**The draft genome sequence of *Hyphochytrium catenoides*; a fungal analogue.**

Guy Leonard1, Aurélie Labarre1, David Milner1, Adam Monier1, Darren Soanes1, Jeremy Wideman1, Finlay Maguire1, Matthew W. Brown2, Sam Stevens1, Bill Wickstead3, Divya Sain4, Jason E. Stajich4, Neil Hall5, Konrad Paszkiewicz1, & Thomas A. Richards1, 6

1 Biosciences, University of Exeter, Geoffrey Pope Building, Exeter, EX4 4QD, UK.

2 Department of Biological Sciences, Mississippi State University, Mississippi State, MS 39762, USA.

3 Nottingham

4 Riverside

5 Earlham

6 Canadian Institute for Advanced Research, CIFAR Program in Integrated Microbial Biodiversity, Toronto, ON, M5G 1Z8

Corresponding Author: Thomas Richards

Biosciences,

University of Exeter,

Geoffrey Pope Building,

Exeter, EX4 4QD, UK.

Running Title: **genome sequence of the ‘Pseudofungus’ *Hyphochytrium catenoides***

Keyword:

**Abstract**

Eukaryotic microbes feed by four primary mechanisms: phagotrophy, intracellular infection, photosynthesis (cellular consumption and digestion of other cells) and osmotrophy (secretion of enzymes, breakdown of organic material and uptake of nutrients). Traits associates with the later three functions arose independently multiple times in the eukaryotes. Fungi successfully coupled trophic functions with filamentous growth in the evolution of this group. Pseudofungi (oomycetes and the hyphochytriomycetes) also followed a similar evolutionary trend representing a pattern of convergent evolution. Both the fungi and the pseudofungi encompass a large diversity of microbial forms, including many important parasites of plants and animals, suggesting filamentous osmotrophic traits underpin subsequent evolution of parasitic characteristics. However, sequencing efforts have largely focused on host associated microbes, providing few genome datasets of free-living relatives. This has hampered the study of the evolution of filamentous osmotrophic systems decoupled the evolution of parasitic characteristics. Here we report the draft genome sequence of *Hyphochytrium catenoids;*the first representative of a free-living hyphochytriomycete ‘pseudofungus’. This protist forms complex lifecycle involving diverse filamentous structures and a flagellated spore with a single atypical anterior-tinselated flagellum. Phylogenetic analysis confirms that *Hyphochytrium* branches sister to the oomycetes. We find a minimal ‘footprint’ of genes that have phylogenetic ancestry putatively arising from a plastid endosymbiosis but identify a footprint of a large DNA virus introgression with related genes present in oomycete genomes.  We use the genomic data to investigate the ancestry of oomycete parasitic characteristics and metabolic traits demonstrating these features are derived and likely evolved recently within the oomycete diversification. While comparative genomics reveals clear differences in the repertoire of gene families associated with filamentous growth (i.e. vesicle trafficking systems, cell wall synthesis pathways, motor proteins, photoreceptors) between the fungi and the pseudofungi, providing a picture of the gene repertoire evolution that underpinned the convergent evolution of filamentous osmotrophic growth.**Introduction**

Stramenopiles (1) (also known as heterokonts (2)) are highly diverse branch of protist, that encompass a multitude of biological forms including huge multicellular kelps ‘seaweeds’, photosynthetic micro-algae through to a variety of microbial pathogens, some of which --oomycetes for example-- feed and grow like fungi and cause important diseases of aquatic animals and plants (3, 4). The stramenopiles are a phylogenetically robust group (e.g. (5)) defined by the presence of two motile flagella, a ‘standard’ smooth posterior flagellum and a ‘tinsel’ anterior flagellum with a tripartite rigid tubular mastigoneme (“hairs”) (2). However, examples of secondary flagellum loss have occurred during the radiation of this group, for example, in the hyphochytrids (6) which have lost a smooth posterior flagellum but retained a tinsel anterior flagellum. As such comparative genomics of hyphochytrids in comparison to other Stramenopiles provides a tool for understanding the diversification of the protein repertoire associated with eukaryotic flagella characteristics.

Many stramenopiles are photosynthetic and retain a plastid of secondary endosymbiotic ancestry (7, 8). Some commentators have hypothesised that this plastid is derived from an ancient event and as such the plastid has a shared ancestry with the plastid organelles present in dinoflagellates and Apicomplexa (see (9)). This hypothesis has been challenged on a number of levels; either by phylogenomic data which supports a pattern of recent multiple (or serial) eukaryote–eukaryote endosymbioses (10-12) implying later plastid acquisitions, or by evidence suggesting heterotrophic stramenopiles, specifically oomycetes, lack a statistically significant footprint of genes of endosymbiotic ancestry (13). As such the ancestry of secondary ‘plastid’ endosymbiosis relative to the radiation of the stramenopiles remains controversial.

Environmental sequencing, specifically of marine environments (e.g. (14)), has increased the known phylogenetic diversity of this group suggesting that the stramenopiles encompass one of the most diverse higher-level groups within the eukaryotes (15). Representatives of these groups remain uncultured with little gene/genome sampling, therefore the biology of these groups is under-explored. Furthermore, genome-sequencing efforts in the stramenopiles have largely focused on photosynthetic algae (e.g. (16, 17)) or oomycete pathogens (e.g. (18, 19)), leaving the large diversity of heterotrophic free-living stramenopiles under sampled.

To address this imbalance we have sequenced the genome of *Hyphochytrium catenoides* (ATCC 18719) a free-living hyphochytrid protist that forms hyphal-like networks and spores with only a single anterior tinsel flagellum (6, 20). The hyphochytrids are thought to branch sister to the oomycetes (4, 21) and both these groups seem to lack a plastid organelle and photosynthetic function. However, both groups grow as filamentous/polarized cells feeding osmotrophically by extracellular secretion of digestive enzymes coupled to nutrient uptake (4, 6, 20). These characteristics mean that they resemble fungi and as such have previously been named Pseudofungi (4). We used the genome sequence data to confirm the phylogenetic branching position of the hyphochytrids, investigate characters shared with the oomycetes and identify the genes encoding cellular characteristics absent or shared with fungi that typify filamentous/osmotrophic growth. We also use the genome data to investigate proteome evolution associated with loss of the posterior flagellum in the hyphochytrids including the diversity of photoreceptors putatively encoded. As part of this analysis we also conducted a phylogenomic survey to search for a remnant footprint of plastid endosymbiosis in the *Hyphochytrium* genome. These data provide unique sampling of a free-living Stramenopile in order to facilitate further genomic and cellular analysis.

**Results**

**Genome assembly and gene model prediction**

Using one lane of paired-end (100 bp) Illumina HiSeq data, trimmed and cleaned (using TagCleaner (22) and PRINSEQ (23)) and subsequently digitally normalised with KHMER (24) in combination with two paired-end Illumina GAiix 76 bp lanes treated in the same way, we assembled a draft genome of *Hyphochytrium catenoides* (ATCC 18719). To improve the genome assembly, we used a range of different read trimming settings and assembly programs choosing a ‘working assembly’ based on a number of optimised metrics. This approach demonstrated that the ‘assembly hypothesis’ given by the program Platanus v1.2.1 (parameters detailed in methods) produced an assembly with the highest N50 and fewest number of scaffolds. The assembly methodology used in Platanus has been reported to improve the assembly of genomes with multiple ploidy (25).

Previously we had sequenced a transcriptome from the same culture strain of *Hyphochytrium* (26) using 454 FLX sequencing of cDNA reads and assembled it in Newbler 2.5 (ref) using the default cDNA settings. We cleaned 70 sequences from this assembly by removing contigs of less than 100 bp in length (excluding the polyA regions) and/or contigs that consisted of predominately repeat motifs. This resulted in 6,202 transcript sequences with an N50 of 1,044 bp. To investigate how ‘complete’ our genome assembly was we utilised three approaches: i) we compared the genes identified from the transcriptome using BLASTn (e-values from 1e-10 to 1e-100 in increments of 10) to the genomic assembly, ii) we used the Core Eukaryotic Genes Mapping Approach (CEGMA) pipeline (27) and iii) due to the author’s discontinuing CEGMA we trialled BUSCO - Benchmarking Universal Single-Copy Orthologs - which aims to give a similar metric, but is instead based on the presence of ‘near-universal single-copy orthologues’ in Eukaryotes. These analyses demonstrated that the genome assembly was predicted to be respectively 97.8%, 91.5%, and 52% complete in terms of gene sampling (Fig. S1). We suspect that both CEGMA and BUSCO underestimate the completeness of genomes as the core gene list is derived from a subset of eukaryotic genomes that does not fully sample a diverse collection of eukaryotic genomes, this is a similar concern with BUSCO (which only samples from Fungi and Metazoa in version 1.2) which inevitably gives a much lower estimation of completion. Specifically, Fig. 1a/b shows the CEGMA/BUSCO scores across the stramenopiles, demonstrating a 27/16 complete/partial gene family absence in the *H. catenoides* assembly according to CEGMA. This is a moderate level of absence compared with other stramenopile genomes (e.g. *Blastocystis hominis* and *Ectocarpus siliculosus* with 76/53 and 71/31 complete/partial gene family absences respectively). Using the CEGMA and transcriptome comparison approaches described, we then investigated the completeness of the assembly when scaffolds below 1 Kbp, 5 Kbp, and 10 Kbp were removed (Fig. S1). This analysis demonstrates that the removal of scaffolds less than 1 Kbp has a negligible effect on the predicted completeness of the genome. As such we have submitted the >=1Kbp scaffold assembly along with the predicted proteome as a draft genome to the EMBL EBI database for community access. See Fig. 1b for details of assembly in-comparison to other eukaryotic genomic sequences.

K-mer counting using Jellyfish along with two publically available scripts (estimate\_genome\_size.pl and the website GenoScope [see https://github.com/josephryan/estimate\_genome\_size.pl and [http://qb.cshl.edu/genomescope/])](http://qb.cshl.edu/genomescope/%5D)) were used to predict a haploid genome size of between 54.2 Mbp and 132.8 Mbp with the estimate\_genome\_size.pl tool with the >=1Kbp assembly equating to 65.7 Mbp across 4,758 scaffolds with a scaffold N50 size of 35.57 Kbp and an L50 of 399 scaffolds. The average sequencing coverage of this assembly was estimated to be 312x as calculated using the ‘estimate\_genome\_size.pl’ tool. However, using the ‘genomeCoverageBed’ from BEDTOOLS (ref) the average coverage over the >=1Kbp scaffolds is 610x. Extraction and purification of long strands of DNA was not possible even though we tried multiple DNA extraction protocols (possibly due to carbohydrate contamination) preventing sequencing using a long-read technology and pulsed-field gel electrophoresis. As such we used a RT PCR method for estimation of genome size (28) suggesting a haploid genome size of 43.73 Mb (S.E.M. = 1.41). These data suggest we recovered a useful draft genome sequence of *H. catenoides*.

To identify any prokaryotic contamination in the >=1 Kbp scaffold assembly we first conducted BLASTn searches of the assembly using prokaryotic SSU and LSU rDNA sequences as search seeds (e.g. *Escherichia coli* [CP012802] and *Sulfolobus acidocaldarius* [NR\_043400 & NR\_076363]). This analysis only returned sequences of similarity to the *H. catenoides* mitochondria genome assembly (discussed below), suggesting that no prokaryotic SSU and LSU rRNA sequences are present. Next, we subjected all 4,758 genome scaffolds to a BLASTx analysis against a database of 65 eukaryotic and 164 representative prokaryotic complete predicted proteomes (Table S1) with a gathering threshold of 1e-10. This approach did not identify any scaffolds that did not have at least one top hit to a eukaryotic genome for a subsection of the scaffold. Indeed, only 87 of the scaffolds had > 50% of the subsections with a top BLAST hit to a prokaryotic genome and only 20 of the scaffolds had > 70% of their top BLAST hits to a prokaryotic genome. These 20 scaffolds were inspected manually, 11 of these showed the presence of putative spliceosomal introns and/or other genes more similar to other eukaryotic genes. For the remaining nine scaffolds (totalling 31.8 Kbp) we could not exclude them as possible prokaryotic contamination (listed in Table S2).

Comparisons of GC content versus read coverage coupled with BLASTn analysis to identify likely taxonomic affiliation of a contig (e.g. blobology (29)) has emerged as useful tool for identifying contamination of genome sequencing projects (30). We undertook this approach on both the > 1 Kbp scaffold assembly and the total assembly, and the graphs did not identify any suspect traces of contamination, however, they do show the presence of the mitochondrial genome as an aberrant cluster of ‘blobs’, i.e. with lower than average GC content (Fig. S2a-d).

A fourth round of checks for contamination were conducted by using tetramer counting of the >1Kbp scaffold dataset for the building of Emergent Self Organising Maps (ESOMs) (31). These use similarities in the 4-mer frequencies to build, by way of an artificial neural network, an emergent ‘map’ of the input space properties of the data. Two runs of the software developed in (31) were completed (see Fig. S3a+b); a) the *Hyphochytrium* scaffolds only, and b) the *Hyphochytrium* scaffolds (orange) along with the scaffolds from eight other ‘small’ genomes which were added to the tetramer frequency dataset, (Bacteria (blue): *E. coli, Mycobacterium tuberculosis;* Archaea (grey): *Methanococcus vanniellii, Sulfolobus solfataricus;* Fungi (purple): *Encephalitozoon intestinalis, S. cerevisiae;* Archaeplastida (green): *Ostreococcus tauri;* Protist (red): *Cryptosporidium hominis*). The maps produced in Fig. S3 show no indication of overlap or features indicative of contamination.

