

Reviewed Preprint

Revised by authors after peer review.

[About eLife's process](#)

Reviewed preprint version 2

April 26, 2024 (this version)

Reviewed preprint version 1

December 19, 2023

Sent for peer review

September 21, 2023

Posted to preprint server

September 16, 2023

Signatures of transposon-mediated genome inflation, host specialization, and photoentrainment in *Entomophthora muscae* and allied entomophthoralean fungi

Jason E. Stajich , Brian Lovett, Emily Lee, Angie M. Macias, Ann E. Hajek, Benjamin L. de Bivort, Matt T. Kasson, Henrik H. De Fine Licht, Carolyn Elya 

Department of Microbiology and Plant Pathology, University of California-Riverside, Riverside, CA United States • Emerging Pests and Pathogens Research Unit, USDA-ARS, Ithaca, NY, United States • Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, United States • Division of Plant and Soil Sciences, West Virginia University, Morgantown, WV, United States • Department of Entomology, Cornell University, Ithaca, NY, United States • Section for Organismal Biology, Department of Plant and Environmental Sciences, University of Copenhagen, Copenhagen, Denmark • Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, United States

 https://en.wikipedia.org/wiki/Open_access

 Copyright information

Abstract

Despite over a century of observations, the obligate insect parasites within the order Entomophthorales remain poorly characterized at the genetic level. This is in part due to their large genome sizes and difficulty in obtaining sequenceable material. In this manuscript, we leveraged a recently-isolated, laboratory-tractable *Entomophthora muscae* isolate and improved long-read sequencing to obtain a largely-complete entomophthoralean genome. Our *E. muscae* assembly is 1.03 Gb, consists of 7,810 contigs and contains 81.3% complete fungal BUSCOs. Using a comparative approach with other available (transcriptomic and genomic) datasets from entomophthoralean fungi, we provide new insight into the biology of these understudied pathogens. We offer a head-to-head comparison of morphological and molecular data for species within the *E. muscae* species complex. Our findings suggest that substantial taxonomic revision is needed to define species within this group and we provide recommendations for differentiating strains and species in the context of the existing body of *E. muscae* scientific literature. We show that giant genomes are the norm within Entomophthoraceae owing to extensive, but not recent, Ty3 retrotransposon activity, despite the presence of machinery to defend against transposable elements (RNAi). In addition, we find that *E. muscae* and its closest allies are enriched for M16A peptidases and possess genes that are likely homologs to the blue-light sensor *white-collar 1*, a *Neurospora crassa* gene that has a well-established role in maintaining circadian rhythms. We find that *E. muscae* has an expanded group of acid-trehalases, consistent with trehalose being the primary sugar component of fly (and insect) hemolymph. We uncover evidence that *E. muscae* diverged from other entomophthoralean fungi by expansion of existing families, rather than loss of particular domains, and possesses a potentially unique suite of secreted catabolic enzymes, consistent with *E. muscae*'s species-specific, biotrophic lifestyle.

Altogether, we provide a genetic and molecular foundation that we hope will provide a platform for the continued study of the unique biology of entomophthoralean fungi.

eLife assessment

This **important** study reports on the genome evolution of a poorly studied fungal group. By combining long-read sequencing and different bioinformatic analyses, the authors show that the giant genome of *Entomophthora muscae* expanded due to extensive transposable element activity. The strength of evidence is **compelling** and the authors are to be commended for their multiple comparative analyses of gene content along with transparently written and visualized techniques, data curation, and methods. This paper will be of relevance to fungal biologists as well as to evolutionary biologists interested in the study of genome size dynamics.

Introduction

Fungal insect pathogens play vital roles within ecosystems and have significant agricultural and economic impacts for human beings (Lovett and St Leger, 2017 [↗](#)). While several ascomycete entomopathogens have been extensively studied (e.g., *Beauveria bassiana*, *Metarhizium anisopliae*, etc.), many others, especially those within the fungal phylum Zoopagomycota (formerly Zygomycota), have received comparatively little attention and are poorly understood. Zoopagomycota is the earliest diverging lineage of non-flagellated fungi (Spatafora et al., 2016 [↗](#)) and consists of saprotrophs and animal and fungal parasites (Fig. 1A [↗](#)). In particular, fungi within the order Entomophthorales (subphylum Entomophthoromycotina) are obligate and often highly-specialized insect parasites that drive epizootic events and have massive impacts on local insect populations. For example, a 1989 outbreak of *Entomophaga maimaiga* was observed to cause population collapses of the invasive spongy moth, *Lymantria dispar*, across the northeastern US (Hajek et al., 1990 [↗](#)). In 1983, an outbreak of *Entomophthora muscae* in the black dump fly (*Ophyra aenescens*) was found to affect nearly the entire population (Mullens et al., 1987 [↗](#)). Owing to their deadliness for specific insect hosts, many entomophthoralean fungi have attracted attention with the hope of developing more targeted pesticides (Pell et al., 2001 [↗](#)). Yet, little progress has been made, due in part to the enduring mysteries of their biology.

In addition to ecological importance and promise for agricultural applications, many entomophthoralean fungi change the behavior of their hosts so dramatically that infected individuals are often referred to as “zombies” (de Bekker et al., 2021 [↗](#)). For example, *Entomophthora muscae* drives infected fly hosts to climb a nearby surface (a phenomenon referred to as “summit disease”) (Evans, 1989 [↗](#); Lovett et al., 2020b [↗](#); Roy et al., 2006 [↗](#)), extend its proboscis and raise its wings immediately prior to death (Elya et al., 2018 [↗](#); Krasnoff et al., 1995 [↗](#)). Likewise, *Entomophaga grylli* and *Pandora formicae* lead grasshoppers and ants to die clinging to or biting the tops of plants, respectively (Boer, 2008 [↗](#); Marikovsky, 1962 [↗](#); Pickford and Riegert, 1964 [↗](#)). *Massospora* spp. employ more active modes of transmission by their living, manipulated cicada hosts (Lovett et al., 2020a [↗](#)), including hypersexual behaviors that increase the rate of contact transmission (Cooley et al., 2018 [↗](#)).

A major challenge in understanding the biology of entomophthoralean fungi is the difficulty of culturing these fungi in the laboratory. While some species can be grown *in vitro* (Freimoser et al., 2000 [↗](#); Grundschober et al., 1998 [↗](#); Hajek et al., 2012 [↗](#); Holdom, 1983 [↗](#); Hua and Feng, 2003 [↗](#)) or, less commonly, inside lab-reared insects (Elya et al., 2018 [↗](#); Mullens, 1986 [↗](#)), such methods are only available for a few species. Genomes for these organisms have also been difficult to

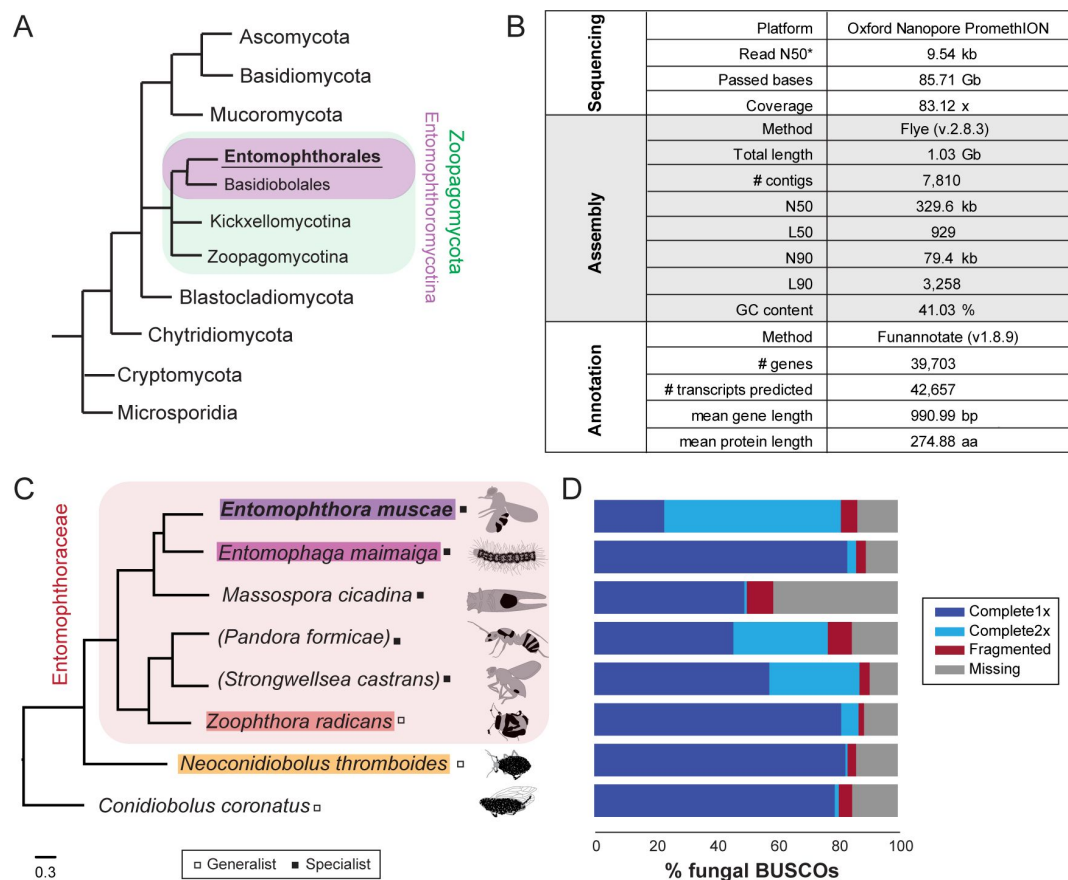


Figure 1.

A new *Entomophthora muscae* genome assembly and comparative entomophthoralean datasets.

A) Fungal cladogram, based on (Spatafora et al., 2016). All unshaded taxa are fungal phyla. Green-shaded branches are subphyla within phylum Zoopagomycota, purple-shaded branches are orders within subphylum Entomophthoromycotina. Branch lengths are not proportional to phylogenetic distance. B) Overview of sequencing, assembly and annotation statistics for new *E. muscae* genome assembly. Asterisk indicates that this value is for reads that passed the base-calling threshold only, not all bases sequenced. C) Cladogram presenting the evolutionary relationships of the Entomophthoraceae species considered in this study. Red-shaded branches are genera within the family Entomophthoraceae. Phylogenetic tree was constructed with FastTree using a concatenated set of conserved protein coding genes. Parentheses around *Pandora formicae* and *Strongwellsea castrans* indicate that transcriptomic datasets were used for these species; for all other species genomic datasets were used. Tree branch length is proportional to phylogenetic distance (substitution rate given in legend below). Species whose genomes comprised our core analysis set are colored. Classification of host specificity of each fungus (i.e., specialist or generalist) is denoted with a box to the right of the species name. Specialist species infect a narrow host range; generalists can infect a broad range of species. An example depiction of a host killed by each of these fungi is drawn to the right. Host tissues are gray, with fungal conidiophores depicted in black. Hosts (top to bottom): fruit fly (*Drosophila melanogaster* adult), spongy moth (*Lymantria dispar* larva), periodical cicada (*Magicicada septendecim* adult), ant (*Formica exsecta* adult), cabbage maggot fly (*Delia radicum* adult), Bagrada bug (*Bagrada hilaris* adult), aphid adult (Aphididae), and planthopper (*Delphacodes kuscheli* adult). D) BUSCO completeness estimates for the predicted proteome corresponding to species listed in C, using fungi_odb10 BUSCO set (Simão et al., 2015).

acquire owing to their extreme size, high proportion of repeat content, and the difficulty of obtaining sufficient high quality, high molecular weight DNA (Gryganskyi and Muszewska, 2014; Stajich JE. et al., 2022). Here, we describe a long-read based genome assembly of the fly pathogen *Entomophthora muscae* isolated from fruit flies (strain *E. muscae* ‘Berkeley’, (Elya et al., 2018)). We used this genome in a comparative analysis with other available entomophthoralean transcriptomic and genomic datasets that addressed four main biological questions: 1) Why are the genomes so large? 2) What are patterns in predicted functionality across Entomophthorales? 3) What elements are unique to *E. muscae*? 4) Are phylogenetic determinations made using morphological characteristics consistent with those based on molecular data? In assembling the *E. muscae* genome and using it to address these initial questions, we hope to provide biological insight into the Entomophthorales as well as provoke additional exploration into these understudied fungi. The answers we uncovered each reveal new avenues of research that will lead to a better understanding of the evolution of entomophthoralean insect pathogens, particularly the influence of vastly proliferated transposable elements, in support of future work and applied use of these insect pathogens.

Results

A long-read assembly of the E. muscae genome in the context of a comparative Entomophthorales dataset

Using a modified genomic DNA extraction protocol for filamentous fungi in the genus *Trichoderma* (Elya and Lee, 2022), we extracted high molecular weight DNA from a *in vitro*-grown *Entomophthora muscae* culture inoculated from a single sporulating *E. muscae*-killed fruit fly. Oxford PromethION sequencing of the resultant genomic library yielded 85.71 Gb of sequenced bases (N50 = 9.54 kb) that passed quality control thresholds (Fig. 1B). These reads were assembled with Flye (v2.8.3) and self-polished using Medaka (v1.2.6) to yield a 1.03 Gb genome consisting of 7,936 contigs (N50 = 301.1 kb). Additional scaffolding using 10x Genomics linked-read data (SRR18312935) (Bronski et al., 2018) resulted in a final contig count of 7,810 (N50 = 329.6 kb) (Fig. 1B). The assembly was annotated with a custom Funannotate-based pipeline, which included assignment of putative protein functions based on a number of enzyme (e.g., MEROPS (Rawlings et al., 2018), CAZy (Drula et al., 2022)), protein family (e.g., Pfam (Mistry et al., 2021), InterPro (Paysan-Lafosse et al., 2023)) and ontology (e.g., GO (Ashburner et al., 2000; Gene Ontology Consortium et al., 2023), EggNog (Huerta-Cepas et al., 2019), COG (Galperin et al., 2021; Tatusov et al., 2000)) databases as well as prediction algorithms for secretion signals (SignalP (Almagro Armenteros et al., 2019)) and transmembrane domains (see Methods for full details). The final annotation predicted 39,703 genes and 42,657 total transcripts (Fig. 1B).

To glean biological insights from this new genome, we used a comparative approach to place our new genome in the context of other newly-available entomophthoralean data. We collated a dataset consisting of the latest available genomic and transcriptomic data for entomophthoralean fungi, including genomes for the specialist pathogens *E. muscae* (this study), *Entomophaga maimaiga* (ARSEF 7190 v1.0) and *Massospora cicadina* (Stajich JE. et al., 2022), transcriptomes for the specialists *Pandora formicae* (Malagocka et al., 2015) and *Strongwellsea castrans* (this study) and genomes for the generalists *Zoophthora radicans* (Amses et al., 2022), *Neoconidiobolus thromboides* (Chang et al., 2022) and *C. coronatus* (NRRL 28638; Fig. 1C).

