Announc.* **3,** 10–11 (2015).

86. Hacquard, S. *et al.* Survival trade-offs in plant roots during colonization by closely related beneficial and pathogenic fungi. *Nat. Commun.* **7,** 11362 (2016).

87. Marchler-Bauer, A. *et al.* CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* **45,** D200–D203 (2017).

88. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28,** 3150–3152 (2012).

89. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).

90. Maddison, W. & Maddison, D. Mesquite: a modular system for evolutionary analysis Version 2.75. (2007).

91. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Le, S. V. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35,** 518–522 (2017).

92. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14,** 587–589 (2017).

93. Crooks, G., Hon, G., Chandonia, J. & Brenner, S. WebLogo: a sequence logo generator. *Genome Res* **14,** 1188–1190 (2004).

Legends for supplementary materials:

Supp. Fig. 1. Maximum likelihood phylogenetic tree of Neocallimastigomycota using Chytridiomycota as the outgroup. All bootstrap values (out of 100) are labeled on the branches.

Supp. Fig. 2. Presence (dark gray) and absence (light gray) of the homologous gene families across the genomes (and transcriptomes) of Neocallimastigomycota and Chytridiomycota. The 4,824 gene families were selected as universal homologous genes that present at least 22 out of the 27 Neocallimastigomycota genomes (and transcriptomes) with missing no more than 1 of the 5 included Chytridiomycota genomes. In addition, it also includes the unique gene families that are strictly absence from all Chytridiomycota but encoded by the Neocallimastigomycota (missing no more than 5 out of the 27 taxa).

Supp. Fig. 3. Mid-point rooted phylogenetic tree of the plant-like “Rhamnogal_lyase” domain encoded by the Neocallimastigomycota (red). Labels are same as Figure 3.

Supp. Table 1. List of domains and potential functions that are found either unique or lost in Neocallimastigomycota (comparing to Chytridiomycota).

bioRxiv preprint doi: https://doi.org/10.1101/401869; this version posted August 31, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supp. Table 2. Genome information of the animal hosts and diet plants used in the study to infer the genetic elements in Neocallimastigomycota that have a foreign origin.