Analysis using RepeatMasker (32) and the ‘2015-08-07’ repbase library demonstrated that the > 1 kb genome assembly contained 9.53% repeat regions where 1.79% of the repeat sequence was assigned to transposable elements. The *Hyphochytrium* genome contained a full set of transfer RNAs including an additional tRNA for selenocysteine. Putative genes and their predicted proteins were identified using a combined pathway as described in the methods. This approach identified 18,481 putative gene models (of which 406 represented different splice forms), a total gene number similar to the mean (15,946) for other sequenced stramenopiles (Fig. 1b). The number of introns and exons reported by the program Genome Annotation Generator (GAG) was 67,332 and 85,813 respectively, with an average exon length of 228 and intron length of 208.

**Genome ploidy and evidence of sexual reproduction**

As mentioned previously the N50 of the genome assembly was much improved by the use of Platanus - an assembly algorithm optimised for multi-ploidy genomes. To further investigate evidence of ploidy in our *H. catenoides* culture we mapped ~101 million reads to the 65.7 Mbp assembly identifying 1,393,505 single nucleotide polymorphisms (SNPs) with 1,332,610 (96%) of the SNPs identified consisting of a two-way nucleotide polymorphism (i.e. 58.81/41.19% mean character split). We also took all scaffolds and plotted SNP frequency against scaffold size demonstrating the majority of the scaffolds are clustered around a SNPs frequency of ~0.0275 (Fig. S4) suggesting that this variation is consistent and not specific to a subset of chromosomes for example in the case of an aneuploidy. Using Jellyfish (33) to map kmer coverage and the web server GenomeScope (<http://qb.cshl.edu/genomescope/)> we showed the presence of two peaks in the coverage frequency, which is also consistent with the reads mapping to a diploid genome (Fig. S5). Taken together these data are consistent with the *H. catenoides* sampled for sequencing containing a diploid variant during its complex and multifaceted life-cycle (6, 20).

We also checked the predicted proteome for evidence of the seven meiosis specific gene families previously shown to be present across a wide diversity of eukaryotes (34). Using reciprocal BLAST searches, we found evidence that genes putatively encoding all seven meiosis specific genes were present on the assembled genome and that these genes were also present in the transcriptome gene set and therefore likely expressed in the culture conditions used to grow *H. catenoides* prior to sequencing. These include SPO11 (the *H. catenoides* putative homologues are available at EMBL accession number XXXXXX), HOP1 (XXXXXX), HOP2 (XXXXXX), MND1 (XXXXXX), DMC1 (XXXXXX), MSH4 (XXXXXX), MSH5 (XXXXXX). Sexual recombination has rarely been observed in Hyphochytriomycota cultures, with only Johnson (1957) identifying cellular forms suggestive of zygote production as a result of fusion in the resting spore development of *Anisolpidium ectocarpii* (35). A range of different sexual reproductions systems have been identified in the oomycetes (36) collectively these data suggest meiosis is present across the Pseudofungi.

**Phylogenetic position of *Hyphochytrium* and traits shared across the Pseudofungi.**

*Hyphochytrium* has been shown to branch as a sister-group to the oomycetes in rRNA gene phylogenies (e.g. (3, 21)). To test this phylogenetic relationship, we generated a concatenated amino acid alignment of 245 putative orthologues (85 taxa and 71,862 amino acid sites) and including a comprehensive sampling of eukaryotic taxa based on previously published papers (37, 38).  The resulting tree topology (Fig. 1a & see XXXX for full tree file) demonstrated that *H. catenoides* forms a sister-branch to the oomycete radiation with 100% bootstrap support consistent with the Pseudofungi hypothesis, i.e. the hyphochytridiomycetes are a sister-group to the oomycetes and therefore share a common evolutionary trend towards fungal-like osmotrophic feeding and polarized cell growth (3, 4). Our tree places the Pseudofungi as a sister group to the photosynthetic stramenopiles (i.e. the Ochrophyta) consistent with previously published phylogenetic analysis based on three nuclear encoded genes (39) and wider phylogenomic analysis of 339 protein sequences (8) and in contradiction to analyses of mitochondrial gene phylogenies (concatenation of ten genes, 7,479 positions) which have demonstrated that Labyrinthulida (i.e. Bigyra) form a sister-group to the oomycetes (40). The tree recovered here is broadly similar to that reported by Derelle *et al.* which uses a large phylogenomic dataset but with a different taxon set. This work convincingly argued for monophyly of Bigyra (e.g. *Blastocystis + Aplanochytrium* and *Schizochytrium*) although our tree shows that this group is paraphyletic. Derelle *et al.* also recovered the paraphyly result in a subset of their Bayesian analysis and in their ML analysis but demonstrated this is likely due to a long branch attraction artefact (e.g. (41)) associated with the *Blastocystis* branch leading to the misplacement of Opalozoa (e.g. *Blastocystis*). Interestingly, sisterhood of the Pseudofungi and Ochrophyta implies a minimum of two losses of photosynthesis (8) and independent specialization of ‘osmotrophic lifestyles’ in the Bigyra (e.g. *Aplanochytrium* and *Schizochytrium*) and the Pseudofungi (e.g. *Hyphochytrium* and *Phytophthora*)clade within the stramenopiles. However, this scenario implies that the stramenopile lineage was ancestrally photosynthetic (9) a subject of debate (13), which is discussed further below.

**Shared derived traits across the Pseudofungi**

Given the placement of *H. catenoides* as a sister branch to the oomycetes, we were interested to investigate the conservation of cellular, biochemical and genetic traits shared across these Pseudofungal taxa*.* Many oomycetes (e.g. *A. laibachii, H. arabidopsidis* and *Phytophthora spp*) have lost the genes for thiamine biosynthesis while others (e.g. *P. ultimum* and S*. parasitica*) have retained the genes that putatively encode the biosynthesis of this vitamin (42). Reciprocal BLASTp searches identified the presence of a gene putatively encoding thiamine biosynthesis present on the *H. catenoides* genome. This gene putatively encodes a multi domain protein (Hypho2016\_00001922 - BN6634\_16369) composed of domains annotated as thiamine-phosphate pyrophosphorylase and thiamine kinase.

Oomycete plant pathogens, for example *Phytophthora* sp., are sterol auxotrophs and appear to have lost many of the enzymes involved in sterol biosynthesis (43). Sterol biosynthesis pathway has been predicted to function in *Saprolegnia* and a putative CYP51 sterol-demethylase encoding gene was identified from the *Saprolegnia* genome and transcriptome data (19, 44). The protein encoded by this gene is, for example, a target of antimicrobial drugs such as clotrimazole and therefore has been suggested as therapeutic target for treatment of *Saprolegnia* infections of fish (45). Reciprocal BLASTp searches and phylogenetic analyses demonstrated that *H. catenoides* also possesses a putative orthologue (Hypho2016\_00003038 (BN6634\_2538) - Fig. S6a) of the *Saprolegnia* CYP51 sterol-demethylase which appears to be lost in plant pathogenic oomycetes. To confirm that this is a viable drug target we grew *H. catenoides* in the presences of two azole antifungals - clotrimazole and fluconazole – to assess effectiveness of these compounds in inhibiting *H. catenoides* growth. Both antifungal agents were able to inhibit growth of *H. catenoides* (MIC100: clotrimazole 0.25 µg ml-1; fluconazole 4 µg ml-1  - Fig. S6b), indicating that the *H. catenoides* is susceptibility to azole compounds consistent with hypothesis that *H. catenoides* has a functioning CYP51 protein.

There has been considerable effort to sequence a number of oomycete genomes which has largely focused on parasitic taxa (e.g. (18, 19, 46-49). This work has also in-part focused on identifying candidate effector proteins (secreted proteins that perturb host function for the benefit to the invading parasite (50) and which often contain N-terminal RxLR amino acid motifs (51-53)) or lectin proteins that bind host molecules. Searches of the *H. catenoides* genome demonstrate there is only one putative protein of unknown function with a candidate RxLR motif (Table 1). In addition, *H. catenoides* lacked several gene families linked with the evolution of plant pathogenic traits in the oomycetes, i.e.: NPP1 or NEP-like (necrosis inducing *Phytophthora* protein (54, 55)), Elicitin proteins (56), Cutinase (57), pectin esterase and pectin lyase (58, 59). While the animal parasite *Saprolegnia parasitica* was noted to show enrichment of Notch proteins and Ricin lectins and presence of other galactose-binding lectins and the bacterial toxin-like gene family (haemolysin E) (19). While the Notch protein and Ricin lectin gene families are present in *H. catenoides* they show no evidence of enrichment comparable to *S. parasitica* and the galactose binding lectin and haemolysin E gene families are absent. Protease gene families show no general enrichment in comparison to other stramenopiles (Table 1).

Comparative secretome analysis demonstrated that*H. catenoides* contains a lower proportion of secreted proteins compared with other stramenopiles, comparable with the paraphyletic obligate biotrophs *Albugo laibachii* and *Hyaloperonospora arabidopsidis* (Fig. 2)*.* The *H. catenoides* predicted proteome contains a moderate-to-low proportion of carbohydrate active enzymes when compared to other stramenopiles (60). Interestingly, *H. catenoides* has very few secreted carbohydrate active enzymes as compared to all other stramenopiles, suggesting that *H. catenoides* has a low diversity of extra-cellular carbohydrate processing functions and is therefore dependent on a limited subset of extracellular fixed carbon sources (Fig. 2). To test this observation, we grew *H. catenoides* cultures in 190 different carbon sources using OmniLog PM1 and PM2 plates (61). These data demonstrated that significant increase in growth rate was observed upon the addition of only α- or β- cyclodextrin (*p* value = 0.01 & 0.01), dextrin (*p* value = 0.02), Tween 40 or 80 (*p* value = 0.03 & 0.03), or melibionic acid (*p* value = 0.03). Of note, dextrin/cyclodextrins are products of enzymatic activity upon starch, a typical component of *H. catenoides* growth medium (YpSs), and may be indicative of the environment in which this organism is typically found. The addition of Tween 40 or Tween 80 has been shown to improve yield in other organisms (62) and may result from direct accumulation of fatty acids, or altered membrane permeability affecting nutrient uptake. In contrast to many oomycetes (ref) *H. catenoides* demonstrates a limited utilisation of diverse carbon sources. These data are consistent with the hypothesis that the evolution of a wide diversity of secreted carbohydrate active enzymes is associated with evolution of parasitic lifestyle within the oomycete lineages (e.g. (19, 26, 63, 64)) although this pattern could also be the product of secondary loss in the *H. catenoides* lineage.

Seidl et al. have demonstrated a significantly higher number of genes encoding unique protein domain combinations in the plant pathogenic oomycetes compared to fungi with similar lifestyles (65). This analysis identified 53 domain architectures that were unique to and conserved across *P. infestans, P. ramorum, P. sojae* and *H. arabidopsidis.* Domains are often recombined by a process of gene fusion and/or domain ‘shuffling’ (66), gene fusion characters although subject to sources of homoplasy such as gene fission (67), can represent synapomorphic traits, i.e. characters that define evolutionary groupings. We searched the *H. catenoides* genome for evidence of the 53 multi-domain gene fusions previously identified as conserved in *P. infestans, P. ramorum, P. sojae* and *H. arabidopsidis* (65). We found that 12 of these 53 domain architectures were also present in *H. catenoides,* while searches of eukaryotic genomes and searches of the GenBank nr database confirmed this domain combination was unique to, and present across, the Pseudofungi (Table S3). Of note, we found a fusion of a putative callose synthase enzyme domain and a putative membrane transporter gene present across the Pseudofungi, suggesting that domain fusion has led to a unique coupling of substrate transportation and enzymatic processing and that this modification occurred prior to the radiation of the Pseudofungi.

Using OrthoMCL (68) combined with a custom pipeline developed in our lab, (see https://github.com/guyleonard/orthomcl\_tools & <http://dx.doi.org/10.5281/zenodo.51349)>), which maps putative orthologues classified from 74 genomes (Table S4) to a species phylogeny, we identified candidate orthologue gene sets specific, given current genome sampling, to the Pseudofungi (Table S5). These were confirmed as Pseudofungi specific by BLAST searches of the GenBank nr and EST databases. This analysis identified six gene families that putatively encode Pseudofungi specific domain combinations, five of which were unique and one that overlapped with the data reported above (see (65) and Table S3). The analysis of Pseudofungi specific orthologues and comparison of conserved novel domain protein domain architectures reported in Seidl et al. (2011) identifies four group-specific protein domain architectures encoding an AGC family Serine/Threonine Kinase domain. These results demonstrate a novel diversification in this domain family and functional Serine/Threonine Kinase repertoire in the Pseudofungi. Indeed, previous work has demonstrated large expansions of kinase families present in oomycete genomes (19) (Table 1).