To assess completeness of these datasets, their corresponding predicted proteomes were analyzed for the presence of Benchmarking Universal Single-Copy Orthologs (BUSCOs) using a fungal reference set (odb10) (Manni et al., 2021). For *E. muscae*, this analysis detected 81.3% of complete fungal BUSCOs and 5.4% fragmented BUSCOs in the *E. muscae* proteome; 13.3% of fungal BUSCOs were not detected (Fig. 1D). For the other entomophthoralean fungi, percentage of complete-copy BUSCOs ranged from 50.3% (*M. cicadina*) to 87.4% (*S. castrans*), with a median value

of 82.4% across this group (**Fig. 1D**). Notably only the *E. muscae* genome had more double-copy (58.2%) than single-copy (23.1%) BUSCOs. The next highest proportions of double-copy BUSCOs were found in the transcriptomic datasets of *P. formicae* (31.1%) and *S. castrans* (29.7%) which includes partial transcripts and isoforms in the duplication count. Based on BUSCO scores and phylogenetic positions, we selected three other genomes aside from *E. muscae* to use as a core set for genomic analysis: *Entomophaga maimaiga*, *Z. radicans* and *N. thromboides* (**Fig. 1C**).

While the high number of duplicate BUSCOs in the transcriptomes could be accounted for by contaminant host transcripts (each of these transcriptomes was assembled from fungus-killed hosts), this explanation could not account for the high number of duplicate BUSCOs in *E. muscae*, which was assembled from the DNA of fungal cells cultured *in vitro*. One possible explanation for many duplicate BUSCOs in *E. muscae* is that the assembled genome was functionally diploid, not haploid, which has been suggested previously for another *E. muscae* isolate from houseflies (*Musca domestica*) (De Fine Licht et al. 2017). To address this, we counted the occurrence of kmers (29-mers or 33-mers) throughout our assembly (Jellyfish v.2.3.0 (Marçais and Kingsford, 2011)) and plotted the coverage of kmers against their observed frequency (GenomeScope v2.0 (Ranallo-Benavidez et al., 2020)). We observed a large peak ~50x coverage, corresponding to single copy (haploid) kmers and a smaller peak at ~100x coverage, corresponding to dual copy (diploid) kmers. This small ~100x peak suggests that our assembly is heterozygous at a low level (0.023%, 0.026% respectively) but does not support the hypothesis that our genome is fully diploid (**Fig. 1E-S1A,B**). We also addressed the possibility that differences in how genomes were annotated (i.e., annotation “pipeline”) could lead to inflated gene counts, owing to differences in the stringency of calling genes. In order to address this, we annotated the genomes for *E. maimaiga* and *Z. radicans*, which had been annotated with other pipelines, with the pipeline used for *E. muscae*. While our pipeline did predict more genes than the original annotation pipelines for these assemblies (23,807 vs. 14,701 for *E. maimaiga*; 18,761 vs. 14,479 for *Z. radicans*), this difference could not fully account for the discrepancy in gene count (Supplemental Table 1). This suggested that differences in annotation approaches are not sufficient to explain the high duplicate gene count in *E. muscae*.

Proliferation of transposable elements has led to huge genomes within Entomophthorales

At 1.03 Gb, *E. muscae*’s genome is very large compared to the vast majority of other fungal genomes (**Fig. 2A**). Similarly, genomes of the most closely-related entomophthoralean fungi for which genomes are available (*M. cicadina*, *E. maimaiga* and *Z. radicans*) are also extremely large, each exceeding 500 Mb. However, the impressive sizes of *E. muscae* and other entomophthoralean genomes do not appear to be a consequence of increased gene number (**Fig. 2B**). The number of genes predicted in entomophthoralean genomes ranges from 8,867 (*N. thromboides*) to 14,701 (*E. maimaiga*) to 39,711 (*E. muscae*). Gene counts for *M. cicadina* are also within the low end of this range (7,532), but this number is expected to be an underestimate given the fragmented nature of the assembly. In comparison, fungi with genomes between 10-100 Mb have between ~3,000-33,000 genes. Thus, as in other fungi whose genome sizes rank within the top 1% (e.g., the mycorrhizal fungi *Gigaspora margarita* and *G. rosea*, the rusts *Phakospora pachyrhizi*, *Austropuccinia psidii* and *Hemileia vastatrix* and the bioluminescent mushroom *Mycena olivaceomarginata*), the predicted number of genes in entomophthoralean fungi is within the range observed for fungi with more typically-sized genomes (<100 Mb).

Unlike gene content, repeat content across entomophthoralean genomes reveals a clear trend: fungi that shared a common ancestor after divergence from *Conidiobolus* have highly repetitive genomes (**Fig. 2C**). The genomes of *E. muscae*, *E. maimaiga* and *M. cicadina* consist of ~90% repeated sequences (90.9%, 90% and 92.4%, respectively); for *Z. radicans* the proportion of repeated sequences genome-wide is about 20% less, at 71%. Meanwhile, *N. thromboides*’ genome is predominantly non-repetitive (only 13.5% sequences are repeated). The bulk of this repetitive content consists of transposable elements, which are categorized as Class I or Class II based on

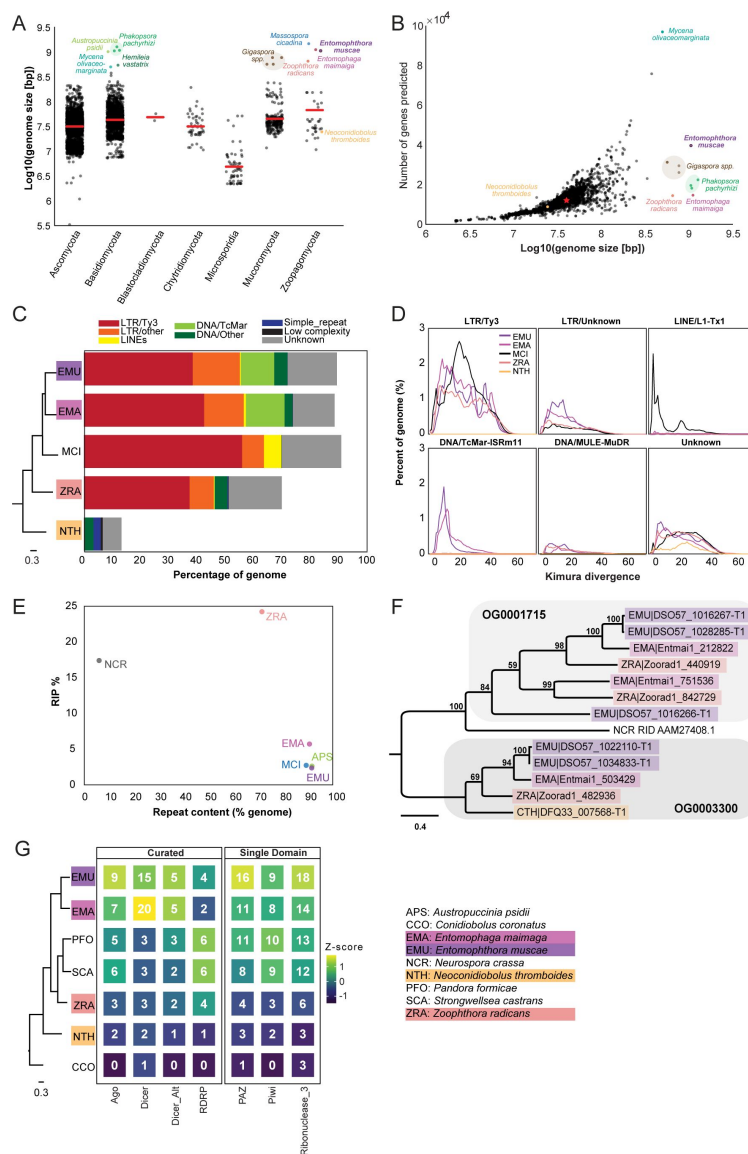


Figure 2.

Entomophthoralean genomes are enlarged (relative to those of other fungi) due to proliferation of transposable elements.

A) Genome assembly sizes of fungal phyla. Red lines indicate mean values per phylum. Species in the core Entomophthorales genomic analysis set are labeled (color-coded per Fig. 1C). B) Observed genome sizes versus predicted gene number in sequenced fungal genomes. Excluding genomes in excess of 500 Mb, the median genome size and number of genes across this dataset is 37.1 Mb and 11,843, respectively (red star). For both A and B, genomes exceeding 500 Mb are indicated by dots colored by species identity. Data for A and B available in Supplementary File S1. C) Repeat element composition within *E. muscae* (EMU), *E. maimaiga* (EMA), *M. cicadina* (MCI), *Z. radicans* (ZRA) and *N. thromboides* (CTH) as determined by RepeatMasker (Smit et al., 2013-2022). Only repeat elements that exceeded 0.1% of the genome for at least one species are shown. Cladogram modified from Fig. 1C. D) Landscapes of DNA repeat elements with those comprising less than 1% of DNA repeats in each genome binned as other. E) Percentage of genome in which RIP was detected versus percent of genome comprised of repeat content. F) Protein phylogeny representing relationships among two methyltransferase-containing orthogroups and RID from *Neurospora crassa*: a cytosine methyltransferase required for RIP (Freitag et al., 2002). Scale bar indicates 0.3 substitutions per site. G) Counts of selected Pfams associated with RNAi pathway components across genomes. Curated counts include only candidates with the expected combination (and frequency) of Pfams for each of the listed RNAi pathway proteins. Cladogram modified from Fig. 1C.

their mechanism of action. Class I transposons (retrotransposons) “jump” via transcription to an RNA intermediate that is then reverse transcribed to DNA and integrated in a target genomic site. Class II transposons (DNA transposons) can “jump” by excision from donor DNA and reinsertion into a target genomic site or “copy” when single-stranded excisions are repaired by host machinery. The majority of repeat content across these four family Entomophthoraceae genomes consists of Class I retrotransposons, specifically Ty3 (formerly called Gypsy (Wei et al., 2022 [↗](#))) long terminal repeat (LTR) retrotransposons. Ty3 elements comprise 39.1%, 43.2%, 56.8% and 38.0% of the *E. muscae*, *E. maimaiga*, *M. cicadina* and *Z. radicans* genomes, respectively, while only 0.01% of the *N. thomoboides* genome. In addition, the most closely-related species pair in this set, *E. muscae* and *E. maimaiga*, also have a sizeable fraction (12% and 13.9%, respectively) of Tc1 mariner DNA (Class II) transposons populating their genomes. *M. cicadina* is unique among the other entomophthoralean fungi in having a substantial fraction (6.3%) of L1 long, interspersed nuclear elements (LINEs) while effectively lacking any DNA transposons.

Kimura divergence estimates for most populous repeat elements of the *E. muscae*, *E. maimaiga*, *M. cicadina*, *Z. radicans* and *N. thomoboides* genomes show trends consistent with phylogenetic relationships between these fungi (Fig. 2D [↗](#)). All but *N. thomoboides* have expanded Ty3 and unknown LTR elements with comparable Kimura divergences (~0-50), indicating similar proliferation of these elements occurred with approximately concurrent timing. *M. cicadina* uniquely shows a burst of Ty3 LTR expansion around divergence time 20 and an expansion of LINE elements, occurring after Ty3 expansion. *E. muscae* and *E. maimaiga* show an increase in TcMar elements more recently than either of the Ty3 and LINE expansions. In addition, *E. muscae*, *E. maimaiga* and *Z. radicans* share an expansion of DNA-MULE-MuDR elements that is not observed in *M. cicadina*. All five genomes also possess putative repeat elements currently classified as “Unknown”. These elements appeared roughly concurrently.

Fungi are known to counter the proliferation of TEs by mutating cytosine to thymine in repetitive regions via a process called Repeat Induced Point (RIP) mutations (Selker, 2002 [↗](#); van Wyk et al., 2020 [↗](#)). This process occurs during meiosis in fungi from Ascomycota and Basidiomycota, where homology between DNA regions directs this transition (Hood et al., 2005 [↗](#); van Wyk et al., 2020 [↗](#)). We calculated dinucleotide indices to look for evidence of RIP in entomophthoralean fungi, along with *Neurospora crassa*, an ascomycete fungus in which RIP has been extensively documented (Freitag et al., 2002 [↗](#); Galagan et al., 2003 [↗](#); Lewis et al., 2009 [↗](#); Selker, 2002 [↗](#)), and *Austropuccinia psidii* (Tobias et al., 2021 [↗](#)), a basidiomycete rust with a genome of similar size to entomophthoralean fungi (Fig. 2E [↗](#)). Consistent with previous studies, *N. crassa* showed high levels of RIP (17.4%) (Cambareri et al., 1991 [↗](#)). Surprisingly, RIP in *Z. radicans* occurred at an even higher rate (24.2%) than in *N. crassa*. We also detected RIP at low levels in the rust *Austropuccinia psidii* (2.6%), *E. muscae* (2.4%), *M. cicadina* (2.7%), and *E. maimaiga* (5.7%). This analysis argues that RIP occurs in *Z. radicans* and could be occurring at a low level in other entomophthoralean fungi as well.

A cytosine methyltransferase named RIP defective (RID) is required for RIP (Freitag et al., 2002 [↗](#)), one of several families of DNA methyltransferases in Fungi (Bewick et al., 2019 [↗](#)). RID contains a DNA methyltransferase region that is recognized by HMMer as two adjacent Pfam DNA_methylase (PF00145) domains, so we leveraged this characteristic domain profile to identify proteins that share this domain architecture. This revealed two orthogroups (OG0001715 and OG0003300) containing candidates with two DNA_methylase domains. These orthogroups were aligned with RID protein from *Neurospora crassa*, and *N. crassa* RID is sister to the OG0001715 clade in a tree of these proteins, albeit at low (12-14%) sequence identity (Fig. 2F [↗](#)). OG0001715 is composed of proteins from *Z. radicans*, *E. maimaiga* and *E. muscae*. Each of these species has two paralogs in this orthogroup, except *E. muscae*. *E. muscae* intriguingly has an additional, partial paralog, (DSO57_1016266) of about half the size, which is in tandem with another, complete *E. muscae* OG0001715 paralog (DSO57_1016267). This analysis suggests that *E. muscae*, *E. maimaiga* and *Z. radicans* contain multiple proteins with similar domain architecture as RID, with some direct

evidence for duplication of these methyltransferases. We report here on our search for methyltransferases with similar domain signatures to RID, but note that these candidates cannot be interpreted as homologs of RID: their function warrants further investigation.

Many eukaryotic organisms also employ RNA interference (RNAi) to control retrotransposon activity. RNAi occurs when double-stranded RNAs are processed into small interfering RNAs (~20-30 nucleotides long) that are used to guide degradation (or, in the case of mRNA, repression of translation) of complementary RNA. In Fungi, there are three proteins essential for RNAi: 1) RNA-dependent RNA polymerase (RDRP), which processes ssRNA into dsRNA, 2) Dicer, the enzyme that processes dsRNA into siRNA, and 3) Argonaute, a protein that loads siRNA to form an RNA-induced silencing complex (RISC) that identifies target RNAs by complementarity (Nicolás and Garre, 2016 [\[4\]](#)). We performed a domain-based analysis to assess if RNAi machinery had been lost in entomophthoralean fungi, which could allow transposable elements to proliferate (Fig. 2G [\[4\]](#)). This analysis suggests that at least one homolog of each of the core components of the fungal RNAi pathway (RDRP, Dicer and Argonaute) is present across our core species, with *E. muscae* and *E. maimaiga* predicted to have several homologs of each of these genes. Thus, loss of RNAi pathway genes seems unlikely to account for transposable element proliferation in these species.

Trends in protein domains across Entomophthorales

We next compared the coding regions of the entomophthoralean datasets in order to identify functional trends across these fungi. Since all of these species are insect pathogens, we expected to observe common themes amongst their metabolic and secreted functions related to their utilization of host resources and interactions with the host immune system. First, we performed protein domain analysis, looking broadly at protein families (Pfam), and more specifically at functional domains involved in metabolizing carbohydrates (using CAZy) and proteins (using MEROPS). For these analyses we used all available entomophthoralean datasets (Fig. 1C [\[4\]](#)) except for the *M. cicadina* genome. *M. cicadina* was excluded because the fragmented nature of its genome assembly precluded an accurate prediction of this fungus' proteome, a key requirement for this analysis (Supplemental Table 2).

Protein family domains (Pfam)

Pfam families encompass diverse proteins with a broad range of functions (El-Gebali et al., 2019 [\[4\]](#)). Enrichment analysis among these domains revealed a unique enrichment in *E. muscae* of retroviral families (i.e., TE-related), specifically Asp_protease (PF09668), Asp_protease_2 (PF13650), gag-aspartate (PF13975), RVP_2 (PF08284) and dUTPase (PF00692) domains (Fig. 3A [\[4\]](#)). The domain zf-CCHC (PF00098) was found to be similarly enriched in *E. muscae* and may represent a retroviral regulatory protein. The largest family that was enriched in *E. muscae* was Lipase_3 (PF01764; 348 in *E. muscae*; 5-fold higher than the median across genomes: 66), which was also enriched in *E. maimaiga* and *Z. radicans* (132 and 190, respectively).

A much smaller number of Pfam domains were found to be significantly underrepresented in *E. muscae* (20) compared to the number of domains found to be significantly overrepresented (105) (Fig. 3B [\[4\]](#)). Proportionally more of these underrepresented families were not predicted to be secreted (17/20, compared to 45/105 of overrepresented Pfam families, $p=0.00055$ per chi-squared test). Three of these not-secreted, underrepresented families in *E. muscae* were underrepresented by greater than 5-fold: alpha/beta hydrolase fold (abhydrolase_3; PF07859; 6.5-fold lower than median), bifunctional feruloyl and acetyl xylan esterases (BD-FAE; PF20434; 20-fold) and endonuclease/exonuclease/phosphatase (Exo_endo_phos_2; PF14529; 13-fold). Alpha/beta hydrolase folds are found as catalytic domains in many different enzymes, including BD-FAE (PF20434) proteins, which specifically act on complex xylans. The BD-FAE family was significantly underrepresented in *E. muscae*, *E. maimaiga* and *Z. radicans* (1, 2 and 3 BD-FAE-containing

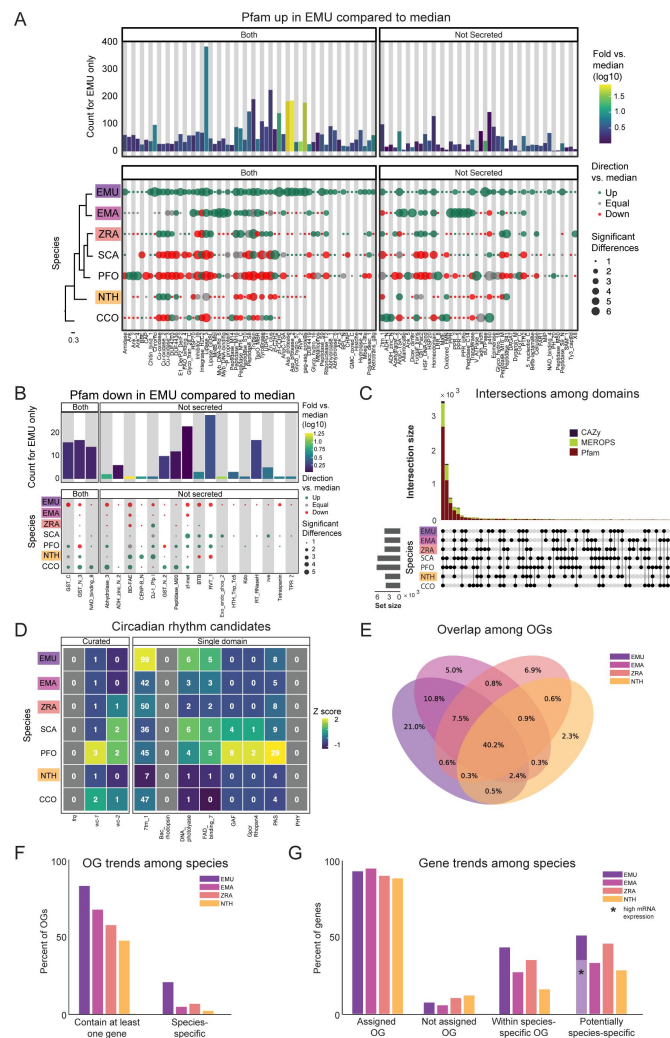


Figure 3.

Comparison of domain architecture/gene content across Entomophthorales.

A) Pfams significantly overrepresented in *E. muscae* (EMU) compared to other species analyzed (*E. maimaiga* (EMA), *Z. radicans* (ZRA), *S. castrans* (SCA), *P. formicae* (PFO), *N. thromboides* (NTH) and *C. coronatus* (CCO)). Bars represent the counts for *E. muscae* colored by fold-versus-the-median across all genomes. Point size represents the number of significant pairwise comparisons among other genomes and these are colored according to whether the value is above, below or equal to the median value across all genomes. Cladogram modified from Fig. 1C. B) Pfams significantly underrepresented in *E. muscae* compared to other species analyzed. Plot format as in (A). C) Combined UpSet plots showing the intersection among genomes for all domain categories (CAZy, Pfam and MEROPS; bar colors). D) Counts of selected Pfams associated with circadian proteins across genomes. Curated counts include only candidates with the expected combination of Pfam domains for each of the listed circadian proteins. (For example, to be considered a curated *wc-1* candidate, a gene needed to have one each of GATA, PAS_3 and PAS_9 domains.) E) Venn diagram depicting intersections between predicted OGs among *E. muscae*, *E. maimaiga*, *Z. radicans* and *N. thromboides*. Values only within a single ellipse indicate the abundance of species-specific genes. F) Occupancy trends by species for 11,711 OGs recovered from amalgating *E. muscae*, *Entomophaga maimaiga*, *Z. radicans* and *N. thromboides* gene models. "Contain at least one gene": percentage of OGs that contain at least one gene from a given species. "Species-specific" is the percentage of OGs that only contain genes from one species. G) Percentage of genes for each species found in each OG type as percentages of total genes annotated in the genome. "Assigned OG" are genes that clustered with an OG; "Not assigned OG" are genes that did not cluster with any OG. "Within species-specific OG" reflects the percentage of genes that fall within an OG that is only populated by genes of the given species. "Potentially species-specific" is the sum of genes "Not assigned OG" and "Within species-specific OG". The light purple bar marked with a black asterisk indicates the percentage of genes that are potentially species-specific with evidence of high expression in an *in vivo* dataset (expression < 5, pooled dataset of 27 whole fly samples exposed with *E. muscae*).

proteins, respectively), compared to *C. coronatus* and *N. thromboides* (23 and 25, respectively). *E. muscae*, with only a single predicted BD-FAE-containing protein, was additionally significantly lower than *P. formicae* and *S. castrans*, which had 20 and 23, respectively.

We compared the broad Pfam composition of these seven genomes (**Fig. 3C**). 2,684 Pfam domains were present in all genomes. *S. castrans* and *P. formicae* shared a unique overlap of 1,142 domains, each with 170 and 485 unique domains, respectively. *E. muscae*, by comparison, only had 14 unique Pfam domains. The trend of *P. formicae* and *S. castrans* sharing many domains was seen across all domain types. However, the proteomes for these fungi were uniquely generated from transcriptomic data, so we were concerned this pattern may reflect a methodological, not biological, signal. To test this, we recapitulated our Pfam domain analysis incorporating a proteome generated from transcriptomic data collected from *E. muscae*. The results of this analysis are summarized in an UpSet plot (**Fig. 3C**-S1), in which the *E. muscae* transcriptomic dataset does not meaningfully group with *P. formicae* and *S. castrans*. This suggests that the unique subset shared between *P. formicae* and *S. castrans* is likely due to their evolutionary history, not a methodological artifact.

Carbohydrate metabolic domains (CAZy)

The CAZy data set (named for Carbohydrate-Active Enzymes) provides insight into which carbohydrates can be metabolized. The majority (55) of CAZy domains were shared among all fungi, with *S. castrans* and *P. formicae* sharing a large set of CAZy enzymes (28) absent in the other species. This unique pattern for *S. castrans* and *P. formicae* was consistent across all types of domains and shows a clear signal on the UpSet plot comparing domains (**Fig. 3C**). *E. muscae* only had three unique CAZy domains (**Fig. 5A**): GH134 (containing endo- β -1,4-mannanases), GH5 (glucanases/cellulases) and PL12 (heparin-sulfate lyases). Six CAZy domains were found to be significantly enriched in *E. muscae*, including AA11 (lytic chitin monooxygenases), AA7 (glucosylated oligosaccharide oxidases), CBM18 (chitin-binding and chitinases), CE16 (acetyl esterases), CE4 (acetyl xylan esterases) and GT1 (UDP-glucuronosyltransferase) (**Fig. 3C**-S2A). AA11 and CE16 were notably significantly enriched in *E. muscae*, *E. maimaiga* and *Z. radicans* and significantly lower in *P. formicae* and *S. castrans*.

Peptidase domains (MEROPS)

All genomes included in our domain analysis shared 691 MEROPS peptidase domains (**Fig. 3C**). *S. castrans* and *P. formicae* shared 448 MEROPS domains absent in the other genomes, with each having a large number of unique MEROPS domains (173 and 136, respectively). *E. muscae* by contrast had about a third to a quarter as many unique MEROPS peptidases (37). Among *E. muscae* MEROPS, 92 were found to be significantly different in comparisons among genomes, with 74 MEROPS domains significantly enriched and 9 domains having significantly fewer proteins (**Fig. 3C**-S2B). A clear trend for *E. muscae*, *E. maimaiga* and *Z. radicans* is enrichment in M16A metallopeptidases (i.e., MER0001214, MER0001218, MER0002283, MER0002345, MER0002423, MER0003386, MER0003823, MER0011096, MER0011744, MER0015259, MER0169735) compared to *S. castrans*, *P. formicae* and *N. thromboides*. Considering proteins containing these 10 M16A domains, *E. muscae* (62 on average), *E. maimaiga* (19) and *Z. radicans* (44) were comprised of approximately 11-fold more proteins (43 vs 4) compared to *S. castrans* (5), *P. formicae* (3) and *N. thromboides* (2). Two metallopeptidases were uniquely enriched in *E. muscae*: MER0001120, a pro-collagen C-peptidase, and MER0016735, an M23 peptidase. Among serine peptidases, a lipid monooxygenase (MER0031618), two S33 prolinases (MER0036050 and MER0036051) and three kexin/kexin-like (MER0000364, MER0000374 and MER0001604/krp1) domains were found to be significantly enriched. Additionally, *E. muscae* was uniquely enriched in the monoglyceride lipase MER0045883. Considering cysteine peptidases, *E. muscae* had a high number of metacaspases across all 10 families, though fewer than the number seen in *E. maimaiga* (~50 accessions for each family).

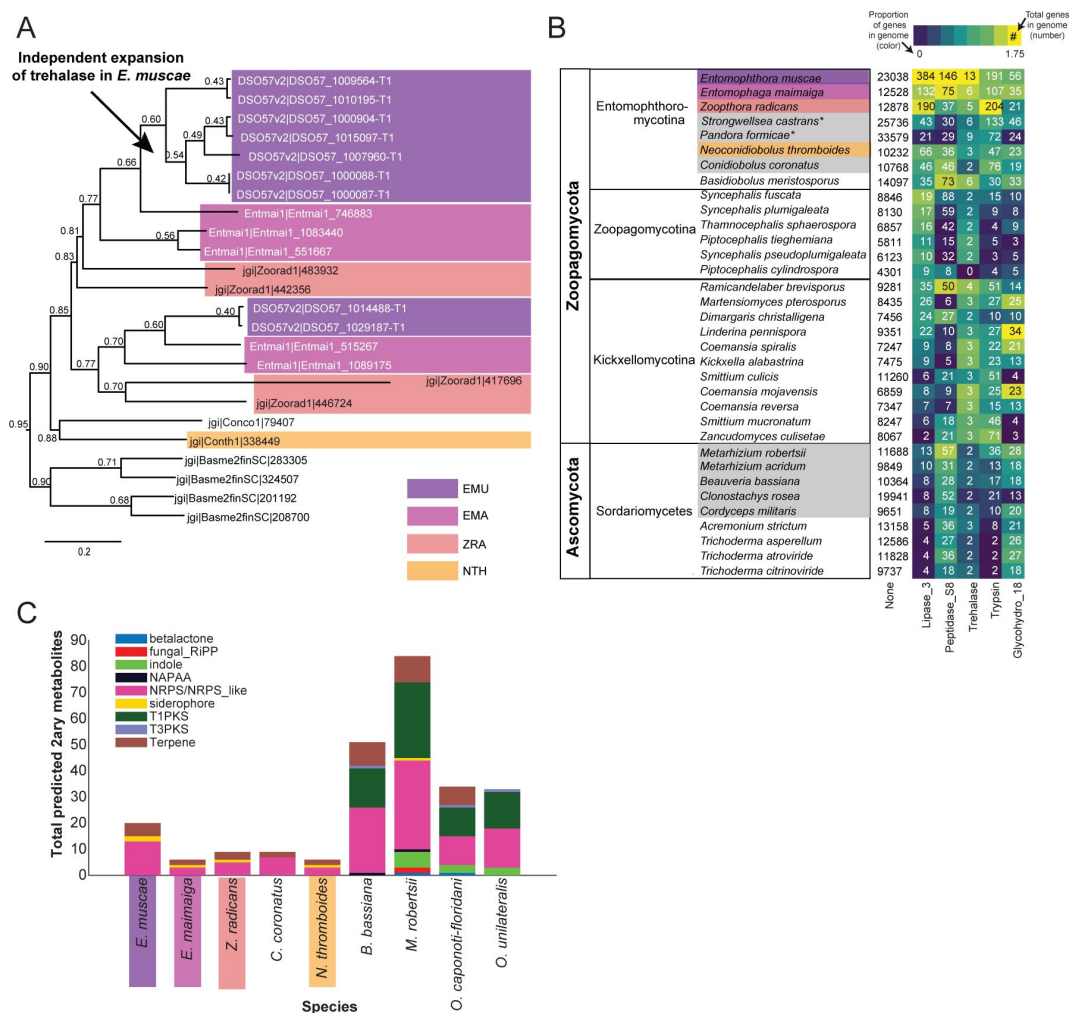


Figure 4.

Gene family expansion and secondary metabolite production of *E. muscae* and other insect pathogens.