**Evidence of viral integration through-out the Pseudofungi**

The comparative genomic analysis of Pseudofungi specific orthologues described above identified that *H. catenoides, Phytophthora cinnamomi, Phytophthora parasitica* and *Pythium ultimum* putatively encode viral major capsid protein (MCP). These genes have high sequence identity to each other and branch with African swine fever virus (MCP proteins) on a phylogenetic but which are highly divergent when compared with other viral MCP proteins (Fig. 3a+b). To further explore if the *H. catenoides* genome assembly contained additional viral-like genes symptomatic of additional viral-like genes, identifying 45 putative genes of viral ancestry, 23 were present on one scaffold and 15 present on a second scaffold (Table S6). All of these genes showed highest similarity to large DNA viruses, i.e. nucleocytoplasmic large DNA viruses such as Mimiviridae, Marseilleviridae, Phycodnaviridae, Asfarviridae and Poxviridae (Table S6). The genes of viral ancestry were found in linkage with genes of *H. catenoides* provenance. To confirm the genome assembly, we designed PCR primers nested to cross both the 5’ and 3’ sides of the *polB*, *mcp*, *mg96* putative genes and sequenced the product confirming the genome assembly (Table S7). The genome assembly demonstrated that the viral *mcp* gene was on the same DNA contig. as a putatively native *H. catenoides* histone encoding gene. To confirm this assembly and linkage between ‘host’ and viral gene we conducted a bridging PCR resulting in an amplicon of 2,837 bp and sequenced this amplicon confirming that the *mcp* and histone genes are linked and on the same stretch of DNA (Table S7). To investigate if the virus is active, we conducted RT-PCRof the *polB, mcp*, *mg96* and *rps3* virus confirming *polB, mg96 and rps3* are expressed in our culture conditions suggesting that the virus is transcriptionally active (Fig. 3b)*.*

Lack of transcription of the mcp encoding gene suggests that a complete virus or a viral factory is not currently being manufactured in the culture conditions tested (Fig. 3), although we note that the MCP protein was present in our transcriptome sequencing data conducted from a separate culturing experiment. [Office1] Electron microscopy also failed to observe icosahedral structures typical of viral particles or an intracellular viral factory (69) (see Fig. S7). [Office2] These data combined with evidence of viral gene integration in the oomycetes suggests a hitherto un-sampled diversity of large DNA viruses found infecting, or integrated within the genome of, Pseudofungi, consistent with other data which suggests the Pseudofungi have been subject to viral transduction/integration (70).It has also been shown that many different lineages of the stramenopiles have similarly retained fragments of viral genomes (ref) suggesting a wider under sampled diversity of stramenopile large DNA viruses. It is tempting to speculate that this may be a mechanism driving HGT seen in the oomycetes (see (70)), given that large DNA viruses have been shown to be vectors for host and foreign genes (ref).

**Organelle genomes in the *H. catenoides* assembly?**

Using common mitochondrial marker genes as seed sequences for tBLASTx searches we searched the *H. catenoides* for a mitochondrial genome demonstrating the organelle genome assembly was highly fragmented. By searching both the Platanus and an alternative Ray assembly (71) we identified and assembled, two mitochondrial genome contigs of 42,239 bp and 6,915 bp using Sequencher (72). We then used multiple primers for targeted PCR amplification and joined the two fragments of the mitochondrial genome on one side resulting in a linear fragment of 48,124 bp in size (Fig. S8). The *H. catenoides* mitochondria is reported to be circular based on mapping analyses (73). We used multiple combinations of PCR primers to try and join the other side into a circular genome, but could not amplify a bridging region. Other pseudofungal mitochondrial genomes have also proved difficult to complete, likely due to secondary structural factors (e.g. (74)). Annotation showed that the mitochondrial genome contained 38 protein genes, 25 transfer RNAs and 2 ribosomal RNA encoding genes arranged in an inverted repeat of the ribosomal RNA region (73). A mitochondrial genome size of 54 Kbp and the presence of a ~14 Kbp inverted repeat had previously been identified using physical genome mapping and is a feature present in some oomycetes (73, 74). The genome assembly also demonstrated the presence of the atp1 gene encoded in both the mitochondria (XXXXX) and the nucleus (Hypho2016\_00005578 - BN6634\_14994). This gene is found in the nucleus of a range of other stramenopiles including oomycetes, Eustigmatophyceae, and *Cafeteria* (74, 75). Phylogenetic analysis revealed that the Atp1-like protein encoded in the nucleus formed a monophyletic clade with similar proteins from other nuclear atp1 stramenopile genes (Fig. S9). TargetP v1.1 analysis demonstrate that all proteins from this clade, except the *Nannochloropsis* protein, have a putative N terminal mitochondrial targeting peptide. These results indicate that this clade may represent a stramenopile-specific nuclear transfer of mitochondrial Atp1 that has since diverged in sequence and possibly function.

Many stramenopile lineages harbour plastid organelles with an associated genome while others, including the oomycetes appear to have lost (9) or, alternatively never had (13) the organelle/genome. To search for the possible plastid genome, we took the proteome encoded by the plastid of *Chlamydomonas reinhardtii* (69 proteins) and *Thalassiosira pseudonana* (141 proteins) as seed sequences for a BLASTx search of the entire genome assembly (including contigs and scaffolds below 1 Kbp in size). The majority of the BLAST searches recovered no candidate plastid gene sequences. On a few cases candidate plastid genes were identified and searched back against the GenBank nr database. These BLAST searches demonstrated that the recovered were mitochondrial ribosomal proteins. This analysis, therefore, suggests that there is no plastid genome present in the *H. catenoides* genome assembly.

**Footprint of a plastid gene ancestry in the *H. catenoides* genome assembly?**

The established approach for identifying endosymbiotic ancestry of protist lineages that appear to have no endosymbiotic organelles is to identify genes of endosymbiotic ancestry present on the nuclear genome. This approach has worked well for identifying cryptic endosymbiotic ancestry for both the plastid (76, 77) and mitochondrial (78-80) organelles. For genes derived from the endosymbiont that gave rise to the mitochondria these genes should show eukaryotic orthologues branching with or within a cluster of alpha proteobacterial genes. For genes derived from the endosymbiont that gave rise to the plastid organelle these genes should show eukaryotic orthologues branching with or within a cluster of cyanobacterial genes. Using a phylogenomic pipeline described in the methods, we constructed a phylogeny for all gene clusters which included both a *H. catenoides* and a gene of a photosynthetic eukaryote and included a wide sampling of similar genes sampled from both eukaryotic and prokaryotic taxa. This process resulted in 8970 preliminary phylogenetic trees, which we searched for a phylogenetic relationship that demonstrated *H. catenoides* genes branching with photosynthetic eukaryotes or eukaryotes with photosynthetic ancestry which in turn branched within a cluster of bacterial sequences. Following the example of others (81, 82) we reasoned that cyanobacterial signature was likely mixed up by horizontal gene transfer in the prokaryotes and so any *H. catenoides* gene cluster that showed the relationship described above (i.e. [[*H. catenoides* + Photo-Euk]bacteria]]) potentially represented a relict gene ancestrally derived from the endosymbiosis that gave rise to the plastid. For 101 candidate phylogenies that showed this relationship, the alignment was edited and masked manually and the taxon sampling checked using additional BLAST searches of NCBI databases before the phylogeny was recalculated. This process identified four gene families where *H. catenoides* protein sequences (in three cases branching with other Pseudofungi/Stramenopiles taxa) branched with photosynthetic eukaryotic taxa sister to bacterial taxa. In all four cases (Fig. S10a-d) we did not detect a putative homologue of these gene families in eukaryotic taxa thought not to have had a plastid endosymbiotic ancestry (e.g. Fungi, Metazoa and Amoebozoa). In only one case did the eukaryotic sequences branch with the cyanobacteria, although the bootstrap support for this was moderate (Fig. S10d). We then used TargetP (83) searches to investigate if the putative orthologues of the *H. catenoides* protein sequence encoded by Archaeplastida photosynthetic eukaryotes had putative plastid targeting N-terminal peptide, demonstrating that none of these protein sequences contained such evidence. Interestingly, TargetP identified both the *Arabidopsis thaliana* and *Oryza sativa* putative Heme oxygenase domain containing protein (Fig. S10d) had evidence of a N-terminal mitochondrial targetingpeptide.

Taken together these data suggest that the *H. catenoides* protein sequences contain a very small footprint of genes that could be putatively of plastid associated endosymbiotic ancestry but could also represent distinct patterns of ancestry decoupled from plastid endosymbiotic ancestry. It is likely that our survey protocol could miss a number of genes of endosymbiotic ancestry. This is because pipeline generated trees often have inappropriate taxon sampling (i.e. to large) and rarely show significant resolution. To partially control for this problem, we reran many of our candidate genes, where the resulting alignment was very large, using a subset of genome sampling. It is difficult to conclude if four gene families identified here, which in all cases do not show a strongly supported cyanobacterial ancestry for the eukaryotic genes, confirm a plastid ancestry for the Pseudofungi. Indeed, others have argued that similar numbers of genes which show potential endosymbiotic ancestry do indeed indicate a plastid endosymbiotic ancestry for different protist groups that branch with photosynthetic protists but have no evidence of a plastid organelle (76, 77). It is difficult to prove a negative in this case because it is certainly possible that the majority of plastid/cynobacterial derived lineages have been lost while any genes retained may be so small in number and with hugely complicated phylogenetic histories that it would be near impossible to recover using a pipeline tree building methods. As such these results are inconclusive with regards to photosynthetic/plastid ancestry of *H. catenoides*.

**Remodelling of the heterokont in *Hyphochytrium*.**

The stramenopiles (also known as Heterokonta i.e. meaning possessing of two unequal flagella) were formally described as a phylum based on the presence of two motile flagella, a ‘standard’ smooth posterior flagellum and an anterior flagellum with tripartite rigid tubular mastigonemes (“hairs” or described as ‘tinselate’) (2). The stramenopile phylogenetic group includes subgroups that have lost one or both of the flagella (2), for example the zoospore of *Hyphochytrium* builds a single, anterior ‘tinselate’ flagellum (6) while oomycetes build the standard stramenopile flagella pair. The phylogenetic placement of the *Hyphochytrium* (Fig. 1a) therefore pinpoints a loss of the posterior smooth flagellum in the ancestor of the hyphochytrids. To explore the consequence of the loss of this organelle in *H. catenoides*, in terms of gene/protein repertoire, we used a comprehensive list of proteins putatively associated with flagellum function (85) to survey the *Hyphochytrium* genome. This list comprises 592 amino acid sequences, 355 are found in both the Opimoda and Diphoda (86) major phylogenetic groupings suggesting they are universal flagella proteins (Table S8 & Fig. 4A). 330 of the 355 universal flagella proteins (UFPs) are also present in the predicted proteome of *H. catenoides*, suggesting that the majority (93%) of the UFPs have been retained and likely encode a function associated with the anterior tinsel flagellum consistent with the hypothesis that this organelle is a modified homologue of the posterior smooth flagellum.

Flagellum specific proteomic analysis of the stramenopile brown alga *Colpomenia bullosa* identified fourteen proteins specific to the posterior flagellum and three specific to the anterior flagellum (85). Homology searches suggested that the three *C. bullosa* anterior flagellum proteins are also conserved in *H. catenoides,* while twelve of the fourteen posterior ‘specific’ flagellum proteins from *C. bullosa* were also present, suggesting these proteins function in the anterior-tinsel flagellum in *H. catenoides* (Fig. 4a)*.* One of the *C. bullosa* posterior specific flagellum proteins absent in *H. catenoides* and the oomycetes include the PAS/PAC sensor hybrid histidine kinase (CBJ26132.1), a putative photo-sensor associated with a swelling in posterior flagellum of brown algae (85), discussed further below.

Twenty-nine of the flagellum proteins were present in the oomycetes and other eukaryotic groups but absent in the *H. catenoides*. This constitutes 8% of the 355 UFPs, which may represent genuine losses or absences, although this is a draft genome so absences may be due to incomplete genome sampling and assembly. If these genes are genuinely absent from the *H. catenoides* genome these results suggest a recent loss, putatively consistent with loss of the posterior flagellum without the function of these proteins being transferred to the anterior-tinsel flagellum (Fig. 4a)*.* Interestingly, these losses include a putative homologue of the DRC1 (P0DL09.1) subunit of the nexin-dynein regulatory complex which functions to regulate dynein motors in *Homo sapiens* and *Chlamydomonas reinhardtii* (87) and Radial Spoke protein 7 (RSP7) a protein that functions in flagellum structure and beating in *Chlamydomonas reinhardtii* (88).Further, analysis of the radial spoke protein repertoire putatively encoded in the *H. catenoides* genome identified that the anterior flagella radial spoke complex had a number of genes missing, like many other eukaryotes, but that RSP7 was the only radial spoke proteome loss specific to the loss of the posterior flagellum in the *Hyphochytrium* lineage (Fig. 4a/b), specifically this protein is putatively encoded in the oomycetes but has been separately lost in the Opisthokonta (e.g. *Homo sapiens*)*.* Collectively, these data suggest the *Hyphochytrium* anterior flagellum has a highly reduced flagellum spoke complex.In *Chlamydomonas* (88) both RSP11 and RSP7 have been shown to encompass a RIIa domain and putatively interact with the outer doublet microtubule end of the radial spoke through binding an AKAP domain and indeed both proteins have been shown to be located proximate to RSP3 at in the stalk of the radial spoke complex (89). An association of RIIa domain and an AKAP domain at the spoke stalk seems critical to flagellum function. Interestingly, comparative analysis suggests that neither RSP7 or RSP11 are conserved across flagellum bearing eukaryotes with only *Chlamydomonas*, *Batrachochytrium* and *H. catenoides* retaining RSP11 in our comparative dataset (Fig. 4a/b)*.* Domain analysis of the putative RSP3 and RSP11 confirmed these proteins contain an AKAP and a RIIa respectively and suggesting that the anterior flagellum of *H. catenoides* is dependent on this specific protein-protein interaction to join the radial spoke to the outer doublet.