A) A family of genes encoding extracellular trehalase enzymes (PF01204) is expanded in *E. muscae* (EMU) compared to other Zoopagomycetes: *E. maimaiga* (EMA), *Z. radicans* (ZRA), *N. thromboides* (NTH), *C. conidiobolus* (Conco1), and *B. meristosporus* (Basme2finSC). B) Total number of genes and number of genes encoding Lipases (*Lipase_3*), Subtilisin-like serine peptidases (*Peptidase_S8*), Trehalases (*Trehalase*), Trypsins (Trypsin), and Chitinases (*Glycohydro_18*) in representative fungal species of Zoopagomycota and Ascomycota. Fungal species in gray are insect pathogens and the four Entomophthoromycotina species are outlined in the same colors as (A). Numbers inside heatmap refer to the number of genes that encode a given Pfam domain, and color scale refers to proportion of genes with a given Pfam compared to the total number of genes in the genome (in percentages). C) Predicted secondary metabolite production for select entomophthoralean genomes (*E. muscae*, *E. maimaiga*, *Z. radicans*, *C. coronatus* and *N. thromboides*) and common ascomycete entomopathogens (*B. bassiana*, *M. robertsii*, *O. caponoti-floridani*, *O. unilateralis*), as predicted by AntiSMASH. Color indicates metabolite class.

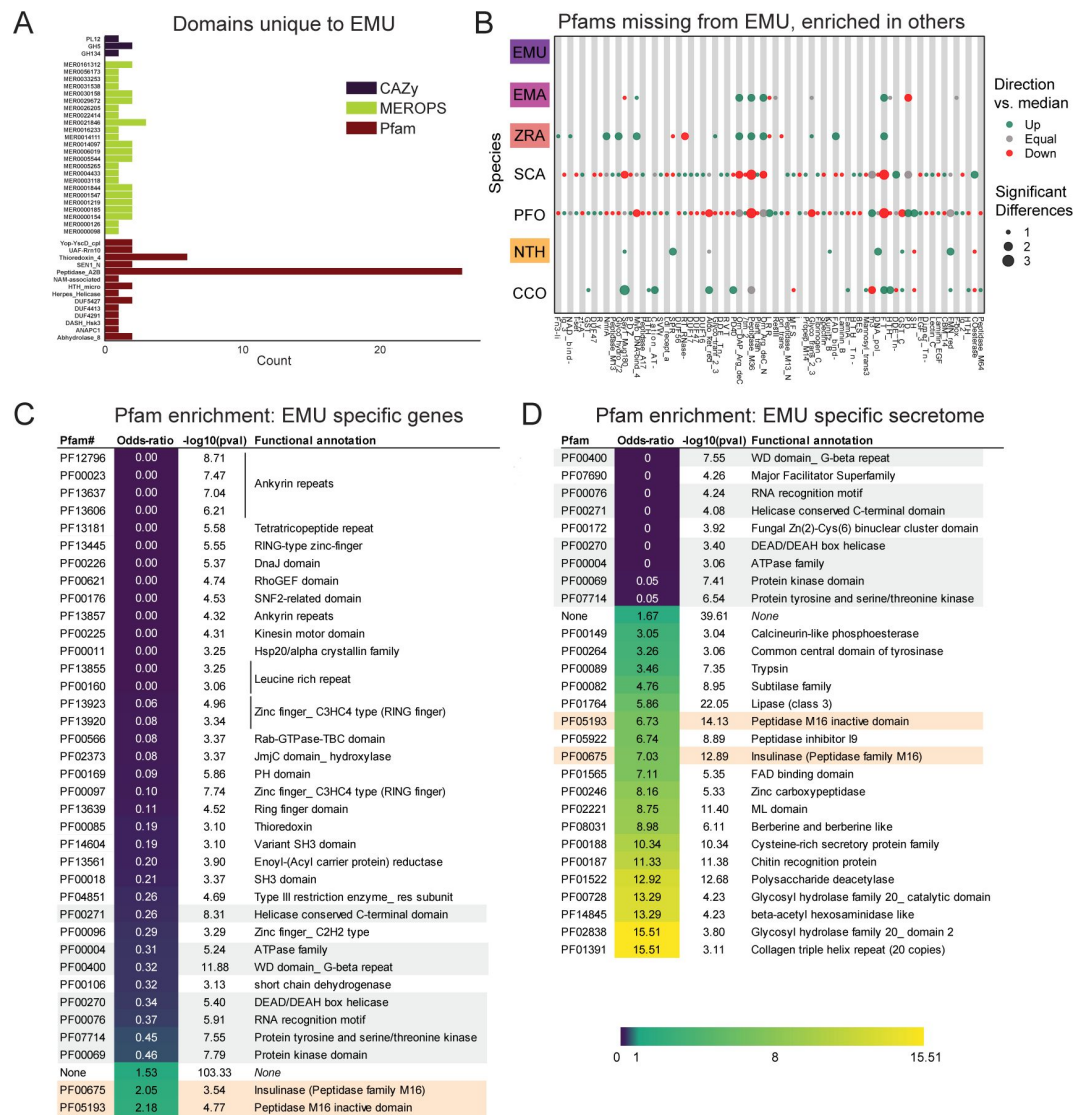


Figure 5.

Unique features of *E. muscae* compared to *E. maimaiga*, *Z. radicans* and *N. thromboides*.

A) Domains unique to *E. muscae*. B) Pfam domains that are missing in *E. muscae*, but enriched in other entomophthoralean fungi. C) Significantly enriched Pfam domains ($p \leq 0.001$) within genes that are potentially *E. muscae*-specific (both genes that did not cluster with any orthogroup and genes that cluster with orthogroups that are species-specific; $N = 9,150$ genes). D) Significantly enriched Pfam domains ($p\text{-value} \leq 0.001$) within potentially *E. muscae*-unique genes encoding proteins predicted to be secreted ($N=1,685$ genes). Odds-ratios are colored according to the scale bar to bottom right. Two Pfam domains (PF00675 and PF05193; highlighted orange in C & D) are overrepresented in potentially *E. muscae*-specific *E. muscae* genes both genome-wide and within the predicted secretome. Pfam domains highlighted in gray are underrepresented across both of these sets.

Only a handful of domains were found to be significantly underrepresented in *E. muscae*, including aminopeptidase H11 (MER0002003), cystinyl aminopeptidases (MER0002060) and ERAP2 aminopeptidases (MER0002968) (Fig. 3 [↗](#)-S2C).

Evidence for circadian pathways

We also examined genes that may be involved in the ability to sense environmental light cues and maintain circadian time, as these processes are likely involved in fungal manipulation of host behavior (de Bekker et al., 2021 [↗](#); de Bekker and Das, 2022 [↗](#)). The circadian clock of *Neurospora crassa* (Ascomycota) is the most highly studied and best understood circadian clock among Fungi (see (Cha et al., 2015 [↗](#)) for overview). In *N. crassa*, two photosensitive proteins, White Collar 1 and 2 (WC-1 and WC-2) form a heterodimeric complex (called the White Collar Complex, WCC). In the presence of light, WCC undergoes a conformational change and becomes a transcriptional activator of light-regulated genes, including the gene *frequency* (*frq*) that encodes an intrinsically-disordered protein. FRQ protein forms a heterodimeric complex with its partner, FRQ-interacting RNA Helicase (FRH), and this complex, called FFC (FRQ-FRH Complex), inhibits the activity of WCC. After protein synthesis, FRQ becomes progressively more phosphorylated, eventually leading to its degradation and reduced concentration of FRQ and FFC. When FRQ titers drop below a certain threshold, WCC activity can proceed unimpeded, thus closing an oscillating cycle of transcription and translation with a periodicity of about 24 hours.

To identify potential components of circadian oscillators in our entomophthoralean datasets, we checked for putative homologs of *wc-1*, *wc-2* and *frq* and domains commonly found in circadian proteins, per-ant-sim or PAS (PAS_3: PF08447) and the structurally-similar domain GAF (GAF: PF01590) (Fig. 3D [↗](#)). All of the analyzed entomophthoralean datasets had at least one copy each of a gene encoding for a protein with the same domains found in *N. crassa* WC-1, with *C. coronatus* and *P. formicae* containing multiple genes (the latter even having two candidate WC-1 genes in tandem, Pfor_14152, Pfor_14153). For *wc-2*, all but *E. muscae*, *Entomophaga maimaiga* and *N. thromboides* also had at least one candidate gene. None of these species contained a gene encoding for a protein with the hallmark domains of *frq* (Fig. 3D [↗](#)).

In addition to finding evidence for at least one copy of a *wc-1* homolog for all of the species analyzed, we also searched for genes encoding for known light-sensitive domains (rhodopsin family [7tm_1; PF00001], bacteriorhodopsin [Bac_rhodopsin; PF01036], DNA photolyase [DNA_photolyase; PF00875], the FAD-binding domain of DNA photolyase [FAD_binding_7; PF03441] and phytochrome [PHY; PF00360], GPCR rhodopsin 4 [GpcrRhopsn4; PF10192])). We did not detect any bacteriorhodopsins or phytochromes within our datasets, but we found several instances of genes containing the other light-sensitive domains, with *E. muscae* having far more than the rest (99 genes with rhodopsin family domains as well as 11 photolyases or FAD-binding domains thereof) (Fig. 3D [↗](#)). This analysis suggests that each of these fungi detect light. As light is the predominant cue for setting the phase and period of circadian clocks, this is consistent with the notion that these fungi may also be able to keep time.

Orthogroup analysis reveals core functionalities across Entomophthorales fungi

We next assessed gene homology among the core genomes in our comparative dataset: *E. muscae*, *Entomophaga maimaiga*, *Z. radicans* and *N. thromboides* (Supplemental Table 2). For this analysis, we used OrthoFinder to identify orthogroups (sets of genes that are descended from a single gene in the last common ancestor of all species within our analysis set, including both orthologs and paralogs, referred to henceforth as OGs) among the pooled genes from this set of genomes (Emms and Kelly, 2015 [↗](#)). We reasoned that such an analysis could be a strategy to identify core genes important for an obligate entomopathogenic lifestyle, and conversely, could reveal potentially

unique genes in each of the species examined, specifically in our focal species *E. muscae*. We recovered 17,111 OGs. 6,878 OGs (40.2%) contained representatives in all four species (**Fig. 3E**), and we refer to these as “core” OGs.

To gain some insight into the function of the core OG genes in *E. muscae*, we performed a Pfam annotation enrichment analysis. Twelve Pfam domains were significantly overrepresented in these genes (**Fig. 3**–S3) by factors ranging from 1.52 to 2.35-fold. These domains are found in proteins across a variety of principal cellular functions, spanning DNA replication, transcription, macromolecular assembly and signal transduction and included kinases (PF0069, PF07714), RNA recognition and binding proteins (PF00076), helicases (PF00400, PF00271, PF04851), endonucleases (PF04851), ATPases (PF00004), Beta-propeller (PF00400), phosphatidyl inositol-binding (PF00169), RNA splicing (PF00176), and DNA replication (PF00226). Genes that lacked a Pfam assignment altogether were underrepresented (odds-ratio 0.46), consistent with core OG genes belonging to conserved cellular processes among these four species. Sixteen Pfam domains were significantly underrepresented, ranging from being completely absent to 18.3 – 1.83-fold underrepresented (odds-ratio 0.055 to 0.55, respectively). Protein domains that were completely absent included chitin recognition (PF00187), transferase (PF02458), multicopper oxidases (PF00394 and PF07732) and polysaccharide deacetylase (PF01522). Underrepresented domains tended to be related to metabolism and transcription factors (PF00046 and PF00172). However, underrepresented genes also included the retrotransposon gag protein, which is involved in activity of retrotransposons (such as LTRs), and rhodopsin family genes.

The next largest overlap consisted of 10.8% of OGs that are shared only between the most closely related species in this set, *E. muscae* and *Entomophaga maimaiga* (**Fig. 3E**). OGs shared between *E. muscae*, *Entomophaga maimaiga* and *Z. radicans*, but not the relative outlier *N. thromboides* totaled 7.5%, with the next largest overlapping set consisting of 2.4% of OGs that are shared between *E. muscae*, *Entomophaga maimaiga* and *N. thromboides*, but not with *Z. radicans*. Other overlapping sets individually accounted for no more than 0.9% of all OGs. The highest proportion of species-specific OGs (OGs that are populated by genes of a single species) was found for *E. muscae* (21%), followed by *Z. radicans* (6.9%), *Entomophaga maimaiga* (5%) and *N. thromboides* (2.3%). Accordingly, we observed that *E. muscae* genes also populate the most OGs overall, though has genes in multi-species OGs no more frequently than *Entomophaga maimaiga* (**Fig. 3F**). These observations are consistent with the *E. muscae* genome containing about three times as many annotated genes as any other fungus considered.

Across these four species, 88% to 94.4% of genes were assigned to OGs (**Fig. 3G**). With respect to species-specific OGs, *E. muscae* has the most: 43.5% of *E. muscae* genes are found in these OGs. The species with the least, *N. thromboides*, had just 16.3%. For each species, the genes that either failed to cluster with an OG (were not assigned) together with the genes that are found in species-specific OGs comprise a set of genes that are potentially unique to that species. If we consider all of the annotated genes in the *E. muscae* assembly, *E. muscae* has the highest percentage of species-specific genes among this species set at 50.9% (**Fig. 3G**). However, if we filter our annotated genes using a pooled *in vivo* transcriptomics dataset (Elya et al., 2018), we find that the percentage of species-specific genes for *E. muscae* falls to 35%, within the range of the other fungal genomes analyzed.

Gene expansions and specializedmetabolite production in *E. muscae* and *Entomophthorales*

Because *E. muscae* is an obligate insect-pathogen (e.g., can only complete its life cycle within live flies), we investigated the presence of canonical entomopathogenic enzymes in the genome. We found that *E. muscae* appear to have an expanded group of acid-trehalases compared to other entomopathogenic and non-entomopathogenic *Entomophthorales* (**Fig. 4A**), which correlates with the primary sugar in insect blood (hemolymph) being trehalose (Thompson, 2003). The

obligate insect-pathogenic lifestyle is also evident when comparing the repertoire of lipases, subtilisin-like serine proteases, trypsins, and chitinases in our focal species versus Zoopagomycota and Ascomycota fungi that are not obligate insect pathogens (Fig. 4B). Sordariomycetes within Ascomycota contains the other major transition to insect-pathogenicity within the kingdom Fungi (Araújo and Hughes, 2016). Based on our comparison of gene numbers, Entomophthorales possess more enzymes suitable for cuticle penetration than Sordariomycetes (Fig. 4B). In contrast, insect-pathogenic fungi within Hypocreales possess a more diverse secondary metabolite biosynthesis machinery as evidenced by the absence of polyketide synthase (PKS) and indole pathways in Entomophthorales (Fig. 4C).

Features of *E. muscae* potentially distinct among Entomophthorales

To investigate unique characteristics of the *E. muscae* proteome, we identified all domains that were either unique to (Fig. 5A) or missing from (Fig. 5B, Fig. S5-1) *E. muscae*. Among the 41 domains unique to *E. muscae* (Fig. 5A), the Pfam family Peptidase_A2B (PF12384), a component of Ty3 retrovirus-like elements, was by far the most common (26 proteins). Many Pfam families that were uniquely found in *E. muscae* were apparently related to retrotransposons. Thioredoxin_4 (PF13462) was the second-most abundant unique Pfam family (6 proteins). Among other *E. muscae*-specific domain families, two proteins were annotated as containing abhydrolase_8 domains (PF06259), perhaps accommodating for the apparent significant reduction in abhydrolase_3 domains in *E. muscae* (Fig. 3B). Nearly all MEROPS families that were unique to *E. muscae* were serine peptidases. A single metallopeptidase (MER0001219; M16A family Ax11 peptidases), threonine peptidase (MER0026205; T3 family) and cysteine peptidase (MER0014097; C19 family) were found uniquely in *E. muscae*. The most abundant MEROPS family that was unique to *E. muscae* was a chymotrypsin family (3 proteins; MER0021846). Only three CAZy domains were uniquely found from *E. muscae*: GH134 (containing endo- β -1,4-mannanases), GH5 (glucanases/cellulases) and PL12 (heparin-sulfate lyase).