**Photoreceptors**

Several stramenopile species have been shown to encode a range of photoreceptor proteins and to initiate a series of responses to light including phototaxis (90). Specifically, the zoospores of some stramenopile algae can show positive and negative phototaxis (91) associated with a flavoprotein photoreceptor (92) putatively the helmchrome located in the posterior flagellum (85) and associates with ‘flagellar swelling’ and a stigma (91). *H. catenoides,* has lost the posterior flagellum along with the flagellum swelling (6). Consistent with this loss the putative photosensitive response system is also absent (Fig. 4a)*.*

A number of additional putative photo-responsive proteins have also been reported from *Ectocarpus* (17). Using these data and other seed sequences (e.g. (85, 93)) we searched the *H. catenoides* genome for putative homologues of these and other genes encoding photo-responsive protein. Reciprocal blast searches demonstrated that the *H. catenoides* genome contained putative homologues of the flavoproteins Cryptochrome (XXXXX), Cryptochrome DASH (XXXXX) and Photolyase (XXXXX) gene families like other stramenopile species (Fig. 5A)*.* The transcriptome data demonstrates that these genes are transcribed. In addition, these analyses identified three putative type I (microbial) rhodopsins (XXXXX, XXXXX & XXXXX) in the assembly, the first putative representative of this gene family from a stramenopile (Fig. 5a/b). The putative type I rhodopsins identified showed variation in the amino acid alignment, however the 11-*cis*- retinal binding pocket, specifically the Schiff base is conserved (Fig. 5a/b). Furthermore, reciprocal blast searches of both the genome and the transcriptome sequence datasets confirmed the presence of genes putatively encoding the latter two steps of the retinal biosynthesis pathway (*β*-Carotene- 15, 15’Dioxygenase (XXXXX) and 11-*cis*-retinol dehydrogenase (XXXXX) the steps in the pathway that converts the vitamin *β*-Carotene into 11-*cis*-Retinal, the critical cofactor for rhodopsin function as a light responsive protein. These data suggest *H. catenoides* is capable of light triggered development, physiological responses and/or phototaxis.

**Gene families encoding hallmarks of fungal characteristics in pseudofungi?**

One aim underpinning the motivation for sequencing the *H. catenoides* genome was to investigate the genes / phenotypes that underpin the fungal lifestyle, which may also be present in the *H. catenoides* genome and other pseudofungi. Most fungi grow as filamentous cells, reinforced by robust cell walls composed of sugar polysaccharides such as chitin. A suite of cellular systems allowing fungi to grow as polarised cells, laying down cell wall and feeding on extracellular substrates by a combination of exocytosis of enzymes and cell wall material and endocytosis and transporter mediated uptake of target nutrients. Fungal filamentous structures such as hyphae grow almost exclusively from the tip of the hyphal structure (94) allowing fungi to ‘grow as they feed.’ This feature combined with a robust cell walls means they can generate high turgor pressures, ramify into recalcitrant material, feed osmotrophically and maximize metabolic rates (95, 96)[book chapter]. Homologous cellular systems also drive bud growth in *Saccharomyces cerevisiae* allowing researchers to use *S. cerevisiae* to study proteome function involved in polarized growth in fungi (for review see (94, 97)). The proteins that are known to control this system are illustrated on Fig. 6a and involve key complexes the Exocyst and the Polarisome. These systems are therefore very important for establishing the temporal and spatial control of polarized cell growth in fungi. Comparative analyses show the Exocyst system and Sec4 orthologues are conserved across a diversity of eukaryotes including *H. catenoides*, while the Polarisome and associated proteins seem to be specific to fungi (given current taxon sampling) and demonstrating that specific elements of polarized cell growth control are not present in pseudofungi, suggesting these filamentous microbes accomplish polarized growth using a combination of different cytoskeleton proteome functions.

Like fungi (98) and many other eukaryotes (99-105), *H. catenoides* also produce chitin as cell wall material (106). Oomycetes have also been shown to produce chitin in their cell walls (107). This is consistent with previous data that suggest that chitin synthesis and deposition as a cell wall material predates the diversification of many major lineages of the eukaryotes (106). Interestingly, *H. catenoides* has a similar repertoire of chitin synthesis and digestion as found in the oomycetes (i.e. chitin synthase division I gene family), while another group of stramenopiles, the diatoms, which also produce chitin (108), have a variant chitin gene repertoire, namely chitin synthase division II and a chitinase (GH19) not present in pseudofungi (Fig. 6). This suggests that chitin production as a cell wall component is universal and ancient, but the genes that control the synthesis and remodelling of this structural sugar have been reconfigured numerous times. Specifically, pseudofungi seem to lack all chitin synthase division II which are numerous and diversified in the fungi, suggesting another key difference, along with absence of the polarisome, which contrast filamentous growth mechanisms between fungi and pseudofungi.

**Conclusion**

We provide a draft genome of the free-living stramenopile pseudofungus *Hyphocytrium catenoides* as a reference for comparative biology specifically with a view to understanding the evolution of filamentous growth and osmotrophic feeding. *H. catenoides* branches sister to the oomycetes that contains many important pathogenic groups. These data demonstrate that *H. catenoides* does not encode many of the gene families found in oomycetes that have been associated with pathogenic function, suggesting that these characteristics are more recent adaptation/acquisitions within the oomycetes. Our data also demonstrates that *H. catenoides* and the Pseudofungi more widely possess a range of features associated with filamentous growth and osmotrophic feeding in the fungi. These include the Exocyst vesicle trafficking system, sterol biosynthesis pathway, and a repertoire of chitin cell wall synthesis systems common to fungi. In contrast, Pseudofungi have lost or never possessed a Polarisome complex and lack chitin synthase II and chitinase I, identifying clear differences between these two filamentous osmotrophic groups.

*H. catenoides* also provides a useful model for understanding the diversification of eukaryotic flagella, as it is the only sequenced organism with a single anterior tinsel flagellum and the genome sequence reported, and which allows us to reconstruct what elements of the flagellum proteome are retained in a system where the posterior flagellum has been lost, demonstrating, for example, that *H. catenoides* tinsel flagellum has encompass an atypical flagellum spoke complex.

Our genome analysis also identified evidence of a viral integration into the *H. catenoides* genome, interestingly this analysis also showed evidence of viral MCP protein present within multiple oomycete genomes suggesting the Pseudofungi have been exposed to multiple large DNA virus infections and subsequent gene introgression and therefore identifies a possible mechanism driving HGT in these microbes that cannot perform phagotrophy, a mechanism through to be important for HGT in eukarhyotes (109). We hope this data will provide an important resource stimulating further comparative genomics analysis.

**Materials and Methods**

**Cell Culture in preparation for sequencing.**

*Hyphocytrium catenoides* (ATCC 18719) was inoculated onto Emerson YpSs agar and cultured and cell mass prepared for DNA and RNA extraction as described previously (26). DNA samples were checked for contamination using an environmental DNA (eDNA) SSU PCR approach (26) using both eukaryotic 18S PCR primers 1F (5’CTGGTTGATCCTGCCAG-3’) and 1520R (5’-CTGCAGGTTCACCTA-3’) (e.g. (110) and prokaryotic 16S PCR primers PA (5’-AGAGTTTGATCCTGGCTCAG-3’) and PH (5’-AAGGAGGTCATCCAGCCGCA-3’) (e.g. (111).

**Genome and Transcriptome Sequencing, Assembly and Validation and ORF Calling**

One lane of paired-end (100 bp) Illumina HiSeq data was generated along with two lanes of paired-end (76bp) Illumina GAiix at the Exeter Sequencing Service generating 2x 212,760,559 HiSeq reads along with 2x 15,266,599 and 2x 16,274,715 GAiix reads. At the time of sequencing we were unable to assemble this much data. Therefore, after trimming and cleaning (using TagCleaner and PRINSEQ) of the data, we subsequently digitally normalised it with KHMER in order to discard redundant data, sampling variation and remove errors. This reduced the number of reads to 415,241,668 HiSeq along with 28,964,302 and 30,961,514 GAiix; a reduction of 13,436,262 reads in total. The raw reads are deposited in NCBI and EBI with accessions as: Illumina GAiix = SRX033129 and Illumina HiSeq = ERS1151585 respectively. An initial assembly, using the program Ray v2.2.1 (112) was generated using command detailed in supplementary file X, and generated 29,448 scaffolds, with a total of 107,387,882bp, and an N50 of 8,746bp. Next we investigated the possibility that *H. catenoides* was a diploid using an assembly program that allows for multiple ploidy. The program Platanus v1.2.1 was used to produce an assembly with 53,358 scaffolds incorporating 68,330,525bp, and an N50 of 29,450bp.

The Platanus assembly was subsequently filtered into 4 datasets; all scaffolds, scaffolds >= 10Kbp, scaffolds >=5Kbp and scaffolds >=1Kbp, see Fig. S1, in order to test the effects or the N50 statistic and gene recovery rate by removing short and erroneous scaffolds/contigs. We determined that the set of scaffolds >= 1Kbp did not affect our predicted proteome complement and increased the N50. The filtered >=1Kbp Platanus assembly, along with the mitochondrial assembly, are deposited in EBI with the accessions; Study ID: PRJEB13950, Scaffolds: FLMG01000001-FLMG01004758 and Mitochondria: LT578416. Further data, such as the full assembly and other filtered datasets, can be accessed here: <https://github.com/guyleonard/hyphochytrium>.

Gene prediction was conducted by using CEGMA to predict which of the 246 core genes are present in our *Hyphochytrium* >=1Kbp scaffolds, these (113) predicted CEGs are then used in the training step of the program SNAP (see http://korflab.ucdavis.edu/software.html) to generate a set of *ab initio* gene models. The program GeneMark-ES (114) is also run independently on the >=1Kbp scaffold data which produces another set of gene models. Both these sets of gene models are in the form of a hidden Markov model (HMM). A first pass of the pipeline MAKER is then run with the default settings, incorporating the gene models from SNAP & GeneMark-ES whilst also deriving alignment statistics from the 454-transcriptome assembly with tBLASTn, repeatmasker and exonerate. The output is a set of gene models in GFF3 format. A second round of SNAP is then performed with the new predictions (after the GFF3 has been converted to a HMM) and the program AUGUSTUS (115) is run in *ab initio* mode using the MAKER first pass predictions (i.e. AUGUSTUS is not run against any 'similar/related' taxa). Both outputs of SNAP (run 2) and AUGUSTUS are then fed back into MAKER for a second run with stricter settings (gene predictions are available here: <https://github.com/guyleonard/hyphochytrium/tree/master/gene_predictions>). The final output is a GFF3 file, transcripts and protein FASTA files. The resulting gene predictions were then BLAST searched against the SwissProt database along with InterproScan to assign functions. The results were then used with the program ANNIE (116) to provide the correct format of annotation information to the program GAG (113) which is used to convert genome information (scaffold nucleotides, gene prediction, annotation, etc) for database deposition. The resulting genome data is therefore submitted as an update of a prior BioProject sequence submission (26), to do this we used the ‘gff3toembl’ program from PROKKA (117)

Previously we had sequenced a transcriptome from the same culture strain of *Hyphochytrium* (26) using 454 FLX sequencing of cDNA reads and assembled it with Newbler 2.5 (ref) using the default cDNA settings. We cleaned 70 sequences from this assembly by removing contigs of less than 100 bp in length (excluding the polyA regions) and/or contigs that consisted of predominately repeat motifs. This resulted in 6,202 transcript sequences with an N50 of 1,044 bp. The transcriptome assembly is based on 454 FLX+ cDNA reads. Assembled in Newbler 2.5 using the standard settings for cDNA. The reads were also assembled in Trinity but resulted in significantly [nearly double] more contigs and a much lower N50 score.

***H. catenoides* genome qPCR size estimation**

50 ml of a *H. catenoides* culture, grown in YpSs for 7 days at 25°C, was centrifuged for 3 mins at 3200 x *g*. The supernatant was removed and genomic DNA was extracted from the remaining cells using a PowerSoil DNA isolation kit (MO BIO Laboratories). The genome size *H. catenoides* was estimated using a qPCR based method to quantify the haploid genome size (28). An *rps3* PCR standard was amplified using primers Hcat\_rps3\_F (5’-GAGGGCTACATGGTCAAGA-3’) and Hcat\_rps3\_R (5’-CCTTTGGCTCGATGATGGTG-3’). Each 25 μl reaction consisted of 0.5 U Phusion polymerase (New England Biolabs), 1x HF buffer, 400 μM dNTPs, 2 μM each primer and 1 μl *H. catenoides* genomic DNA (11.6 ng μl-1). Cycling conditions consisted of an initial denaturation of 5 mins at 98°C, followed by 30 cycles of 10 s at 98°C, 30 s at 61.0°C and 30 s at 72°C, then a final extension of 5 mins at 72°C. The 185 bp PCR product was purified by gel extraction (Thermo Scientific GeneJET Gel Extraction kit) and eluted using elution buffer. Concentration of the purified product was determined using a Qubit dsDNA HS assay kit (Thermo Fisher Scientific). Real-time PCR was used to quantify the number of copies of *rps3* present in each genomic DNA sample. Quantitative PCR was performed in a StepOnePlus Real-Time PCR system (Thermo Fisher Scientific). Reaction conditions were optimised using a gradient PCR and a standard curve was determined using dilutions of *H. catenoides* genomic DNA and analysed using StepOne Software v2.3 (slope: -3.367; y-intercept: 33.841; efficiency: 98.15%). Each 20 μl PCR contained 10 μl PowerUp SYBR Green Master Mix (Thermo Fisher Scientific), 500 nM each primer (Hcat\_rps3\_F and Hcat\_rps3\_R, sequences as above) and 1 μl template DNA. Template was either *H. catenoides* genomic DNA or the PCR standard. Standards were diluted (10-1 to 10-7) from an initial concentration of 24.7 ng μl-1 and performed in triplicate, while three independent genomic DNA samples were run in quintuplicate. Cycling conditions were as follows: UDG activation for 2 mins at 50°C and DNA polymerase activation for 2 mins at 95°C, followed by 40 cycles of 15 secs at 95°C and 1 min at 60°C. ROX was used as a reference dye for analysis of CT values. Each reaction was followed by melt-curve analysis, with a temperature gradient of 60°C to 95°C at 0.3°C s-1, to ensure presence of only a single amplicon. The PCR standards were used to create a calibration curve (*y = 8x1010 x e-0.67x;* R2 = 0.99992); CT values from amplifications of genomic DNA templates were then applied to this curve and the ‘mass’ of a haploid genome was calculated (28). This value was then used to calculate the haploid genome size, using 660 g mol-1 as the mean molar mass of a base pair consistent with the AT proportion of the *H. catenoides* genome of XX%.

**Mitochondrial genome assembly**

As indicated above the genome of *Hyphochytrium catenoides* was sequenced using Illumina HiSeq with Illumina GAiix and assembled using both Ray and Platanus. Contigs of putative mitochondrial origin, from both assemblies, were identified by BLAST searches against the mitochondrial genome of *Phytophthora infestans* (NC\_002387.1). The contigs from the genome assemblies were visualized, linked and edited using the program SEQUENCHER (<https://www.genecodes.com>) resulting in two contigs (one of 42,239 bp and 6,915 bp). However, we were unable to circularise the genome using these two fragments, possibly because of the two repeat regions (73). Therefore, regions spanning the gaps in the mtDNA super-contigs were amplified by polymerase chain reaction (PCR) with primers specific to the flanking sequences. Purified PCR products were sequenced using Sanger chemistry (externally at XXXXX). This allowed the two contigs to be joined, however it still did not result in a circular genome. Genes were identified on the final assembly using a number of bioinformatics tools: ORFs were located using the GETORF tool from EMBOSS (118) and the MITOS program (119) based on a BLASTX-based similarity search. The translated products were identified by BLASTp. Genes coding for rRNAs and tRNAs were annotated using tRNAscan-SE (120, 121). All this information was then collated and used with the circular genome reconstruction package CGView (121). A blast comparison allowed us to compare the *Hyphochytrium* mitochondrial genome with three other stramenopile species: *Phytophtora infestans* [U17009], *Ectocarpus*  [xxxxxx], *Saprolegnia ferax* [NC\_005984], see Figure S7.

**Search for *H. catenoides* representatives of key Oomycete gene families**

Using Pfam searches (Pfam release 29.0) with default defined e-value cut-offs we searched the *H. catenoides* predicted proteome for: NPP1-like proteins (Pfam domain: PF05630), elicitin (PF00964), cutinase (PF01083), pectin methyl esterases (PF01095), pectate lyase (PF03211), polygalacturonase (PF00295) PAN lectin (PF00024), Ricin lectin (PF00652), Jacalin lectin (PF01419) Galactose binding lectin (PF00337), Legume lectin (PF00139), Legume-like lectin (PF03388), ABC transporters (PF00005), Protein kinase (PF00069 & PF07714), Notch protein (PF00066) and Haemolysin E (PF06109). In addition, the *H. catenoides* predicted proteome was searched against the MEROPS database <https://merops.sanger.ac.uk/> to identify putative protease inhibitors and proteases and the CAZymes analysis (122) toolkit (using Pfam) at <http://mothra.ornl.gov/cgi-bin/cat/cat.cgi?tab=PFAM1> to identify putative carbohydrate interacting proteins**.** Predicted proteins containing putative RxLR motifs and Crinkler domains were identified using the pipelines described in the following publications (123, 124).

**Secretome analysis**

Putatively secreted proteins were predicted using a custom pipeline (<https://github.com/fmaguire/predict_secretome/tree/refactor>) which predicts sequences as secreted by taking the intersection of those sequences predicted to have a signal peptide (via SignalP 4.1 (125)), no TM domains in their mature peptide (via TMHMM 2.0c (126, 127)), a signal peptide that targets for secretion (via TargetP (83)) and belonging to the extracellular ‘compartment’ (as predicted by WoLFPSort 0.2 (128)).  The CAZY database (129) was downloaded, converted into a BLASTDB and searched using the predicted proteome and secretomes using BLASTP with an expectation of 1e-5. Hit tallies were then summed, proportions calculated and data plotted in Python via the Pandas and Seaborn packages.

**Phylogenetic analysis of individual gene families**

Unless otherwise stated all phylogenetic analyses were conducted using the following protocols. Using BLASTp we used the seed sequence to identify putative homologues across a locally maintained database of eukaryotic and prokaryotic genome-derived protein datasets (Table. S9) with a gather threshold of 1e-10. The Multiple Sequence Comparison by Log-Expectation (MUSCLE) program (v3.8.31) (130) was used to produce a multiple sequence alignment for each set of proteins. Alignments were then manually corrected and masked in SeaView (version 4.2.4) (131). Sequences that required a high level of site exclusion (due to the sequence not aligning or not masking well) or where they formed long-branches in preliminary analysis were removed. The phylogenies were calculated using parallelised-PTHREADS RAxML (version 7.7 -Pfeiffer and Stamatakis 2010) with 1,000 (non-rapid) bootstrap replicates and using the substitution matrix and gamma distribution identified using ProtTest3 (version 3.2.1). In some cases, the invariant sites parameter was also included in the model (if indicated in the ProtTest3 analysis). Details of the alignment sampling and model parameters used are given in the accompanying figure legend.

**Multi-gene concatenated phylogenetic analysis to identify the branching position of *H. catenoides***

Using the methods established in (37) we built a concatenated amino acid alignment of 245 putative orthologues resulting in a masked data matrix of 85 taxa and 71,862 amino acid sites. This dataset encompassed a wide sampling of eukaryotes based on previous analysis (e.g. (37, 38)).  Using this alignment, we calculated a maximum likelihood (ML) tree using  RAxML (132) with 60 ML searches under PROTGAMMALGF model of sequence substitution. 1000 boostrap replicates were then calculated using the IQ-TREE software using the LG\_G4+C20+F substitution model.

**Identification of genes of plastid ancestry**

We used a sequence similarity search combined with phylogenetic reconstruction pipeline (e.g. (133)) to identify gene families which include *H. catenoides* genes which have a phylogenetic ancestry demonstrating the gene was derived from the cyanobacterial ancestor that gave rise to the plastid organelle. Initially we constructed a database of (xxx Table - Guyxxx) genomes and clustered these genes into putative orthologs using OrthoMCL xxxGuyxxx, finally retaining only the groups containing *H. catenoides* genes. Next we resampled sequences from a wider database of 1205 taxa (Table S9) using BLASTp searches (134) to recover up to three similar sequences from each genome using a gathering threshold of 1e-10. We then filtered these clusters to those only containing both a *H. catenoides* gene and genes from photosynthetic or ancestrally-photosynthetic eukaryotic taxa. These sequences were then aligned using MAFFT (135), masked using TRIMAL (136) and a phylogeny was calculated from the data matrix using FASTTREE2 (137). The resulting phylogenies were manually inspected for a phylogeny that showed *H. catenoides*/Pseudofungal/Stramenopile genes which: a) branched within the Archaeplastida radiation, b) branched with genes of photosynthetic eukaryotes and within a bacterial radiation, or c) branched with cyanobacterial genes. This process required re-running of the phylogenetic pipeline for many gene clusters either reducing gene sampling or removing long branch sequences. A subset of 101 gene cluster phylogenies putatively showed a phylogenetic relationship consistent with criteria *a-c* described above. The alignments from these clusters were then manually refined, the taxon sampling checked using manual BLAST searches of the NCBI nr database and phylogenies recalculated using the RAxML approach described above.

**Testing for CYP51 sterol-demethylase drug sensitivity**

Azole susceptibility was assessed using a modified protocol of Warrilow *et al.,* (2014). Briefly, fluconazole and clotrimazole were dissolved in dimethyl sulfoxide (DMSO) to a stock concentration of 25.6 mg ml-1. Dilutions were then made with DMSO to prepare 100x stock solutions. These stocks were diluted in PYG medium to a final volume of 5 ml, each containing 100 µl of *H. catenoides* liquid culture (grown in YpSs at 25°C shaking for 7 days) to achieve final azole concentrations of 256, 128, 64, 32, 16, 8, 4, 2, 1, 0.5, and 0.25 μg ml-1 and with control samples containing 1% (v/v) DMSO. Cultures were incubated, in triplicate, for 7 days at 25°C with 200 rpm shaking and MIC100 was scored manually by assessing for presence/absence of hyphal growth.

**OmniLog ‘Phenotype Microarrays’**

100 ml *H. catenoides* cultures were grown in PYG (1.25 g L-1 peptone, 1.25 g L-1 yeast extract, 3 g L-1 glucose) in baffled flasks, 25°C 170 rpm shaking, to minimise aggregation. Cells were recovered by centrifugation at 3200 x *g*, washed twice with water and re-suspended in PY-G (as above, no carbon-source) to a final concentration of approximately 1.5 x 103 cells ml-1. Cells were allowed to recover at 25°C with shaking for 30 minutes before Dye mix D (Biolog) was added to a 1x final concentration. 100 µl of cells was inoculated into each well of PM1 and PM2 carbon-source plates and incubated for 7 days at 25°C. Each growth assay was performed in triplicate from independent cultures.

OmniLog Phenotype Microarray outputs were analysed using OPM (Vaas *et al.*, 2013). Data were aggregated using the ‘opm-fast’ method, analysed using the A parameter (maximum value of OmniLog units reached) and tested by t-test. Significant p-values were extracted if they resulted in increased growth in comparison to the negative control well A01.

**Confirmation of viral integration on *H. catenoides* contigs and reverse transcriptase PCR of viral genes**

To confirm that the viral genes were assembled correctly and were resident in the *H. catenoides* genome fragments, PCRs across the 3’ and 5’ junctions of the putative viral open reading frame for three of the viral genes *polB*, *MCP*, and *mg96* were performed. 25 µl PCR reactions (1x Phusion HF buffer, 400 µM dNTP mix, 200 nM each primer, 0.5 U Phusion polymerase) were performed with the following cycling conditions: initial denaturation of 5 mins at 98°C, followed by 30 cycles of 10 s at 98°C, 30 s at 56-64°C and 1 min at 72°C, then a final extension of 5 mins at 72°C. These were purified using a GeneJET PCR Purification Kit or GeneJET Gel Extraction kit (Thermo Scientific) and sequenced to confirm that each product matched the expected amplicon. To confirm that the *mcp* gene was on the same contig as the histone H3 gene, we performed a PCR across these two genes (expected amplicon of 2837 bp) using the same conditions as above, except with an annealing temperature of 64 °C and with a 3 min extension. The PCR product was purified and A-tailed using *Taq* polymerase; then cloned using the StrataClone PCR Cloning Kit (Agilent Technologies). The resulting vector was sequenced using T3/T7 primers, with primer-walking to confirm the entire 2. 8 kb sequence.

To investigate if the viral derived genes are actively transcribed in ourculture conditions, we conducted RT-PCRof the *polB, mcp*, *mg96* and *rps3* virus confirming *polB, mg96 and rps3* are expressed in our culture conditions suggesting that the virus is transcriptionally active. RNA was extracted from *H. catenoides* using RNA PowerSoil Total RNA Isolation (MoBio). Residual genomic DNA was removed using RQ1 RNase-Free DNase (Promega) and Taq PCR was performed to confirm absence of DNA. Reverse-transcriptase PCR was then performed using a Qiagen OneStep kit according to the manufacturer’s instructions, alongside genomic DNA positive and no-template controls. The following cycling conditions were used: reverse-transcriptase of 30 mins at 50°C and initial denaturation 15 minutes at 94°C, followed by 32 cycles of 1 min at 94°C, 1 min at 50°C and 1 min at 72°C, then a final extension of 10 mins at 72°C. Samples were then analysed on a 2% (w/v) agarose gel.

**WGA staining**

**Figure Legends**

**Fig. 1. (A) Subsection of a 245 orthologous multi-gene concatenated phylogeny (85 taxa and 71,862 amino acid sites) demonstrating the branching position of *Hyphochytrium* within the stramenopiles calculated using Maximum Likelihood methods and summary of genome stats*.*** See Supp. X for full tree file. Dots on the phylogeny indicate 100% bootstrap support, values below 50% have been removed. Cartoons of cells indicate change in flagellum anatomy over the tree. Key indicates generalized lifestyle of the taxa sampled. **(B)** Table of genome statistics for a range of different stramenopiles.

**Fig. 2 Comparison of secreted proteome and putative carbohydrate active proteins across the pseudofungi.** A comparative distribution of the proportion of Oomycete proteomes and predicted secretomes which have carbohydrate digesting activity. The schematic phylogeny at the top indicates the relationship between different oomycete species with the dominant lifestyle of each species indicated by text colour. Specifically, green text (Phytophthora species) indicates hemibiotrophy, blue (Hyaloperonospora and Albugo) indicates obligate biotrophy, teal (Pythium) indicates nectrophy, orange (Saprolegnia) indicates both saprotrophy and necrotrophy, finally, black indicates unknown (Hyphochytrium) or non-oomycete (Ectocarpus and Thalassiosira). The first heat map in white/purple indicates the proportion of proteome of each organism (organised columnwise) which was identified as belonging to a particular CAZY ([www.cazy.org](http://www.cazy.org/)) database category using BLASTP with an expectation of 1e-5. The number listed is the proportion and the colouring relates to magnitude of the listed number (as shown by scale bar). The second heatmap in blue/yellow is largely the same but instead indicates the proportion of the secretome (predicted via a custom pipeline <https://github.com/fmaguire/predict_secretome/tree/refactor>) that is identified as belonging to each of these CAZY categories. Finally, the bottom barplot shows the proportion of the secretome for each organism which is predicted to be secreted by the above mentioned pipeline.