In total, 2,237 Pfam domains were absent in the *E. muscae* genome but observed in other genomes within our core set of entomophthoralean fungi. To assess loss of key domains in *E. muscae*, we considered domains that were missing from *E. muscae*, but were significantly enriched in other entomophthoralean fungi (Fig. 5B, Fig. S5-1). We did not observe any domains across Pfam, MEROPS and CAZy that were both missing from *E. muscae* and overrepresented across all other entomophthoralean fungi, suggesting that *E. muscae* has not lost any functions specific to the Entomophthorales. Our domain analysis suggests that *E. muscae* is unique largely through its protein-domain family expansions, not in its losses of domains. Many domains, of all types in our analysis, that were not enriched (or not present) in *E. muscae* were enriched only in *P. formicae* and *S. castrans*, suggesting this signal is reflective of functionalities unique to those species, rather than functionalities lost in *E. muscae*.

We also performed an enrichment analysis on the orthogroups that were classified as potentially species-specific for *E. muscae*. As in our previous OG analysis, we considered genes to be potentially species-specific if they either failed to cluster with an OG in *E. maimaiga*, *Z. radicans* or *N. thromboides* or clustered with OGs comprising only other *E. muscae* genes. We performed an enrichment analysis of Pfam annotations of genes that met these criteria, comparing the frequency of Pfam occurrence in this set against occurrence within all annotated genes (Fig. 5C). We observed significant overrepresentation of three Pfams among potentially *E. muscae*-specific genes: insulinases (PF00675) M16 peptidases (PF05193) and genes that lacked any Pfam annotation. That *E. muscae* may have specific peptidases is consistent with our understanding of the entomopathogenic lifestyle: these enzyme families are both key in host recognition and invasion as well as host resource utilization (Arnesen et al., 2018). Finding genes lacking any Pfam annotation as another enriched group is consistent with the large number of genes observed in *E. muscae* compared to other species, and suggests that there are novel genes of undescribed/unknown function specific to *E. muscae*.

Pfams that were underrepresented in the set of potentially *E. muscae*-specific genes spanned multiple processes, but a few general themes emerged. These include a lack of proteins with cytoskeletal (e.g., ankyrin [PF12796, PF00023, PF13637, PF13606, PF13857], kinesin [PF00225]), cell signaling (e.g., RhoGEF [PF00621], Rab-GTPase [PF00566], proteins tyrosinase and kinase [PF07714, PF00069], thioredoxin [PF00085], pleckstrin homology domain [PF00169]), and transcriptional regulation functions (e.g., zinc fingers [PF13445, PF13923, PF13920, PF00097, PF13639, PF00096], SNF2-related [PF00176], DEAD box helicase [PF00270], RNA recognition motif [PF00076], JmjC [PF02373]). The underrepresentation of these domains in potentially *E. muscae*-specific genes suggests their functionalities are similar to those in other entomophthoralean species.

We performed a similar enrichment analysis looking at the set of potentially *E. muscae*-unique genes that are predicted to encode secreted proteins (**Fig. 5D**). We observed several Pfams that were underrepresented within the predicted secretome, many which encompassed functions that are not expected to serve the fungus in an extracellular context (e.g., domains involved in gene transcription [PF00076, PF00172, PF00270] and ATP production [PF00004]). We also observed a few underrepresented Pfam domains involved in signaling [PF00069, PF07714]). Pfams could be underrepresented in *E. muscae*-specific secreted proteins either because similar proteins are present across other entomophthoralean fungi considered or because proteins containing these domains are not typically secreted. We suspect the Pfams involved in signaling are underrepresented due to their similarity to proteins in other entomophthoralean fungi, while Pfams involved in intrinsic (i.e., not extracellular) processes are underrepresented due to a both unlikelihood of being secreted and conserved function among species.

Pfams that were significantly overrepresented in the predicted *E. muscae*-unique secretome included various catabolic enzymes (e.g., phosphoesterase [PF00149], tyrosinase [PF00264], proteases including trypsin [PF00089], subtilase [PF00082], peptidase inhibitor I9 (known constituents of subtilisins) [PF05922], zinc carboxypeptidase [PF00246], insulinase [PF00675], lipase [PF01764], polysaccharide deacetylase [PF01522], glycosyl hydrolases [PF00728, PF02838]) and macromolecular recognition domains (e.g., chitin recognition, ML domain). The observation that catabolic enzymes are overrepresented within the putatively *E. muscae*-specific predicted secretome is consistent with our current understanding of entomopathogen host interactions (Elya and De Fine Licht, 2021). These enzymes are important during fungal infection, where they help degrade chitin-linked proteins in the cuticle when the fungi force their way inside insects. Both during growth on the cuticle and when proliferating inside the living insects *E. muscae* use enzymes such as peptidases, glycosyl hydrolases etc. to obtain nutrients that support fungal growth in the hemocoel (Elya and De Fine Licht, 2021). In addition, we saw enrichment of the collagen triple helix repeat domain (PF01391), FAD binding domain (PF01565) and berberine-like proteins (PF08031). As with our analysis of all *E. muscae*-unique genes, an overrepresentation of genes that were not associated with any Pfam suggests that many *E. muscae* genes have as yet unknown functions.

Morphological characters and sequence data suggest divergent phylogenetic relationships for *E. muscae* species complex strains

To compare the performance of morphology-based identification with DNA sequence, we mined NCBI GenBank for sequence data for members of the *E. muscae* species complex (EMSC) and outgroup taxa within the genus *Entomophthora*. Novel Sanger sequence data was also generated in support of this analysis, which has been deposited in NCBI (i.e., ARSEF_13514, ARSEF_6918, SoCal_c1 and LTE_c1, See Methods). In parallel, BLASTn searches using reference sequences generated from *E. muscae* ARSEF 13514 were used to identify closely related sequence matches for both ITS and 28S. Sequences were aligned to assess length and quality, and metadata was compiled to determine which strains had available sequence data and morphological data. Fourteen strains were chosen for inclusion based on the availability of both DNA sequence data and morphological data.

Two methods of phylogenetic inference, maximum likelihood (ML) and Bayesian inference (BI), for the 2-gene concatenated dataset resolved the EMSC as a well-supported monophyletic group that includes *E. ferdinandii*, *E. muscae* and *E. scatophagae* (Fig. 6A). In addition, *E. schizophorae* and the clade containing *E. syrphi* and *E. aff. grandis* were both monophyletic and well-supported by both phylogenetic methods. Within the EMSC, two well-supported clades were observed, referred to hereafter as EMSC Clade 1 and EMSC Clade 2 (Fig. 6A). Clade 1 contained ARSEF 13514 from *D. melanogaster* (Drosophilidae) and three other strains from Muscidae (Fig. 6B). Clade 2 had host species from three fly families (Anthomyiidae, Muscidae and Scathophagidae) and contained strains from *E. ferdinandii* and *E. scatophagae* in addition to strains identified as *E. muscae*. Strains sequenced for this study from various *Drosophila* spp. also expanded the known ranges of *E. schizophorae* and *E. grandis* (Keller, 2002).

We examined the number of nuclei in, and dimensions of, primary conidia, which are potentially diagnostic morphological characters for these species (Keller, 2002). For the six formally described fly-infecting *Entomophthora* spp. included in the study, there was considerable overlap in the ranges of size and mean number of nuclei from primary conidia (Fig. 6B, 6D). The ranges of at least four species overlap among the included *Entomophthora* species. The published sources of data used for this analysis varied in their measures (mean, maximum, minimum, standard deviation and standard error) for number of nuclei and sample sizes. Nuclei data was missing for three of fourteen strains. Number of nuclei was the most commonly reported morphological character, but this measure alone did not resolve species boundaries, except for *E. schizophorae*.

Primary conidial length and width measurements provided better species-level resolution, but these measurements were only available for four of fourteen strains (Fig. 6C). Reported spore measurement ranges for known species overlaid with those included in this study are not aligned with phylogenetic findings. Strains ARSEF 13514 and ARSEF 6918, which occurred in EMSC Clade 1 and EMSC Clade 2, respectively, overlapped with measurements for *E. ferdinandii*, but not *E. muscae*. Strains ARSEF 6716 and KVL14-17, both members of EMSC Clade 1, had overlapping spore measurements with each other and *E. muscae*, but not *E. ferdinandii*. These spore measurements further support *E. scatophagae* as part of the larger EMSC. They also highlight significant overlap between *E. grandis* and *E. syrphi*.

Discussion

A new entomophthoralean genome

We report a new ~1 Gb assembly of the *E. muscae* genome obtained using a long-read sequencing approach. Despite the comparative ease of obtaining high molecular weight DNA from other organisms, entomophthoralean fungi are recalcitrant to standard DNA extraction procedures. The reasons for this are not yet clear, but, along with the relatively large sizes of entomophthoralean genomes, they have stymied sequencing and assembly efforts within this fungal order. Our success came from having on-demand access to an *E. muscae* culture in the lab and to using a modified extraction protocol (See Methods), which could potentially be applied to other entomophthoralean fungi for future genomic work.

While our analysis suggests that this new genome is fairly complete (81.3% complete BUSCOs), we observed a high level of duplicate BUSCOs. Our assembly is decidedly not fully haploid (Fig. S1A,B). Instead, our data are consistent with duplicate BUSCOs arising from heterozygosity in coding regions. If this is the case, we have overestimated the size of the genome by about 20 Mb (2%). The prevailing thinking in the field is that entomophthoralean fungi are haploid (Humber, 2016), but the ploidy of *E. muscae* is still not fully resolved. Our observation is reminiscent of the functional diploidy previously reported for another *E. muscae* isolate (De Fine Licht et al., 2017).

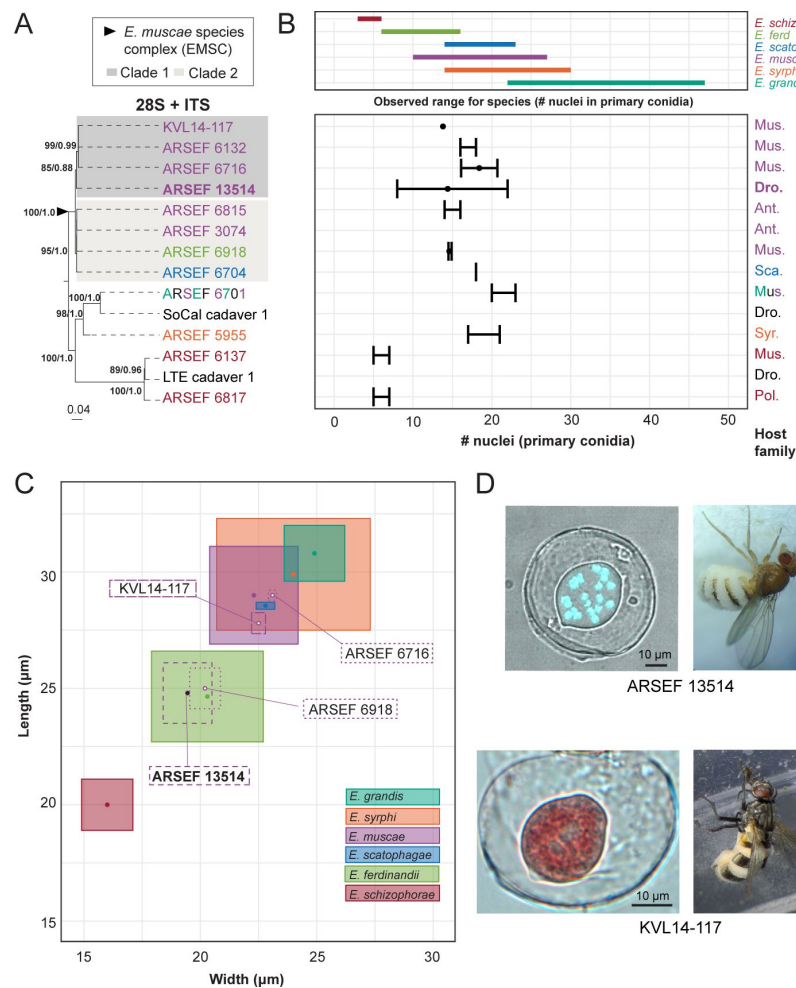


Figure 6.

Phylogenetic and morphological data for *E. muscae* ARSEF 13514, members of the *Entomophthora muscae* species complex, and closely allied fly-infesting *Entomophthora* spp.

Across all panels, species are designated by color (see key in B). A) Concatenated ITS + 28S phylogenetic tree of representative *Entomophthora* spp. including diverse strains across the EMSC. Gray boxes indicate distinct well-supported clades within the EMSC. Topology and branch lengths shown are from the ML analysis. Bootstrap support and posterior probabilities are indicated near each node (ML / BI), and only nodes with >50% support are labeled. ARSEF 6701 is denoted by three colors to indicate that this strain has multiple identifications (*E. grandis* (teal), *E. muscae* (purple) and *E. sp.* (black)). B) Nuclei number of primary conidia among strains (bottom) relative to the known ranges for each of six described species included in this study as defined by (Keller, 2002) (top). Fly family is noted at the far right for each strain (Mus. = Muscidae, Dro. = Drosophilidae, Ant. = Anthomyiidae, Sca. = Scathophagidae, Syr. = Syrphidae, and Pol. = Polleniidae). For Panel A, ITS and 28S sequence data for ARSEF 13514 (ITS & 28S), ARSEF 6918, SoCal cadaver 1 and LTE cadaver 1 are original to this study, while data for the remainder are from the literature (see Supplemental Table 3). C) Primary conidial length and width for strains with published spore measurements, overlaid atop measurements reported for each of the six species. Isolate in bold is *E. muscae* described in this study. D) Primary conidia (left) and fly hosts (right) for *E. muscae* strains ARSEF 13514 (above) and KVL14-117 (below). The *E. muscae* ARSEF 13514 primary conidium is stained with Hoechst 33342 to visualize multiple nuclei contained within conidium. The *E. muscae* KVL 14-117 conidium is stained with aceto-orcein.

A proliferation of transposable elements within Entomophthorales

Given how little is known about the biology of Entomophthorales, we adopted a comparative approach for our initial analysis of the *E. muscae* genome, sourcing available transcriptomic and genomic data from species within this fungal order. The inferred phylogenetic relationship between these species, based on conserved protein sequences (**Fig. 1C**), implied that *Conidiobolus* is paraphyletic. This is consistent with recent taxonomic revisions of that genus (Gryganskyi et al., 2022; Nie et al., 2020).

The second smallest genome (after *Conidiobolus*) in our comparative data set, was *Z. radicans*, which, at 655 Mb, is still >15-fold higher than the mean fungal genome size of all 6,144 genomes considered in **Fig. 2A** (42.7 Mb). The size of these genomes is primarily driven by a prolific expansion of repeat elements, the most numerous of which are Ty3 retrotransposons (**Fig. 2C**). Fungal Ty3 elements are common, and appear to have undergone independent expansions across many fungal lineages (Muszewska et al., 2011a). From the consistent patterns of Ty3 sequence abundance and the estimated divergence time of these sequences (**Fig. 2D**), we speculate that Ty3 expansion was a feature of the last common ancestor of non-*Conidiobolus* Entomophthorales: this proliferation was not recent.