**Fig. 3.**

A. Homologous sequences were identified using 3 psi-blast iterations with H. catenoides putative MCP as query; to remove sequence redundancies, retrieved sequences were clustered at 90% amino-acid identity with cd-hit v4.6. Sequences were then aligned using MAFFT v7 iterative, global homology mode (G-INS-i); alignement sites retained for subsequent phylogenetic analysis were selected using trimal gap distribution mode. Final MCP multiple sequence alignment was composed of 386 sites. ML tree was inferred using iq-tree v1.3 and LG+I+G4+F model (determined as the best-fitting model by Bayesian information criterion). Node supports were evaluated with 100 non-parametric bootstrap replicates. The Mimiviridae clade was used to root the ML tree (unrooted version displayed on the lower left part). B. Reverse transcriptase (RT) PCR showing expression of polB and mg96 viral genes alongside an rps3 positive control. No expression of the mcp gene was detected. RT-PCR was performed on H. catenoides RNA alongside genomic DNA (+) and no template (-) controls, with PCR products run on an agarose gel alongside a 1kb ladder (Promega; 250 bp shown).

**Fig. 4. Comparative genomic analysis of flagellum proteome evolution. (A)** Heat map showing sequence identity profiles for flagella proteins with putative homologues present across a wide diversity of eukaryotes. The wider analysis uses a seed dataset of 592 proteins (see Table S6 for full dataset). The proteins show proteome loss or conservation patterns consistent with retention or loss of proteins that function in the anterior-tinselated flagellum. The heat map identifies 29 proteins present in the oomycetes but absent in *H. catenoides,* suggesting that this gene had been lost at the same proximate point to that where the posterior flagellum was also lost. The analysis also shows 12 proteins (marked\*) identified as posterior flagellum specific in *C. bullosa* that are retained in *H. catenoides***.** Three *C. bullosa* anterior flagellum specific proteins are also retained in *H. catenoides*. The putative radial spoke proteome also shows numerous losses similar to *Homo sapiens* (\*\*), this includes the loss of RSP7 (\*\*\*). **(B)** Shows a cartoon of the radial spoke protein complex identified in *Chlamydomonas* (88) with each shape number referring to the RPS number*.* Black shapes illustrate conserved proteins of the spoke complex, grey are non-conserved proteins, while the white complex refers to RPS7 which although absent in *Homo sapiens* has been lost separately when the posterior flagellum was also lost. Collectively, this suggests the *Hyphochytrium* anterior flagellum has a highly reduced flagellum spoke complex.

**Fig. 5 Photoreceptors putatively encoded on the *Hyphochytrium* genome. (A)** Distribution of known photoreceptor families across a subset of stramenopile genomes. **(B)** Subsection of the Rhodopsin alignment showing the conserved Schiff base. **(C)** Phylogeny of the rhodopsin gene family showing phylogenetic position of the hyphochytrium putative rhodopsins.

**Fig. 6. Comparative genomic analysis of gene families that function in polarised filamentous growth in the fungi. (A)** Cartoon outlining how proteins and complexes involved in polarised growth in *Saccharomyces cerevisiae*:vesicles are delivered from the Golgi (Fig. 5A-i) along cytoskeleton tracks to predetermined sites on the plasma membrane. Cdc42p is activated by Cdc24p (Fig. 5A-ii) promoting (97) assembly of the polarisome complex (Fig. 6A-iii) resulting in the formin Bni1p radiating actin cables (138, 139). Msb3p and Msb4p interact with Spa2 in the polarisome (Fig. 6A-iv) which is thought to recruit Cdc42 from the cytosol at the site of tip growth (140). Post-Golgi secretory vesicles are transported along the actin cables using a type V myosin (myo2) motor protein (141, 142) (Fig. 6A-v), to dock with the exocyst complex in a process dependent on Sec4 and its GEF Sec2 (143, 144)(Fig. 6A-vi) and so the vesicle is guided to its target site on the plasma membrane (145). Cdc42p along with Rho1 is required for localisation of Sec3p, which acts as a spatial marker for the formation of the exocyst (Fig. 5A-vii) while Rho3p and Cdc42p mediate docking of the vesicle (Fig. 5A-viii). Cdc42p plays a key role in regulating these processes in yeast but in Pezizomycotina and Basidiomycete fungi equivalent functions are also performed by Rac1p (146, 147). **(B)** The domain architecture of the 17 proteins associated with polarised growth in fungi. **(C)** The taxon distribution of putative homologues of polarised growth proteins across a representative set of taxa including the Pseudofungi. ‘P’ indicates a putative paralogue relationship as identified using phylogenetic analysis.

**Fig. 7. Comparative genomic analysis of gene families that function in cell wall synthesis. (A)** Micrographs showing the Wheat Germ Agglutinin fluorescentstaining of a chitin cell wall on *Hyphochytrium* developmental structures. **(B)** The domain architecture of eight proteins that function in cell wall synthesis. **(C)** The taxon distribution of putative gene families associated with cell wall synthesis across a representative set of taxa including the Pseudofungi.

**Fig. S1. Comparison of transcriptome to assembly**

**Scaffolds were sorted in to four groups; all, >1Kbp, >5Kbp and >10Kbp in length. For each dataset the N50 statistic was computed, along with gene predictions, CEGMA results and ‘blat’ hits against the transcriptome. The data suggests that removal of scaffolds <1Kbp does not have an appreciable effect on the “completeness” of the Hyphochytrium genome.**

**Fig. S2. Blobology analysis**

**‘Blobplots’ visualize genomes by displaying taxon-annotated GC-coverage plots. A/B: These plots show all scaffolds from the draft assembly with blast annotation from the NCBI ‘nt’ database, in B the transcriptome of Hyphochytrium and its mitochondria have been included. Notice the small circle in the upper-left of the image, these are the mitochondria. C/D: Shows the same images, however, only scaffolds greater than 1Kbp have been included.**

**Fig. S3. Emergent properties graphs**

**Two emergent self-organizing maps show the scaffold data plotted as tetra-mer frequencies represented by a neural network mapping approach. A: To determine contamination, the map was run on the Hyphochytrium scaffolds and does not show any areas suggestive of division within the data. B, To further the contamination check, other eukaryotic genomes and prokaryotes were spiked in to the process, here we see nice separation of all the genomes. Hyphochytrium scaffolds (orange), (Bacteria (blue): E. coli, Mycobacterium tuberculosis; Archaea (grey): Methanococcus vanniellii, Sulfolobus solfataricus; Fungi (purple): Encephalitozoon intestinalis, S. cerevisiae; Archaeplastida (green): Ostreococcus tauri; Protist (red): Cryptosporidium hominis)**

**Fig. S4. SNP frequency**

**A plot of SNP frequency for each scaffold of the draft genome arranged by length. SNPs are more common in shorter scaffolds. Blue line indicates trend-line with standard deviation.**

**Fig. S5. kmer coverage showing two peaks indicative of a diploid genome**

**A GenomeScope graph of the k-mer profile generated by using Jellyfish displaying k-mer coverage vs frequency. The two k-mer peaks are indicative of a diploid genome.**

**Fig. S6. CYP51 sterol-demethylase phylogeny and drug sensitivity assay**

**A RAxML 1000BS phylogeny showing the CYP51 sterol-demethylase gene**

**Fig. S7. Hyphochytrium TEM**

**This figure shows six imaged cells (A, C, G, J, K & L, with micrographs: B, D, E, F, H, I) after twenty cells on one slide were subject to cursory inspection. The cultures were grown in PYG medium (virus infection - that’s a question for David Milner, I was woeful at that). The samples then underwent high pressure freezing using a Leica EM RTS and were left to undergo free substitution using an infiltration mixture of 0.1% uranyl acetate and 1% osmium tetroxide as the fixative overnight. The fixative was removed and the concentration of EPON resin was increased in over 6 increments until embedding occurred at 100% EPON. The resultant palate was cut using a diamond knife and exposed to lead citrate before being mounted on a standard copper grid. All images were obtained using a Jeol JEM 1400 transmission electron microscope.S8. Mitochondrial genome assembly diagram**

**A genomic map of the mitochondrial genome of Hyphochytrium.**

The outer tracks show the position of genes (pink) and tRNAs on both strands, followed by start and stop codons of open reading frames across the assembly. The inner blue/green tracks show blast hits against the genomes of Phytophthora, Ectocarpus & Saprolegnia respectively. The black tracks shows XXX and the blue/green inner circle shows G/C content. Note the empty space at the top of the map showing the region we were unable to amplify/close during assembly.

**Fig. S9. ATP1 gene phylogeny**

A divergent nuclear-encoded Atp1-like protein present in a subset of Stramenopiles.

The nuclear encoded H. catenoides Atp1-like protein sequence was used as a query in BLAST and pHMMer searches. Sequences retrieved were aligned with sequences from Ševčíková et al. 2016 (Genome Biol Evol. 2016 Feb 12;8(3):705-22. doi: 10.1093/gbe/evw027.) using MUSCLE. Sequences were subjected to phylogenetic analysis using RaxML with 500 bootstraps using the LG substitution matrix. Atp1-like sequences predicted to be targeted to mitochondria by TargetPv1.1 are indicated by an encircled "M". Support represented as inset. Bar represents substitutions per site. H. catenoides sequences are in red.

**Fig. S10. Four phylogenies showing pseudofungal protein sequences branching with plastid carrying eukaryotes nested within a cluster of bacterial sequences potentially representing a gene derived from the plastid endosymbiosis.** All phylogenies were calculated using Maximum Likelihood methods with 100 bootstrap replicates as described in the methods.Sequences from all taxa possessing plastid of primary endosymbiotic ancestry (i.e. Archaeplastida) were searched using TargetP for evidence of an N-terminal organelle targeting peptide. All alignment and data matrices are provided at XXXX. **(A)** Pas -Histidine Kinase domain containing protein phylogeny calculated from a data matrix of 46 sequences and 281 alignment positions. **(B)** Cytidylytransferase family domain containing protein phylogeny calculated from a matrix of 38 sequences and 178 alignment positions. **(C)**  Phylogeny of a conserved hypothetical protein calculated from a data matrix of 41 sequences and 143 alignment positions. **(D)** Phylogeny of a putative Heme oxygenase domain calculated from a data matrix of 49 sequences and 186 alignment positions.

**Table 1.** Comparison of pseudofungal genes with function related to lifestyle

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Gene families** | *Hyphochytrium catenoides* | *Albugo laibachii* | *Hyaloperonospora arabidopsidis* | *Phytophthora infestans* | *Phytophthora ramorum* | *Phytophthora sojae* | *Pythium ultimum* | *Saprolegnia parasitica* | *Ectocarpus siliculosus* | *Thalassiosira pseudonana* |
| RXLR | 1 | 0 | 23 | 317 | 102 | 106 | 0 | 0 | 0 | 0 |
| NPP1-like proteins | 0 | 0 | 21 | 27 | 62 | 74 | 7 | 0 | 0 | 0 |
| Elicitin | 0 | 9 | 14 | 43 | 47 | 53 | 44 | 25 | 0 | 0 |
| **Plant Cell wall**  **degrading:** |  |  |  |  |  |  |  |  |  |  |
| Cutinase | 0 | 3 | 2 | 4 | 4 | 16 | 0 | 0 | 0 | 0 |
| Glycosyl hydrolases | 357 | 384 | 242 | 533 | 838 | 1208 | 436 | 415 | 282 | 264 |
| Pectin methyl esterases | 0 | 0 | 4 | 11 | 13 | 19 | 0 | 0 | 0 | 0 |
| Pectate lyase | 0 | 0 | 8 | 36 | 25 | 24 | 16 | 0 | 0 | 0 |
| Polygalacturonase | 0 | 3 | 3 | 24 | 17 | 25 | 6 | 3 | 0 | 0 |
| **Lectins** |  |  |  |  |  |  |  |  |  |  |
| PAN lectin | 4 | 3 | 2 | 5 | 8 | 5 | 11 | 6 | 1 | 0 |
| Ricin lectin | 1 | 1 | 3 | 5 | 9 | 10 | 5 | 57 | 0 | 1 |
| Jacalin lectin | 2 | 0 | 8 | 15 | 23 | 15 | 3 | 4 | 1 | 0 |
| Galactose lectin | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Leguminous lectin | 2 | 1 | 0 | 2 | 1 | 1 | 2 | 0 | 2 | 1 |
| Legume-like lectin | 2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 0 |
| **Protease functions** |  |  |  |  |  |  |  |  |  |  |
| Protease inhibitors | 13 | 11 | 14 | 51 | 35 | 46 | 30 | 28 | 15 | 15 |
| Proteases, all | 428 | 379 | 324 | 450 | 541 | 602 | 482 | 630 | 361 | 367 |
| Serine proteases | 166 | 84 | 106 | 170 | 182 | 189 | 200 | 248 | 112 | 140 |
| Metalloproteases | 92 | 91 | 80 | 98 | 100 | 91 | 107 | 129 | 88 | 101 |
| Cysteine proteases | 115 | 124 | 92 | 140 | 116 | 113 | 121 | 208 | 117 | 85 |
| **Others** |  |  |  |  |  |  |  |  |  |  |
| ABC transporters | 81 | 36 | 49 | 148 | 171 | 175 | 158 | 138 | 70 | 58 |
| Protein kinases | 243 | 305 | 217 | 423 | 398 | 430 | 232 | 690 | 330 | 160 |
| Notch protein | 3 | 0 | 1 | 1 | 1 | 1 | 1 | 18 | 11 | 2 |
| Haemolysin E | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 |

**Table SX. Identification of putative homologues of known flagellum associated proteins.** This analyses includes comparison with the *Colpomenia bullosa* flagella proteomes (85)

1. Patterson DJ. Stamenopiles: chromophytes from a protistan perspective. In: Green JC, Leadbeater BSC, Diver WL, editors. The chromophyte algae, problems and perspectives. Oxford: Claredon; 1989. p. 357-79.

2. Cavalier-Smith T. The kingdom Chromista: origin and systematics. In: Round FE, Chapman DJ, editors. Progress in Phycological Research. 4. Bristol: Biopress Ltd; 1986. p. 309-47.

3. Cavalier-Smith T, Chao EE. Phylogeny and megasystematics of phagotrophic heterokonts (kingdom Chromista). J Mol Evol. 2006;62(4):388-420.

4. Cavalier-Smith T. The origin of fungi and pseudofungi. In: Rayer ADM, editor. Evolutionary Biology of the Fungi (British Mycological Society Symposia). Cambridge: Cambridge University Press; 1987. p. 339-53.

5. Burki F, Kaplan M, Tikhonenkov DV, Zlatogursky V, Minh BQ, Radaykina LV, et al. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. Proceedings of the Royal Society of London B: Biological Sciences. 2016;283(1823).

6. Cooney EW, Barr DJS, Barstow WE. The ultrastructure of the zoospore of Hyphochytrium catenoides. Canadian Journal of Botany. 1985;63(3):497-505.

7. Archibald JM. The puzzle of plastid evolution. Curr Biol. 2009;19(2):R81-8.

8. Derelle R, López-García P, Timpano H, Moreira D. A phylogenomic framework to study the diversity and evolution of stramenopiles (=heterokonts). Molecular Biology and Evolution. 2016.

9. Cavalier-Smith T. Membrane heredity and early chloroplast evolution. Trends Plant Sci. 2000;5(4):174-82.

10. Baurain D, Brinkmann H, Petersen J, Rodríguez-Ezpeleta N, Stechmann A, Demoulin V, et al. Phylogenomic Evidence for Separate Acquisition of Plastids in Cryptophytes, Haptophytes, and Stramenopiles. Molecular Biology and Evolution. 2010;27(7):1698-709.

11. Stiller JW, Schreiber J, Yue J, Guo H, Ding Q, Huang J. The evolution of photosynthesis in chromist algae through serial endosymbioses. Nat Commun. 2014;5.

12. Petersen J, Ludewig A-K, Michael V, Bunk B, Jarek M, Baurain D, et al. Chromera velia, Endosymbioses and the Rhodoplex Hypothesis—Plastid Evolution in Cryptophytes, Alveolates, Stramenopiles, and Haptophytes (CASH Lineages). Genome Biology and Evolution. 2014;6(3):666-84.

13. Stiller JW, Huang J, Ding Q, Tian J, Goodwillie C. Are algal genes in nonphotosynthetic protists evidence of historical plastid endosymbioses? BMC Genomics. 2009;10:484-.

14. Massana R, Castresana J, Balagué V, Guillou L, Romari K, Groisillier A, et al. Phylogenetic and ecological analysis of novel marine stramenopiles. Appl Environ Microbiol. 2004;70(6):3528-34.

15. Richards TA, Bass D. Molecular screening of free-living microbial eukaryotes: diversity and distribution using a meta-analysis. Curr Opin Microbiol. 2005;8(3):240-52.

16. Armbrust E, Berges J, Bowler C, Green B, Martinez D, Putnam N, et al. The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. Science. 2004;306:79 - 86.

17. Cock J, Sterck L, Rouze P, Scornet D, Allen A, Amoutzias G, et al. The Ectocarpus genome and the independent evolution of multicellularity in brown algae. Nature. 2010;465:617 - 21.

18. Tyler BM, Tripathy S, Zhang X, Dehal P, Jiang RH, Aerts A, et al. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. Science. 2006;313(5791):1261-6.

19. Jiang RHY, de Bruijn I, Haas BJ, Belmonte R, Löbach L, Christie J, et al. Distinctive expansion of potential virulence genes in the genome of the oomycete fish pathogen *Saprolegnia parasitica*. PLoS Genet. 2013;9.

20. Karling JS. A New Fungus with Anteriorly Uniciliate Zoospores: Hyphochytrium catenoides. American Journal of Botany. 1939;26(7):512-9.

21. Van der Auwera G, De Baere R, Van de Peer Y, De Rijk P, Van den Broeck I, De Wachter R. The phylogeny of the Hyphochytriomycota as deduced from ribosomal RNA sequences of *Hyphochytrium catenoides*. Mol Biol Evol. 1995;12(4):671-8.

22. Schmieder R, Lim YW, Rohwer F, Edwards R. TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. BMC Bioinformatics. 2010;11:341-.

23. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27(6):863-4.

24. khmer-protocols documentation [Internet]. 2013.

25. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Research. 2014;24(8):1384-95.

26. Richards TA, Soanes DM, Jones MD, Vasieva O, Leonard G, Paszkiewicz K, et al. Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. Proc Natl Acad Sci USA. 2011;108(37):15258-63.

27. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23(9):1061-7.

28. Wilhelm J, Pingoud A, Hahn M. Real-time PCR-based method for the estimation of genome sizes. Nucleic Acids Research. 2003;31(10):e56-e.

29. Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. Frontiers in Genetics. 2013;4:237.

30. Koutsovoulos G, Kumar S, Laetsch DR, Stevens L, Daub J, Conlon C, et al. No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. Proc Natl Acad Sci USA. 2016;X(X):600338113v1-201600338.

31. Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, et al. Community-wide analysis of microbial genome sequence signatures. Genome Biology. 2009;10(8):R85.

32. Jurka J, Kapitonov V, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110:462 - 7.

33. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27(6):764-70.

34. Ramesh MA, Malik SB, Logsdon JM, Jr. A phylogenomic inventory of meiotic genes; evidence for sex in *Giardia* and an early eukaryotic origin of meiosis. Curr Biol. 2005;15:185-91.

35. Johnson TW, Jr. Resting Spore Development in the Marine Phycomycete Anisolpidium ectocarpii. American Journal of Botany. 1957;44(10):875-8.

36. Beakes G, Glockling S, Sekimoto S. The evolutionary phylogeny of the oomycete “fungi”. Protoplasma. 2012;249(1):3-19.

37. Brown MW, Sharpe SC, Silberman JD, Heiss AA, Lang BF, Simpson AGB, et al. Phylogenomics demonstrates that breviate flagellates are related to opisthokonts and apusomonads. Proceedings of the Royal Society B: Biological Sciences. 2013;280(1769):20131755.

38. Noguchi F, Tanifuji G, Brown MW, Fujikura K, Takishita K. Complex evolution of two types of cardiolipin synthase in the eukaryotic lineage stramenopiles. Molecular Phylogenetics and Evolution. 2016;101:133-41.

39. Tsui C, Marshall W, Yokoyama R, Honda D, Lippmeier J, Craven K, et al. Labyrinthulomycetes phylogeny and its implications for the evolutionary loss of chloroplasts and gain of ectoplasmic gliding. Mol Phylogenet Evol. 2009;50:129 - 40.

40. Secq M-PO-L, Goër SL-d, Stam WT, Olsen JL. Complete mitochondrial genomes of the three brown algae (Heterokonta: Phaeophyceae) Dictyota dichotoma, Fucus vesiculosus and Desmarestia viridis. Current Genetics. 2005;49(1):47-58.

41. Philippe H. Opinion: long branch attraction and protist phylogeny. Protist. 2000;151(4):307-16.

42. Judelson HS. Dynamics and Innovations within Oomycete Genomes: Insights into Biology, Pathology, and Evolution. Eukaryotic Cell. 2012;11(11):1304-12.

43. Gaulin E, Bottin A, Dumas B. Sterol biosynthesis in oomycete pathogens. Plant Signaling & Behavior. 2010;5(3):258-60.

44. Gaulin E, Madoui M, Bottin A, Jacquet C, Mathe C, Couloux A, et al. Transcriptome of Aphanomyces euteiches: New oomycete putative pathogenicity factors and metabolic pathways. PLoS ONE. 2008;3:e1723.

45. Warrilow AGS, Hull CM, Rolley NJ, Parker JE, Nes WD, Smith SN, et al. Clotrimazole as a Potent Agent for Treating the Oomycete Fish Pathogen Saprolegnia parasitica through Inhibition of Sterol 14α-Demethylase (CYP51). Applied and Environmental Microbiology. 2014;80(19):6154-66.

46. Baxter L, Tripathy S, Ishaque N, Boot N, Cabral A, Kemen E, et al. Signatures of adaptation to obligate biotrophy in the Hyaloperonospora arabidopsidis genome. Science. 2010;330(6010):1549-51.

47. Raffaele S, Win J, Cano L, Kamoun S. Analyses of genome architecture and gene expression reveal novel candidate virulence factors in the secretome of Phytophthora infestans. BMC Genomics. 2010;11(1):637.

48. Levesque CA, Brouwer H, Cano L, Hamilton J, Holt C, Huitema E, et al. Genome sequence of the necrotrophic plant pathogen Pythium ultimum reveals original pathogenicity mechanisms and effector repertoire. Genome Biology. 2010;11(7):R73.

49. Haas B, Kamoun S, Zody M, Jiang R, Handsaker R, Cano L, et al. Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans. Nature. 2009;461:393 - 8.

50. Birch P, Rehmany A, Pritchard L, Kamoun S, Beynon J. Trafficking arms: oomycete effectors enter host plant cells. Trends Microbiol. 2006;14:8 - 11.

51. Birch P, Armstrong M, Bos J, Boevink P, Gilroy E, Taylor R, et al. Towards understanding the virulence functions of RXLR effectors of the oomycete plant pathogen Phytophthora infestans. J Exp Bot. 2009;60:1133 - 40.

52. Dou D, Kale S, Wang X, Jiang R, Bruce N, Arredondo F, et al. RXLR-mediated entry of Phytophthora sojae effector Avr1b into soybean cells does not require pathogen-encoded machinery. Plant Cell. 2008;20:1930 - 47.

53. Dou D, Kale S, Wang X, Chen Y, Wang Q, Wang X, et al. Carboxy-terminal motifs common to many oomycete RXLR effectors are required for avirulence and suppression of BAX-mediated programmed cell death by Phytophthora sojae effector Avr1b. Plant Cell. 2008;20:1118 - 33.

54. Qutob D, Kemmerling B, Brunner F, Kufner I, Engelhardt S, Gust AA, et al. Phytotoxicity and innate immune responses induced by Nep1-like proteins. Plant Cell. 2006;18(12):3721-44.

55. Fellbrich G, Romanski A, Varet A, Blume B, Brunner F, Engelhardt S, et al. NPP1, a Phytophthora-associated trigger of plant defense in parsley and Arabidopsis. The Plant Journal. 2002;32(3):375-90.

56. Yu LM. Elicitins from Phytophthora and basic resistance in tobacco. Proceedings of the National Academy of Sciences of the United States of America. 1995;92(10):4088-94.

57. Belbahri L, Calmin G, Mauch F, Andersson J. Evolution of the cutinase gene family: evidence for lateral gene transfer of a candidate Phytophthora virulence factor. Gene. 2008;408:1 - 8.

58. Kemen E, Gardiner A, Schultz-Larsen T, Kemen AC, Balmuth AL, Robert-Seilaniantz A, et al. Gene gain and loss during evolution of obligate parasitism in the white rust pathogen of Arabidopsis thaliana. PLoS Biol. 2011;9.

59. Blackman LM, Cullerne DP, Hardham AR. Bioinformatic characterisation of genes encoding cell wall degrading enzymes in the Phytophthora parasitica genome. BMC Genomics. 2014;15(1):1-24.

60. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Research. 2014;42(D1):D490-D5.

61. Bochner BR, Gadzinski P, Panomitros E. Phenotype MicroArrays for High-Throughput Phenotypic Testing and Assay of Gene Function. Genome Research. 2001;11(7):1246-55.

62. Taoka Y, Nagano N, Okita Y, Izumida H, Sugimoto S, Hayashi M. Effect of Tween 80 on the growth, lipid accumulation and fatty acid composition of Thraustochytrium aureum ATCC 34304. Journal of Bioscience and Bioengineering. 2011;111(4):420-4.

63. Adhikari BN, Hamilton JP, Zerillo MM, Tisserat N, Lévesque CA, Buell CR. Comparative Genomics Reveals Insight into Virulence Strategies of Plant Pathogenic Oomycetes. PLoS ONE. 2013;8(10):e75072.

64. Zerillo MM, Adhikari BN, Hamilton JP, Buell CR, Lévesque CA, Tisserat N. Carbohydrate-Active Enzymes in <italic>Pythium</italic> and Their Role in Plant Cell Wall and Storage Polysaccharide Degradation. PLoS ONE. 2013;8(9):e72572.

65. Seidl MF, Van den Ackerveken G, Govers F, Snel B. A Domain-Centric Analysis of Oomycete Plant Pathogen Genomes Reveals Unique Protein Organization. Plant Physiology. 2011;155(2):628-44.

66. Doolittle RF. The multiplicity of domains in proteins. Annu Rev Biochem. 1995;64:287-314.

67. Leonard G, Richards TA. Genome-scale comparative analysis of gene fusions, gene fissions, and the fungal tree of life. Proc Natl Acad Sci USA. 2012;109(52):21402-7.