One possible avenue leading to such a proliferation could be loss of defenses against repeat accumulation. Fungi use repeat induced polymorphism (RIP) and RNA interference (RNAi) as their two main strategies to combat the accumulation of repeated elements. We detected several homologues of the core RNAi pathway components in each of our fungal species: Dicer, Argonaute and RNA-dependent polymerase (**Fig. 2G**). This analysis suggests that RNAi pathways are intact and potentially function in a similar manner to other fungi within Entomophthorales, and may even indicate an expansion of this defense mechanism to counteract TE proliferation.

While RNAi occurs across Fungi, RIP has only been observed in Dikarya. Our analyses suggest that RIP may be active in fungi outside of Dikarya (**Fig. 2E**, 2F), but there are major caveats to this finding. We detected DNA methyltransferases with shared domain architecture as RID, a key enzyme for RIP function, but this should be validated with deeper analyses and experimental characterization of genes in these orthogroups. Host population dynamics limit opportunities for pathogen sexual recombination temporally and spatially, and entomophthoralean fungi are hypothesized to reproduce almost entirely asexually (Humber, 2016). If this is the case, these cells would not undergo meiosis, which is the cell-cycle stage where RIP occurs in fungi in the subkingdom Dikarya. Parasexuality (i.e., heterokaryons formed via arbitrary fusion of genetically distinct cells) may represent another opportunity for RIP. If parasexuality occurs in entomophthoralean fungi, a suitable moment might be in the protoplast stage, when cells lack cell walls and could meet *in insecta* after successfully infecting the host. Protoplast production is a necessary step for parasexual fusing in the laboratory setting. However, the existence of parasexuality has been questioned in this group of fungi, as cell-to-cell fusion has only been observed to occur between genetically-identical hyphal bodies prior to resting spore formation (Humber, 2016). Such fusion may still provide an opportunity to silence TEs that accumulate during primarily asexual development. However, parasexuality may also provide an opportunity for TEs to spread among heterokaryotic nuclei, as has been suggested in arbuscular mycorrhizal fungi (Yildirim et al., 2020) and may be particularly amenable for transposons with RNA intermediates (such as Ty3). In support of the presence of RIP in entomophthoralean fungi, all fungi in our set possess genes encoding DNA methylase domains (DNA_methylase: PF00145). If these genes are indeed found to mediate RIP, it would position *E. muscae* as a model for investigating meiosis-independent RIP.

Rust fungi parasitize plants (e.g., *Phakospora pachyrhizi*, *Astropuccinia psidii*; **Fig. 2A**, 2B) and also have very large genomes as a result of rampant TE proliferation. Recent analysis of three isolates of *Phakospora pachyrhizi* found that, similar to Entomophthorales, Ty3 elements comprise

43% of the genome (Gupta et al., 2023) and the expansion of these elements was predicted to coincide with radiation of host species. We do not yet know if TEs in *P. pachyrhizi* and the family Entomophthoraceae expanded via similar mechanisms. The large genomes in these clades are puzzling given that evolution tends to reduce rather than expand parasitic genomes (Wolf and Koonin, 2013). TE expansions may have been triggered by a variety of biotic and abiotic factors and underlie evolutionary choice points (Belyayev, 2014). But what factors would allow an enlarged genome to persist over evolutionary time? We hypothesize that an enlargement of the non-coding genome may set the stage for evolution of novel functionalities, and, in turn, specialization for different hosts and speciation. Rust and entomophthoralean pathogens, which are both marked by Ty3 element proliferation, are known for their extreme degree of host specialization (Sacco and Hajek, 2023). In the absence of frequent sexual recombination, the flexibility permitted by an outsized “genomic canvas” may have been favored by selection and maintained to the present day. Work addressing the basis of TE proliferation across divergent lineages is needed to better understand the evolutionary pressures driving genomic enlargement and maintenance of giant genomes. Considering the observed Ty3 proliferation in Entomophthorales is not recent (Fig. 2D), we can theorize a trajectory to the observed TE-inflation and high degree of host specialization in entomophthoralean fungi: first, a transition away from frequent sexuality to infrequent parasexuality/sexuality (lowering TE defenses), then a proliferation of TEs (generating raw materials for adaptation) and finally a slow, parasexual cessation of TE bursts with selection on genes involved in host-specific infection (strengthening the degree of host specialization).

Entomophthorales share the goal of converting insect tissue to energy; precise approaches vary

Analyses of protein domains (Pfam, CAZy and MEROPS) on our core set of entomophthoralean genomes (plus two available transcriptomes: *S. castrans* and *P. formicae*) revealed several functions enriched in the predicted *E. muscae* proteome (Fig. 3A, Fig. 3S2). Many of the associated proteins are predicted to be secreted and likely metabolize host tissues (lipases, chitinases, insulinases, etc). In addition, trehalase enzymes which act upon the primary sugar of hemolymph (Thompson, 2003), were found to be expanded in *E. muscae*, providing further evidence that this fly pathogen has evolved to efficiently process host tissue. As more entomophthoralean fungi are sequenced, the enzymatic repertoire of these highly specialized pathogens will provide insights into the nutritional needs of these fungi, which may vary with differences in tissue composition of their hosts.

Beyond these protein domain and enzyme analyses, we were very interested in comparing potential circadian machinery among entomophthoralean fungi as, for many of these pathogens, infection and behavioral manipulation of the host follows a strict daily timing (Dustan, 1924; Elya et al., 2018; Hajek and Harris, 2023; Krasnoff et al., 1995; Milner et al., 1984; Nielsen and Hajek, 2006; Pickford and Riegert, 1964). How this timing is achieved is not understood. One study suggests that timing of death of house flies killed by *E. muscae* is not dictated by host circadian machinery, instead suggesting that a fungal clock drives the timing (Krasnoff et al., 1995). In *N. crassa*, the primary fungal model system for circadian biology, two key genes in maintaining circadian rhythms are *frequency* (*frq*) and the blue-light sensitive *white-collar 1* (*wc-1*). While we found many genes encoding light-sensitive protein domains, including several *wc-1* homologs, we did not find homologs of *frq* within our entomophthoralean fungi. However, absence of *frq* does not preclude a functional clock in these fungi; many fungi that lack a *frq* homolog have clear circadian phenotypes under free-running conditions (e.g., Mucoromycotina fungus *Pilobolus sphaerosporus* and ascomycetes *Aspergillus flavus*, *Saccharomyces cerevisiae* and *Cercospora kikuchii*), indicating that there are *frq*-independent mechanisms for keeping time within Fungi that have yet to be discovered (Montenegro-Montero et al., 2015; Salichos and Rokas, 2010). In addition, the ascomycete plant pathogen *Verticillium dahliae* has a homolog of *frq* that does not cycle over circadian time (Cascant-Lopez et al., 2020). Notably, *frq* genes have

only been observed in Ascomycota (except for the subphylum Saccharomycotina), so it is unsurprising that entomophthoralean fungi would lack homologs of this gene (Montenegro-Montero et al., 2015 [↗](#); Salichos and Rokas, 2010 [↗](#)).

As a complementary approach to our domain-based analysis, we also categorized genes from our core genomic set (*E. muscae*, *Entomophaga maimaiga*, *Z. radicans* and *N. thromboides*) into putative orthologous groups (OGs; **Fig. 2E-G** [↗](#)). A plurality of OGs contained genes from all four species (40.2%). The next most common species composition of OGs was *E. muscae* and *Entomophaga maimaiga* (10.8%), followed by *E. muscae*, *Entomophaga maimaiga* and *Z. radicans* (7.5%). Each species also possessed species-specific OGs, with *E. muscae* possessing the most. An enrichment analysis of Pfam annotations for *E. muscae* genes that were assigned to core OGs revealed that many processes that may be host-specific (e.g., chitin recognition, fatty acid, protein and sugar utilization) are underrepresented within the core set (**Fig. 3** [↗](#)-S3). Conversely, genes predicted to serve in signaling pathways (e.g., kinases, pleckstrin homology domains) and basic cellular metabolism (e.g., transcription, translation and respiration) are overrepresented among *E. muscae* genes assigned to core OGs (**Fig. 3** [↗](#)-S3). All told, these results are consistent with specialization in utilizing host tissues as a key driver of species-level differences in these fungi.

We also searched for gene-family-level expansions not just within Entomophthorales (**Fig. 4A** [↗](#)), but across diverse fungal lineages that include other insect pathogenic species (**Fig. 4B** [↗](#)). We observed gene family expansions related to host tissue utilization, some unique to Entomophthorales (e.g., lipases, acid trehalases) and others shared across entomopathogens of different phyla (e.g., peptidases). This partial overlap indicates that there are common themes among insect fungal entomopathogens in conquering their hosts as well as divergent strategies within distinct fungal lineages.

Entomophthoralean fungi likely produce secondary metabolites

Our analysis of predicted secondary metabolites within insect fungal entomopathogens suggests that Entomophthorales may produce non-ribosomal peptides, siderophores and terpenes. A recent report demonstrating the presence of several terpenoid compounds in *E. muscae*-killed house flies (Naundrup et al., 2022 [↗](#)) suggests that this prediction likely missed important classes of metabolites. Current models predict that the quantity and diversity of secondary metabolites is lower in entomophthoraleans than for ascomycete entomopathogens (**Fig. 4C** [↗](#)), but this conclusion has several caveats. At a first glance, our predictions seem consistent with Entomophthorales following a primary growth, rather than a primary toxin-producing, infection strategy (as proposed for certain generalist *Metarhizium* species (Kershaw et al., 1999 [↗](#))), and with the long-held dogma that zygomycete fungi produce few secondary metabolites (Bushley and Turgeon, 2010 [↗](#); Voigt et al., 2016 [↗](#)).

However, secondary metabolite cluster analysis methods have been developed almost exclusively using data from Ascomycota and Basidiomycota. Thus, the lack of predicted secondary metabolites in Entomophthorales could be due to prediction methods not well suited to function outside of Dikarya. Recent evidence for secondary metabolite gene clusters in Mucoromycota opposes the long-held view that zygomycete fungi do not produce these compounds (Koczyk et al., 2021 [↗](#); Voigt et al., 2016 [↗](#)) and suggests that methods to detect metabolites and the enzymes that synthesize them need to be adjusted for non-Dikarya fungi. Indeed, a recent study in *Massospora* experimentally detected several secondary metabolites that, based on current bioinformatic models of biosynthesis, should not exist in this species (Boyce et al., 2019 [↗](#)). In addition, because Zoopagomycota are so uncharacterized with respect to secondary metabolites, it is possible that the apparent lack of standard metabolite classes is accurate, but we failed to identify metabolite clusters that are unique to this phylum and not yet described. We have much to learn about secondary metabolism beyond Dikarya and future efforts should prioritize exploring metabolism of these taxa for the potential discovery of novel metabolic gene cluster architecture or novel metabolite synthesis pathways.

Potentially unique features of *E. muscae* biology

From analysis of Pfam, MEROPS and CAZy domains, we found that *E. muscae* has a handful of protein domains unique within the Entomophthorales species we analyzed. The most numerous of these was Peptidase A2B, which is a protease family involved in processing retrotransposon Ty3 polyprotein into its component parts (Kirchner and Sandmeyer, 1993 [↗](#)). Perhaps this reflects a species-specific transposon for *E. muscae*. However, given the proliferation of Ty3 elements among family Entomophthoraceae fungi in our dataset, we expect that the functional activity of this domain is not restricted to Peptidase A2B. Importantly, we did not find any domains that were missing from *E. muscae* but enriched in other fungi in our core set (Fig. 5B [↗](#), Fig. S5C). Together, these observations are consistent with *E. muscae* having diverged from other entomophthoralean fungi by expansion of existing domain families rather than loss of particular domains.

As an additional approach to identifying unique gene functions in *E. muscae*, we also looked at orthogroups unique to this species, from either the total proteome or the secreted proteome (Fig. 5C, D [↗](#)). M16 peptidases (domains PF00675 and PF05193) were found to be enriched across all proteins as well as within proteins predicted to be secreted. M16 peptidases are metalloendopeptidases with family members that cleave N-terminal targeting peptides that direct proteins to their destination subcellular compartments (Rawlings et al., 2018 [↗](#)). Recent work in the protozoan *Cryptosporidium parvum* implicated insulinase-like proteases (which are also members of the M16 family) in early infection (Zhang et al., 2019 [↗](#)), perhaps suggesting a role for these proteins in early development of *E. muscae*.

Based on our domain enrichment analysis, we expect that many of the unique secreted proteins are involved in the metabolism of fly host tissues. Depending on the host range specificity of this study's *E. muscae* isolate, these genes could encode proteins for metabolizing or recognizing *Drosophila* macromolecules specifically. Looking at just secreted proteins, we saw enrichment for a handful of catabolic domains including peptidase inhibitor I9 (PF05922), a constituent of subtilisins (Arnesen et al., 2018 [↗](#); Muszewska et al., 2011b [↗](#)). This observation is consistent with work finding unique subtilisin-like serine proteases in the entomophthoralean fungi *E. muscae*, *C. incongruus* and *P. formicae* and suggests further subspecialization of these enzymes at the species level (Arnesen et al., 2018 [↗](#)). We also observed an enrichment for chitin recognition protein (PF00187) in the secreted proteome. Both the fly host and *E. muscae* produce chitin, so these proteins could function in either host recognition or in masking *E. muscae* from being recognized by the host immune system upon entry (Cen et al., 2017 [↗](#)).

An enrichment for the collagen triple helix repeat domain (PF01391) caught our attention. In the generalist fungal pathogen *Metarhizium anisopliae*, a collagen-like protein (Mcl1) that is expressed immediately after invasion of the host hemolymph helps fungal cells avoid detection by the host immune system (Wang and St Leger, 2006 [↗](#)). In a similar vein, we wonder if *E. muscae*'s unique collagen triple helix domain may play a role in evading the host immune system.

We also were surprised to observe an enrichment in berberine-like proteins (PF08031) among secreted *E. muscae*-specific OGs. Berberine-like proteins are involved in the synthesis of isoquinoline alkaloids, a class of compounds that includes the analgesics morphine and codeine, and usually require FAD as a cofactor (Kutchan and Dittrich, 1995 [↗](#)). Intriguingly, we also observed an enrichment of FAD binding domains (PF01565). Isoquinolines have been most frequently discovered in plants, but recent work in *Aspergillus fumigatus* discovered that an orphan metabolic gene cluster containing small NRPS-like genes encodes enzymes capable of producing a novel isoquinoline (Baccile et al., 2016 [↗](#)). Isoquinolines have wide-ranging effects (Khan and Suresh Kumar, 2015 [↗](#)), and some isoquinolines have been shown to inhibit the innate immune response. Future work leveraging high-sensitivity metabolomics methods will be needed to determine if isoquinolines are produced by *E. muscae* and, if so, what role they may play in infection and in altering host behavior.

Recommendations for *Entomophthora* species identification

Historically, *Entomophthora* spp. were identified using a combination of morphology, host species, and/or location (MacLeod et al., 1976 [↗](#); MacLeod and Müller-Kögler, 1973 [↗](#)). However, in recent years it has come to light that these features may be unreliable. For example, conidial size of members of the EMSC changes when infecting host flies different from the original source host (Jensen et al., 2006 [↗](#)), meaning that this feature is not, on its own, reliable for identifying the fungal species. Such differences in conidial size across insect hosts have also been observed for another closely allied member of the Entomophthoraceae, *Massospora levispora* (Macias et al., 2020 [↗](#)). Another study examining haplotype diversity in *E. muscae* strains from a single epizootic infecting two fly species in NC, USA uncovered two genetically distinct subpopulations, both with nearly identical primary conidial sizes and numbers of nuclei per conidium (Gryganskyi et al., 2013 [↗](#)).

To explore the utility of DNA sequence data (ITS + 28S) for identification purposes in the EMSC, we assessed the congruence of available sequence and morphology data for EMSC isolates. We recovered two well-supported clades (both in the single gene trees and the concatenated tree) that contain strains previously identified as *E. muscae*, *E. ferdinandii*, and *E. scatophagae* (Fig. 6A-C [↗](#)). Using morphological characters, we found that the number of nuclei in the primary conidia contained almost no phylogenetic signal among members of the EMSC based on the strains we assessed (Fig. 6B [↗](#)). Furthermore, the dimensions of the primary conidia were largely overlapping among nearly all *Entomophthora* species examined, with the exception of *E. ferdinandii* and *E. schizophorae* (Fig. 6C [↗](#)). Interestingly, dimensions of primary conidia for our *E. muscae* strain (ARSEF 13514) overlapped with measurements for *E. ferdinandii* and outside the reported measurements for *E. muscae* despite occupying the same well-supported clade as strains whose spore measurements agreed with *E. muscae*.

Taken together, these results indicate that the morphological characters examined here, which are traditionally used to identify Entomophthorales, are incongruent with molecular data especially within the EMSC. However, this disagreement between morphology and sequence data could reflect 1) the paucity of character data, 2) insufficient taxon sampling that failed to capture the full range of morphological and genetic variation, and/or 3) the presence of misidentified specimens in collections and studies that are potentially masking the presence of morphologically cryptic species (i.e., if researchers relied on nuclei number and conidia measurements to identify their specimens in the absence of DNA sequence data, then any morphology data they report for the species could include data from multiple species). While ITS and 28S loci may have insufficient phylogenetic signal to fully resolve these species boundaries especially within the EMSC, resolution may be achievable with the addition of other loci. In addition, the evolutionary history of this group may include a combination of introgression and rare sexual recombination events that have softened species boundaries within the EMSC, though data supporting this are currently limited (Gryganskyi et al., 2013 [↗](#)). Overall, the EMSC remains monophyletic, but a comprehensive taxonomic revision is needed to define species within the clade.

Many strains of EMSC exist in fungal collections (ARSEF in particular) that have no associated sequence or morphology data. We propose that for each individual specimen, whether novel or from a collection, researchers minimally generate ITS and 28S sequence data and collect morphological data (specifically the length, width, and number of nuclei within primary conidia), so that future research can assess these features in combination to define the species boundaries of this group. We encourage researchers to further deposit voucher specimens, live cultures and/or DNA with fungal collections (e.g., ARSEF) so that future researchers can return to the original sample to examine traits or generate more sequence data if needed. Likewise, historic studies based solely on morphology from herbarium specimens may need to be resampled for both DNA sequencing and morphology.

What's next?

The work presented here is one of the first steps in exploring Entomophthoralean genomes. We hope that the data and questions posed herein can serve as a springboard for generating and testing new hypotheses about these organisms. *Entomophthora muscae* has been under scientific scrutiny since its description in 1855, yet, even under the new light of our genomic investigation, this fungus remains “one of the strangest and most interesting apparitions” (Elya and De Fine Licht, 2021) with clear avenues to enrich our understanding of myriad biological phenomena.

Materials & Methods

E. muscae culture and DNA extraction

E. muscae was isolated for *in vitro* growth from a single sporulating cadaver in 60 mm sterile petri dishes using the ascending sporulation method (Hajek et al., 2012). All cultures were grown in supplemented Grace's Insect Media (ThermoFisher Scientific #11605094) with 5% fetal bovine serum added (ThermoFisher Scientific #10437010) in volumes of 20 mL in 25 cm² vented tissue culture flasks (Corning #353014). Cultures were grown without shaking, with flasks laid on the long edge to maximize surface oxygen exchange, in the dark at room temperature (19-21°C). DNA was isolated from a log-phase culture via methods of (Elya and Lee, 2022). DNA quantity, quality and size distribution were assessed by Qubit dsDNA HS Assay kit (ThermoFisher Scientific #Q32851), Nanodrop spectrophotometric analysis (ThermoFisher Scientific #ND2000) and DNA Genomic ScreenTape analysis (Agilent TapeStation #5067-5365 and #5067-5366) following manufacturer's protocols.

Genomic sequencing and assembly

3 µg of genomic DNA were used to prepare two Oxford Nanopore DNA libraries, one sheared to 20 kb and one unsheared (45-50 kb) and sequenced sequentially on the Oxford Nanopore PromethION platform by the The Bauer Core Facility at Harvard University (Cambridge, MA) to 82x coverage. Reads are accessible through the NCBI Sequence Read Archive (SRA) via accession number SRR18312934. Nanopore reads that passed filtering were assembled using Flye v2.8.3 (Kolmogorov, 2021; Kolmogorov et al., 2019). The resultant assembly was polished with the same Nanopore reads used for assembly using Medaka v1.2.6 (Oxford Nanopore Technologies Ltd, 2021).

The Flye+Medaka assembly was further scaffolded with 10x genomic sequence library reads from the UCB isolate of *E. muscae* (Elya et al., 2018) deposited in NCBI SRA via accession number [SRR18312935](#). This was accomplished by demultiplexing with longranger v2.2.2 (10x genomics), reads mapped with bwa-mem2 (Vasimuddin et al., 2019), followed by assembly scaffolding with Tigmint v1.2.2 (Jackman et al., 2018) and ARCS 1.2.1 (Yeo et al., 2018).

Genome annotation

This updated *E. muscae* UCB genome was annotated with Funannotate v1.8.9 (Palmer and Stajich, 2020) employing the default parameters, which automates processing of RNAseq to produce transcripts, gene prediction with evidence from the transcripts and alignments of proteins, and construct consensus gene models from multiple gene prediction evidence, refine gene models for alternative splicing prediction based on the RNA-seq, and produce product function annotation based on protein sequence homology to databases of domains and annotated proteins. To annotate the genome, first repetitive sequences were masked by building a species-specific repeat library with RepeatModeler v2.0.1 (Flynn et al., 2020). These putative repeat families were further screened manually to remove likely protein-coding genes based on Diamond BLASTX v2.0.8

(Buchfink et al., 2021 [↗](#)) searched against SwissprotDB v2021_04 (The UniProt Consortium et al., 2020) and removing sequences that matched clear non-repetitive but multicopy gene families. The curated repeat library was combined with the RepBase (Bao et al., 2015 [↗](#)) library of fungi repeats with RepeatMasker v4-1-1 (Smit et al., 2013-2022 [↗](#)) to softmask the genome before gene prediction to avoid over predicting transposons as host genes.

To train gene predictors and support gene models, RNA-seq data from previously deposited sequence data in NCBI SRA accession ERR1022665 were used as informant data in the annotation process. Briefly, the RNA-seq processing used Trinity v2.11.0 (Haas et al., 2013 [↗](#)) in Genome Guided mode which aligned reads to the genome with Hisat2 v2.2.1 (Kim et al., 2019 [↗](#)) followed by targeted transcript assembly with the Trinity pipeline on reads clustered to distinct genomic locations. The constructed assemblies total 63,231 which reflect partial transcripts and potential alternative splicing. These sequence assemblies were aligned to the genome with PASA (Haas et al., 2008 [↗](#)) which uses the splice-aware aligner GMAP (Wu et al., 2016 [↗](#)) and additional software to produce gene models from transcript data. This produced 67,902 gene models. The transcripts with full-length Open Reading Frames as scored by Transdecoder v5.5.0 (Haas et al., 2013 [↗](#)) were kept as high quality models for gene prediction training. A total of 2,653 full-length PASA-derived gene models were used as a training set to *ab initio* predictors SNAP v2013_11_29 (Korf, 2004 [↗](#)) and AUGUSTUS v3.3.3 (Stanke et al., 2008 [↗](#)). In addition, the tools GeneMark-ES v4.62 (Ter-Hovhannisyan et al., 2008 [↗](#)) and GlimmerHMM v3.0.4 (Majoros et al., 2004 [↗](#)) were run after these tools performed self-training. Gene models were also predicted by CodingQuarry (Testa et al., 2015 [↗](#)) using the transcript alignments as exon hints. Exon evidence was also generated by DIAMOND alignment of SwissprotDB proteins and polished by Exonerate v2.4.0 (Slater and Birney, 2005 [↗](#)). The transcript and protein-based hints were provided to GeneMark, SNAP, and AUGUSTUS for evidence-guided gene prediction. EvidenceModeler v1.1.1 (Haas et al., 2008 [↗](#)) generated consensus gene models in Funannotate using its default evidence weights. tRNA genes were predicted by tRNAscan-SE v1.3.1 (Lowe and Eddy, 1997 [↗](#)). Putative protein functions were assigned to genes based on sequence similarity to the InterProScan v5.51-85.0 (Blum et al., 2021 [↗](#); Jones et al., 2014 [↗](#)), Pfam v35.0 (Mistry et al., 2021 [↗](#)), EggNOG v2.1.6-d35afda (Cantalapiedra et al., 2021 [↗](#)), dbCAN2 v9.0 (Huang et al., 2018 [↗](#)), and MEROPS v12.0 (Rawlings et al., 2018 [↗](#)) databases relying on NCBI BLAST v2.9.0+ (Camacho et al., 2009 [↗](#)) and HMMer v3.3.2 (Eddy, 2011 [↗](#)). Predicted secreted genes and transmembrane domains were annotated with Phobius (Käll et al., 2004 [↗](#)) and SignalP v5.0b (Almagro Armenteros et al., 2019). A total of 39,711 gene models comprising 42,665 predicted proteins with alternative splicing isoforms, and 793 tRNAs were predicted. Pipeline for annotation is archived in the github repository https://github.com/zygolife/Entomophthora_muscae_UCB [↗](#) and Zenodo archive (Stajich et al., 2023a [↗](#)).

Fungal datasets

A comprehensive overview of all of the fungal datasets used in this paper and their associated figures is provided in Supplemental Table 2. Transcriptomic data for *Strongwellsea castrans sensu lato* were generated from a cabbage fly (*Delia radicum*) caught in a cabbage field (farm name: Søriskgård, Latitude: 55.823706, Longitude 12.171149, date: 04/09/2013). The fly contained a clearly visible large hole on the side of the abdomen characteristic of *S. castrans* infection and was kept alive for a few hours until being snap-frozen in liquid nitrogen. The infected fly was ground to a fine powder in liquid nitrogen before extracting total RNA using a Plant RNEasy Kit (QIAGEN #74904) following manufacturer's instructions. Total RNA was sequenced using Illumina HiSeq 2000 technology and TRUseq library building by Beijing Genomics Institute (BGI-Europe, Copenhagen, Denmark), and transcriptome sequences were assembled in *de novo* mode using Trinity v2.11.0. Raw reads are deposited at the European Nucleotide Archive (ENA) with accession number: ERR12023556. Assembled transcripts are available in zenodo archive under Supporting_data folder (Stajich et al., 2023b [↗](#)).

Phylogeny of species

Phylogenetic relationships of species were determined by identifying conserved proteins from the protein translation of predicted genes or transdecoder ORFs from the transcriptome-only sampled species (*P. formicae* and *S. castrans*) using the PHYling v1.1 (Stajich, 2023) BUSCO / OrthoDB fungi_odb10 marker set. Briefly, the pipeline searches for conserved, generally single copy proteins via HMMer (Eddy, 2011) searches and builds individual and concatenated protein alignments. The phylogenetic tree was constructed from the concatenated alignments with FastTree v2.1.11 (Price et al., 2009) using the ‘-lg -gamma’ parameters. Comparison of this topology matched previous published relationships of these lineages (eg (Boyce et al., 2019; Wang et al., 2023)) which did not justify further exploration with additional likelihood or Bayesian methods.

BUSCO analysis

Genome completeness was computed with BUSCO v5.2.2 (Manni et al., 2021) using the -m genome mode on the masked assembly employing either fungi_odb10 or eukaryota_odb10 marker sets. The analysis was also performed on the predicted proteins using the -m protein mode.

Genome size and gene counts

Data used to generate Fig. 2A and 2B were compiled from NCBI, MycoCosm, the Fungal Genome Size Database (zbi.ee/fungal-genomesize/) and (Mohanta and Bae, 2015) and are available as Supplementary File S1. In some cases, multiple strains of the same species have been sequenced.

Repeat analysis

To identify repetitive sequences in the entomophthoralean genomes we used RepeatModeler v2.0.1 (Flynn et al., 2020) to develop a *de novo* library of elements and classify their likely lineage using default parameters and LTR finding with LTRStruct option. The elements were classified by similarity to RepBase v20170127 (Bao et al., 2015) using the RepeatClassifier component of RepeatModeler. The *de novo* repeat library was combined with all identified fungi repeats in RepBase to produce a composite library to identify repetitive regions in the genomes with RepeatMasker v4.1.1 (Smit et al., 2013-2022) genome masking and transposon exploration. The repeat landscape plots were generated with the RepeatMasker script createRepeatLandscape.pl using default options.

To test for the presence of Repeat Induced Point Mutation patterns in the genome, a RIP index was calculated for 1kb windows with 500 bp offsets using the composite index (Lewis et al., 2009). Summary statistics were calculated to score the number of windows where the composite RIP index was greater than 0 to summarize the % of the genome RIPped. Scripts and summarized values from each genome are part of the https://github.com/zygolife/Emuscae_Comparative repository and archived in zenodo (Stajich et al., 2023b) in the ‘comparative/RIP’ folder.

Domain and comparative analysis

Domain analysis for seven entomophthoralean proteomes (*E. muscae*, *Entomophaga maimaiga*, *Z. radicans*, *P. formicae*, *S. castrans*, *N. thromboides*, and *C. coronatus*) was limited to MEROPS, CAZy and Pfam annotations. These annotations were completed according to a pipeline identifying domains by sequence similarity and motif searches https://github.com/stajichlab/Comparative_pipeline. The identified domain counts per species aggregated as total counts or total number of unique genes with a domain were compiled into a single table for comparison between species.

In RStudio v1.4.1717 (RStudio Team, 2022 [↗](#)), a custom R v4.1.0 (R Core Team, 2021 [↗](#)) script for enrichment analysis is available in https://github.com/zygolife/Emuscae_Comparative [↗](#) repository and archived (Stajich et al., 2023b [↗](#)), which relies on a number of R packages: tidyverse v1.3.1 (Wickham et al., 2019 [↗](#)), fmsb v0.7.2 (Nakazawa, 2022 [↗](#)), grid (an R base package), gridExtra v2.3 (Auguie, 2017 [↗](#)), ComplexUpset v1.3.3 (Krassowski et al., 2022 [↗](#)), UpSetR v1.4.0, ggforestplot v0.1.0 (Scheinin et al., 2020 [↗](#)), broom v0.7.12 (Robinson et al., 2023 [↗](#)) and viridis v0.6.2 (Garnier et al., 2021 [↗](#)). For each domain type, the number of accessions containing each domain was counted. Domains that were present in at least two genomes were used for pairwise enrichment analysis among genomes. For each domain type, enrichment of counts was calculated using `fmsb::pairwise.fisher.enrichment` with Bonferroni correction and a p-value threshold of 0.01 for significance. For each genome, the number of genes found to be significantly different were counted, and the median count for each domain was used to estimate the fold compared to median and a direction (i.e., up or down) compared to median.