68. Li L, Stoeckert CJ, Jr., Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13(9):2178-89.

69. Sharma V, Colson P, Giorgi R, Pontarotti P, Raoult D. DNA-Dependent RNA Polymerase Detects Hidden Giant Viruses in Published Databanks. Genome Biology and Evolution. 2014;6(7):1603-10.

70. Savory F, Leonard G, Richards TA. The Role of Horizontal Gene Transfer in the Evolution of the Oomycetes. PLoS Pathog. 2015;11(5):e1004805.

71. Boisvert S, Raymond F, Godzaridis E, Laviolette F, Corbeil J. Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol. 2012;13(12):R122.

72. GCC. Sequencher. 4.2 ed. Ann Arbor: Gene Codes Corporation; 2003.

73. McNabb SA, Eros RW, Klassen GR. Presence and absence of large inverted repeats in the mitochondrial DNA of Hyphochytriomycetes. Canadian Journal of Botany. 1988;66(12):2377-9.

74. O'Brien MA, Misner I, Lane CE. Mitochondrial Genome Sequences and Comparative Genomics of Achlya hypogyna and Thraustotheca clavata. Journal of Eukaryotic Microbiology. 2014;61(2):146-54.

75. Ševčíková T, Klimeš V, Zbránková V, Strnad H, Hroudová M, Vlček Č, et al. A Comparative Analysis of Mitochondrial Genomes in Eustigmatophyte Algae. Genome Biology and Evolution. 2016;8(3):705-22.

76. Reyes-Prieto A, Moustafa A, Bhattacharya D. Multiple Genes of Apparent Algal Origin Suggest Ciliates May Once Have Been Photosynthetic. Current Biology. 2008;18(13):956-62.

77. Huang J, Mullapudi N, Lancto CA, Scott M, Abrahamsen MS, Kissinger JC. Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in *Cryptosporidium parvum*. Genome Biol. 2004;5(11):R88.

78. Riordan CE, Ault JG, Langreth SG, Keithly JS. *Cryptosporidium parvum Cpn60* targets a relict organelle. Curr Genet. 2003;44(3):138-47.

79. Roger AJ, Svard SG, Tovar J, Clark CG, Smith MW, Gillin FD, et al. A mitochondrial-like chaperonin 60 gene in *Giardia lamblia*: evidence that diplomonads once harbored an endosymbiont related to the progenitor of mitochondria. Proc Natl Acad Sci USA. 1998;95(1):229-34.

80. Hirt RP, Healy B, Vossbrinck CR, Canning EU, Embley TM. A mitochondrial HSP70 orthologue in *Vairimorpha necatrix*: molecular evidence that microsporidia once contained mitochondria. Curr Biol. 1997;7(12):995-8.

81. Martin W, Herrmann RG. Gene transfer from organelles to the nucleus: how much, what happens, and Why? Plant Physiol. 1998;118(1):9-17.

82. Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, et al. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci USA. 2002;99(19):12246-51.

83. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol. 2000;300(4):1005-16.

84. Ajjawi I, Tsegaye Y, Shintani D. Determination of the genetic, molecular, and biochemical basis of the Arabidopsis thaliana thiamin auxotroph th1. Archives of Biochemistry and Biophysics. 2007;459(1):107-14.

85. Fu G, Nagasato C, Oka S, Cock JM, Motomura T. Proteomics Analysis of Heterogeneous Flagella in Brown Algae (Stramenopiles). Protist. 2014;165(5):662-75.

86. Derelle R, Torruella G, Klimeš V, Brinkmann H, Kim E, Vlček Č, et al. Bacterial proteins pinpoint a single eukaryotic root. Proceedings of the National Academy of Sciences. 2015;112(7):E693-E9.

87. Wirschell M, Olbrich H, Werner C, Tritschler D, Bower R, Sale W, et al. The nexin-dynein regulatory complex subunit DRC 1 is essential for motile cilia function in algae and humans. Nature genetics. 2013;45(3):10.1038/ng.2533.

88. Yang P, Diener DR, Yang C, Kohno T, Pazour GJ, Dienes JM, et al. Radial spoke proteins of Chlamydomonas flagella. Journal of Cell Science. 2006;119(6):1165-74.

89. Yang C, Compton MM, Yang P. Dimeric Novel HSP40 Is Incorporated into the Radial Spoke Complex during the Assembly Process in Flagella. Molecular Biology of the Cell. 2005;16(2):637-48.

90. Jékely G. Evolution of phototaxis. Philos Trans R Soc Lond B Biol Sci. 2009;364(1531):2795-808.

91. Kawai H. Green flagellar autofluorescence in brown algal swarmers and their phototactic responses. Bot Mag Tokyo. 1992;105(1):171-84.

92. Kawai H, Nakamura S, Mimuro M, Furuya M, Watanabe M. Microspectrofluorometry of the autofluorescent flagellum in phototactic brown algal zoids. Protoplasma. 1996;191(3-4):172-7.

93. Avelar Gabriela M, Schumacher Robert I, Zaini Paulo A, Leonard G, Richards Thomas A, Gomes Suely L. A rhodopsin-guanylyl cyclase gene fusion functions in visual perception in a fungus. Current Biology. 2014;24(11):1234-40.

94. Sudbery PE. Regulation of polarised growth in fungi. Fungal Biol Rev. 2008;22(2):44-55.

95. Bartnicki-Garcia S. The cell wall: a crucial structure in fungal evolution. In: Rayner ADM, Brasier CM, Moore D, editors. Evolutionary Biology of the Fungi. Cambridge: Cambridge University Press; 1987. p. 389-403.

96. Richards TA, Talbot NJ. Horizontal gene transfer in osmotrophs: playing with public goods. Nat Rev Microbiol. 2013;11(10):720-7.

97. Park H-O, Bi E. Central roles of small GTPases in the development of cell polarity in yeast and beyond. Microbiol Mol Biol Rev. 2007;71(1):48-96.

98. Bartnicki-Garcia S. Cell wall chemistry, morphogenesis, and taxonomy of fungi. Annu Rev Microbiol. 1968;22:87-108.

99. Blanc G, Duncan G, Agarkova I, Borodovsky M, Gurnon J, Kuo A, et al. The *Chlorella variabilis* NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. Plant Cell. 2010;22(9):2943-55.

100. Fuller MS, Barshad I. Chitin and cellulose in the cell walls of *Rhizidiomyces* sp. Am J Bot. 1960;47(2):105-9.

101. Kneipp LF, Andrade AF, de Souza W, Angluster J, Alviano CS, Travassos LR. *Trichomonas vaginalis* and *Tritrichomonas foetus*: expression of chitin at the cell surface. Exp Parasitol. 1998;89(2):195-204.

102. Lin CC, Aronson JM. Chitin and cellulose in the cell walls of the oomycete, *Apodachlya* sp. Arch Mikrobiol. 1970;72(2):111-4.

103. Arroyo-Begovich A, xe, rabez-Trejo A, Ru, xed, z-Herrera J, et al. Identification of the structural component in the cyst wall of *Entamoeba invadens*. J Parasitol. 1980;66(5):735-41.

104. Das S, Van Dellen K, Bulik D, Magnelli P, Cui J, Head J, et al. The cyst wall of *Entamoeba invadens* contains chitosan (deacetylated chitin). Mol Biochem Parasit. 2006;148(1):86-92.

105. Schwelm A, Fogelqvist J, Knaust A, Jülke S, Lilja T, Bonilla-Rosso G, et al. The *Plasmodiophora brassicae* genome reveals insights in its life cycle and ancestry of chitin synthases. Scientific Reports. 2015;5:11153.

106. Clay RP, Benhamou N, Fuller MS. Ultrastructural detection of polysaccharides in the cell walls of two members of the Hyphocytriales. Mycological Research. 1991;95(9):1057-64.

107. Mélida H, Sandoval-Sierra JV, Diéguez-Uribeondo J, Bulone V. Analyses of extracellular carbohydrates in oomycetes unveil the existence of three different cell wall types. Eukaryot Cell. 2013;12(2):194-203.

108. Durkin CA, Mock T, Armbrust EV. Chitin in diatoms and its association with the cell wall. Eukaryot Cell. 2009;8(7):1038-50.

109. Doolittle WF. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet. 1998;14(8):307-11.

110. Lefèvre E, Bardot C, Noël C, Carrias JF, Viscogliosi E, Amblard C, et al. Unveiling fungal zooflagellates as members of freshwater picoeukaryotes: evidence from a molecular diversity study in a deep meromictic lake. Environ Microbiol. 2007;9(1):61-71.

111. Edwards U, Rogall T, Blocker H, Emde M, Bottger EC. Isolation and direct complete nucleotide determination of entire genes. Characterisation of a gene encoding for 16S ribosomal RNA. . Nucleic Acids Res. 1989;17:7843-53.

112. Boisvert S, Raymond F, Godzaridis É, Laviolette F, Corbeil J. Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biology. 2012;13(12):R122.

113. Hall B, DeRego T, Geib S. GAG: the Genome Annotation Generator (Version 1.0) [Software]. <http://genomeannotation.github.io/GAG2014>.

114. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Research. 2008;18(12):1979-90.

115. Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics. 2011;27(6):757-63.

116. Tate R, Hall B, DeRego T, Geib S. Annie: the ANNotation Information Extractor (Version 1.0) <http://genomeannotation.github.io/annie2014>.

117. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068-9.

118. Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, et al. The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic acids research. 2015;43(W1):W580-4.

119. Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, et al. MITOS: Improved de novo metazoan mitochondrial genome annotation. Molecular Phylogenetics and Evolution. 2013;69(2):313-9.

120. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research. 1997;25(5):955-64.

121. Stothard P, Wishart DS. Circular genome visualization and exploration using CGView. Bioinformatics. 2005;21(4):537-9.

122. Park BH, Karpinets TV, Syed MH, Leuze MR, Uberbacher EC. CAZymes Analysis Toolkit (CAT): Web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. Glycobiology. 2010;20.

123. Whisson SC, Boevink PC, Moleleki L, Avrova AO, Morales JG, Gilroy EM, et al. A translocation signal for delivery of oomycete effector proteins into host plant cells. Nature. 2007;450(7166):115-8.

124. Sharma R, Xia X, Cano LM, Evangelisti E, Kemen E, Judelson H, et al. Genome analyses of the sunflower pathogen Plasmopara halstedii provide insights into effector evolution in downy mildews and Phytophthora. BMC Genomics. 2015;16(1):741.

125. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8(10):785-6.

126. Krogh A, Larrson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001;305(3):567-80.

127. Sonnhammer EL, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol. 1998;6:175-82.

128. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, et al. WoLF PSORT: protein localization predictor. Nucleic Acids Res. 2007;35(Web Server issue):W585-7.

129. Cantarel B, Coutinho P, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res. 2009;37:D233 - D8.

130. Edgar R. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792 - 7.

131. Galtier N, Gouy M, Gautier C. SEAVIEW and PHYLO\_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci. 1996;12(6):543-8.

132. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014 30(9):1312-3.

133. Richards TA, Soanes DM, Foster PG, Leonard G, Thornton CR, Talbot NJ. Phylogenomic analysis demonstrates a pattern of rare and ancient horizontal gene transfer between plants and fungi. Plant Cell. 2009;21(7):1897-911.

134. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389-402.

135. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 2005;33(2):511-8.

136. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972-3.

137. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS ONE. 2010;5(3):e9490.

138. Pruyne D, Evangelista M, Yang C, Bi E, Zigmond S, Bretscher A, et al. Role of formins in actin assembly: nucleation and barbed-end association. Science. 2002;297(5581):612-5.

139. Pruyne DW, Schott DH, Bretscher A. Tropomyosin-containing actin cables direct the Myo2p-dependent polarized delivery of secretory vesicles in budding yeast. J Cell Biol. 1998;143(7):1931-45.

140. Tcheperegine SE, Gao X-D, Bi E. Regulation of cell polarity by interactions of Msb3 and Msb4 with Cdc42 and polarisome components. Mol Cell Biol. 2005;25(19):8567-80.

141. Bretscher A. Polarized growth and organelle segregation in yeast: the tracks, motors, and receptors. J Cell Biol. 2003;160(6):811-6.

142. Pruyne D, Legesse-Miller A, Gao L, Dong Y, Bretscher A. Mechanisms of polarized growth and organelle segregation in yeast. Annu Rev Cell Dev Biol. 2004;20(1):559-91.

143. Walch-Solimena C, Collins RN, Novick PJ. Sec2p mediates nucleotide exchange on Sec4p and Is involved in polarized delivery of post-Golgi vesicles. J Cell Biol. 1997;137(7):1495-509.

144. Goud B, Salminen A, Walworth NC, Novick PJ. A GTP-binding protein required for secretion rapidly associates with secretory vesicles and the plasma membrane in yeast. Cell. 1988;53(5):753-68.

145. Walworth NC, Goud B, Kabcenell AK, Novick PJ. Mutational analysis of SEC4 suggests a cyclical mechanism for the regulation of vesicular traffic. The EMBO Journal. 1989;8(6):1685-93.

146. Virag A, Lee MP, Se H, Harris SD. Regulation of hyphal morphogenesis by cdc42 and rac1 homologues in *Aspergillus nidulans*. Mol Micro. 2007;66(6):1579-96.

147. Mahlert M, Leveleki L, Hlubek A, Sandrock B, Bölker M. Rac1 and Cdc42 regulate hyphal growth and cytokinesis in the dimorphic fungus *Ustilago maydis*. Mol Micro. 2006;59(2):567-78.