Count tables were used to determine lists of domains present within each genome for set analysis including UpSet plots (ComplexUpset and UpSetR) and analysis of unique/missing genes. A second set analysis was conducted with Pfam domains including an additional *E. muscae* proteome from RNAseq data (EMU-T) generated from the same strain (NCBI GEO GSE111046).

Putative circadian genes were surveyed using Pfam domain annotations. The following domains were used for this survey: 7tm_1 (PF00001), FRQ (PF09421), GATA (PF00320), PAS_3 (PF08447), PAS_9 (PF13426), PAS (PF00989), Bac_rhodopsin (PF01036), GpcrRhopsn4 (PF10192), DNA_photolyase (PF00875), FAD_binding_7 (PF03441), PHY (PF00360) and GAF (PF01590). These were used to filter single-domain circadian candidates from each genome for our domain analysis. Additionally, we considered the expected domain pattern of known circadian genes to identify curated candidates for FRQ (one FRQ domain), WC1 (one GATA and one PAS_3 and one PAS_9), WC2 (one GATA and one PAS_3 and no PAS_9) and 7tm_1 (rhodopsin; one 7tm_1 domain).

RNAi pathway candidates were similarly surveyed using Pfam domain annotations. Domains included in this survey were: RdRP (PF05183), Dicer_dimer (PF03368), PAZ (PF02170), Piwi (PF02171), Ribonuclease_3 (PF00636) and DEAD (PF00270). Identified candidates containing surveyed domains, but we further employed an expected domain pattern to curate candidates for RNAi pathway proteins: RDRP (containing RdRP domain), Dicer (containing Dicer_dimer), Ago (containing both PAZ and Piwi), Dicer_Alt (containing Ribonuclease_3 and either DEAD or PAZ).

Orthogroup analysis

Orthologous genes were identified between the compared entomophthoralean species by first taking the longest peptide for each genee are part of the to avoid including alternative spliced isoforms in the analyses. Proteins were clustered with OrthoFinder v2.5.2 (Emms and Kelly, 2015 [↗](#)) using Diamond v2.0.6.144 (Buchfink et al., 2021 [↗](#)) with the ultra-sensitive parameter.

Gene expression was estimated by aligning a pooled set of 27 *in vivo* samples (whole female fruit flies that had been exposed to *E. muscae* 24 to 120 hours previous from NCBI #GSE111046 (Supplemental Table 4) to the annotated assembly using Kallisto (v.0.46.1). Genes with fewer than 5 estimated counts across this pooled set were filtered out of the potentially species-specific gene set (a collection of genes that either failed to cluster in any orthogroup across *E. muscae*, *Entomophaga maimaiga*, *Z. radicans* and *N. thromboides* or appeared in an orthogroup that was exclusively populated by *E. muscae* genes).

Orthogroups were used to identify methyltransferases that share RID domain architecture. RID is annotated with multiple Pfam DNA_methylase (PF00145) domains. RID candidate proteins, containing more than one DNA_methylase, were identified with the same approach we used to identify circadian protein candidates. Where necessary, homologs that corresponded with genomes used in orthogroup analysis were identified via BLAST. All proteins within orthogroups

that contained RID candidates (OG0001715 and OG0003300) were grouped in Geneious Prime (2021.2.2) and aligned using MAFFT (7.450; BLOSUM62) with the sequence of RID (AAM27408.1) from *N. crassa*. This alignment was used to produce a protein tree using RAXML (8.2.11; GAMMA BLOSUM62 with 1000 bootstraps). This tree was manually rooted to separate orthogroups, maintaining the clade containing RID and OG0001715. The basal *E. muscae* protein DSO57_1016266-T1 is apparently partial (140 aa) compared to other *E. muscae* proteins in OG0001715 (both are >1200 aa).

Gene family expansion analysis

Protein sequences containing extracellular trehalase enzymes (PF01204) were retrieved from the data and aligned using MAFFT, which was used to calculate a maximum likelihood phylogenetic tree with RAXML using default parameters. In addition, total number of genes and number of genes encoding the PFAM domains Lipases (Lipase_3), Subtilisin-like serine peptidases (Peptidase_S8), Trehalases (Trehalase), Trypsins (Trypsin), and Chitinases (Glycohydro_18) in representative fungal species of Zoopagomycota and Ascomycota were retrieved from JGI Mycocosm and compared to our data of *E. maimaiga* (EMA), *Z. radicans* (ZRA), *N. thromboides* (CTH), *C. coronatus* (Conco1), and *Basidiobolus meristosporus* (Basme2finSC).

Antismash secondary metabolite prediction

To examine secondary metabolism potential for the *E. muscae* genome and its relatives, fungiSMASH v6.0.0 of antiSMASH (Blin et al., 2021 [DOI](#)) was run as part of the funannotate genome annotation steps with the options “--taxon fungi --genefinding-tool none --fullhmmer --clusterhmmer --cb-general --cassis --asf --cb-subclusters --cb-knownclusters.” The predicted clusters were incorporated into the genome annotation submitted to GenBank.

Mining and analysis of ribosomal RNA sequences and morphological data for *Entomophthora muscae* species complex isolates

DNA sequence data for members of the *E. muscae* species complex and other *Entomophthora* spp. were mined from NCBI GenBank. Novel Sanger sequence data were also generated in support of this analysis, which has been deposited in NCBI (i.e., ARSEF_13514, ARSEF_6918, SoCal_c1 and LTE_c1). DNA was extracted from individual EMSC-killed cadaver drosophilids collected across California as per (Elya et al., 2018 [DOI](#)) using a QIAamp DNA Micro Kit (QIAGEN #56304). Extracted DNA was used to template PCR reactions with GoTaq polymerase (Promega #M3001) to amplify ITS (primers emITS-1:TGGTAGAGAATGATGGCTGTTG and emITS-4:GCCTCTATGCCTAATTGCCTTT) and/or LSU regions (primers LR0R-4:GTACCGCTGAACCTAAGC and LR3-1:GGTCCGTGTTTCAAGAC) (James et al., 2006 [DOI](#)). PCR reactions were enzymatically cleaned with ExoSAP-IT (ThermoFisher #78201.1.ML) per manufacturer’s instructions and submitted to ELIM (Hayward, CA) for Sanger sequencing with both forward and reverse primers. Consensus sequences were assembled in UGENE (Unipro). Sequences were aligned to assess length and quality, and metadata was compiled to determine which strains had available sequence data and morphological data (Supplemental Table 3). Morphological data were collected by cross-referencing *Entomophthora* strain IDs with published peer-reviewed literature as well as unpublished/raw data provided by two of this paper’s co-authors. Fourteen strains were chosen for inclusion based on the availability of both DNA sequence data and morphological data.

E. muscae phylogenetic tree

ITS and 28S sequences were aligned separately using MAFFT (Katoh and Standley, 2013 [DOI](#)) on the Guidance2 server (<http://guidance.tau.ac.il/>; (Landan and Graur, 2007 [DOI](#); Sela et al., 2015 [DOI](#))), and individual residues with Guidance scores <0.5 were masked (1.6% of all residues for ITS, 0.1% for 28S). Overall Guidance scores for each locus were 0.955 and 0.995 for ITS and 28S, respectively (alignments with scores approaching 1.0 have high confidence). Nucleotide substitution models

were chosen using corrected Akaike information criterion (AICc) scores in Model Test in MEGA X 10.2.6 (Kumar et al., 2018 [↗](#); Stecher et al., 2020 [↗](#)). Alignments of each individual locus, and a concatenated alignment of the two, were used in a maximum likelihood (ML) analysis (RAxML 8.2.12; (Stamatakis, 2014 [↗](#))) and a Bayesian inference (BI) analysis (MrBayes 3.2.5; (Ronquist et al., 2012 [↗](#))), for a total of 6 analyses. In brief, for ML analyses, an appropriate model was chosen, partitions were applied (for each locus in the concatenated analysis only), 1,000 bootstrap replicates were used, and the best-scoring tree was identified and bootstrapped in a single run. For BI analyses, MrBayes was allowed to select a substitution model for each data set, and rates were set based on results from Model Test. One cold chain and three heated chains were used for each of 2 runs, and the first 25% of generations were discarded as burn-in. Each analysis was set for 1 million generations, and no additional generations were needed because the standard deviation of split frequencies fell below 0.01. Finally, run parameters were checked for convergence in Tracer 1.7.1 (Rambaut et al., 2018 [↗](#)). Trees were viewed and prepared for publication using FigTree 1.4.4 (Rambaut, 2017 [↗](#)) and Inkscape 0.92.2 (<https://www.inkscape.org/> [↗](#)). All resulting trees & alignments are available in zenodo archive (Stajich et al., 2023b [↗](#)).

Supplementary materials

Supplemental Table 2. Summary of fungal isolates and data used.

Supplemental Table 3. Information about strains used in phylogenetic and morphologic studies (related to **Fig. 6** [↗](#)).

File S1. Genome sizes and gene counts across Fungi (related to **Fig. 2** [↗](#)).

Zenodo: *E. muscae* genome annotation repository [↗](#)

Zenodo: *E.muscae* comparative repository [↗](#)

Funding

JES was supported by the National Science Foundation (NSF) (DEB-1441715 & EF-2125066) and the U.S. Department of Agriculture (USDA) (National Institute of Food and Agriculture Hatch projects CA-R-PPA-211-5062-H). Analyses were performed on the UC Riverside High Performance Computing Cluster supported by the NSF (DBI-1429826 & DBI-2215705) and the National Institutes of Health (NIH) (S10-OD016290). AMM is supported by an WVU Outstanding Merit Fellowship. BL was supported by the USDA (USDA-ARS Project 8062-22410-007-000D). BdB is supported by the Alfred P. Sloan Foundation (Research Fellowship), the Ester A. and Joseph Klingenstein Fund (Klingenstein-Simons Fellowship Award), the Richard and Susan Smith Family Foundation (Odyssey Award), a Harvard/MIT Basic Neuroscience Grant, the NIH/NINDS (1R01NS121874-01), and the NSF (IOS-1557913). AEJ is supported by the USDA Forest Service (15-CA-11420004-095). HHDFL is supported by a Sapere Aude: DFF-Starting Grant (8049-00086B) from the Independent Research Fund Denmark and a Carlsberg Foundation Young Researcher Fellowship (CF20-0609). CNE is supported by Howard Hughes Medical Institute (Hanna H. Gray Postdoctoral Fellowship GT11087).

Species	Gene models - original pipeline	Gene models - funannotate pipeline
<i>Entomophaga maimaiga</i>	14,701*	23,807
<i>Z. radicans</i>	14,479**	18,761
<i>M. cicadina</i>	N/A	7,532

*FilteredModels3, per <https://mycocosm.jgi.doe.gov/Entmai1/Entmai1.info.html>

**FilteredModels2, per <https://mycocosm.jgi.doe.gov/Zoorad1/Zoorad1.info.html>

Supplemental Table 1.

Variance in predicted gene models using different annotation pipelines does not explain large gene count predicted in *E. muscae* genome.

Accession	Time_point	Tissue	Treatment
SRR6766803	72 hours	Whole fly	Exposed
SRR6766804	72 hours	Whole fly	Exposed
SRR6766805	72 hours	Whole fly	Exposed
SRR6766806	24 hours	Whole fly	Exposed
SRR6766807	24 hours	Whole fly	Exposed
SRR6766808	24 hours	Whole fly	Exposed
SRR6766809	48 hours	Whole fly	Exposed
SRR6766810	48 hours	Whole fly	Exposed
SRR6766811	48 hours	Whole fly	Exposed
SRR6766812	96 hours	Whole fly	Exposed
SRR6766813	96 hours	Whole fly	Exposed
SRR6766814	96 hours	Whole fly	Exposed
SRR6766815	48 hours	Whole fly	Exposed
SRR6766816	48 hours	Whole fly	Exposed
SRR6766817	48 hours	Whole fly	Exposed
SRR6766818	24 hours	Whole fly	Exposed
SRR6766819	24 hours	Whole fly	Exposed
SRR6766820	72 hours	Whole fly	Exposed
SRR6766821	24 hours	Whole fly	Exposed
SRR6766822	72 hours	Whole fly	Exposed
SRR6766823	72 hours	Whole fly	Exposed
SRR6766824	96 hours	Whole fly	Infected cadaver
SRR6766825	96 hours	Whole fly	Infected cadaver
SRR6766826	96 hours	Whole fly	Infected cadaver
SRR6766827	120 hours	Whole fly	Infected cadaver
SRR6766828	120 hours	Whole fly	Infected cadaver
SRR6766829	120 hours	Whole fly	Infected cadaver

Supplemental Table 4.

SRA accession numbers of *E. muscae* RNAseq data (NCBI GSE111046) used for pooled expression analysis (related to Fig. 3 [↗](#)).

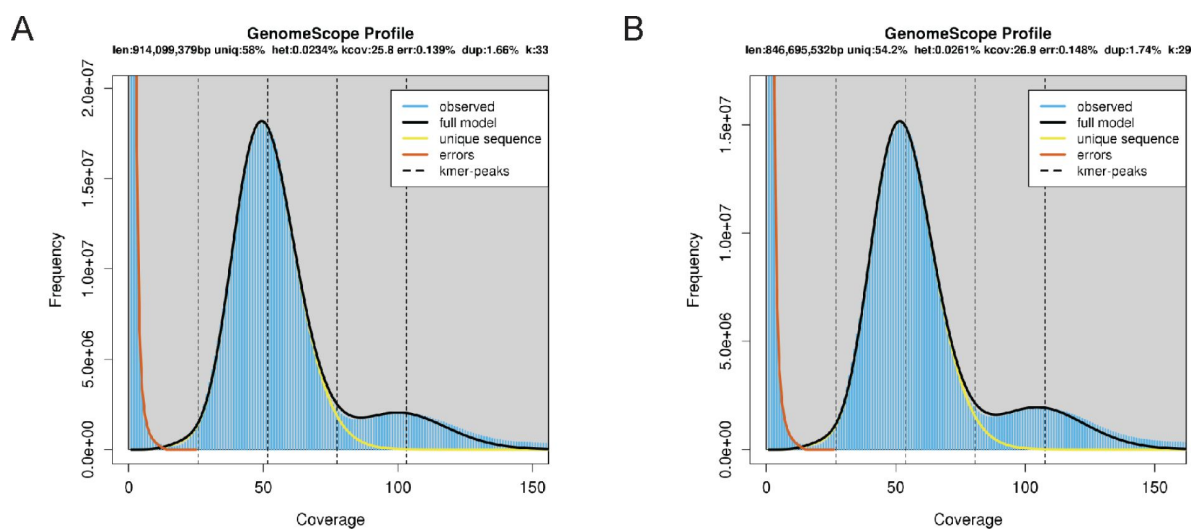


Figure. 1-S1.

Kmer distributions within *E. muscae* genome assembly. A) 33mers; B) 29mers. Jellyfish was used to count kmers and plots were generated with GenomeScope. "Len" indicates estimated assembly size (in bp); "uniq": % of unique kmers observed; "kcov": estimated coverage of assembly; "err": % error kmers; "dup": % duplicated kmers; "het": indicates % of heterozygous bases; "k": kmer size.

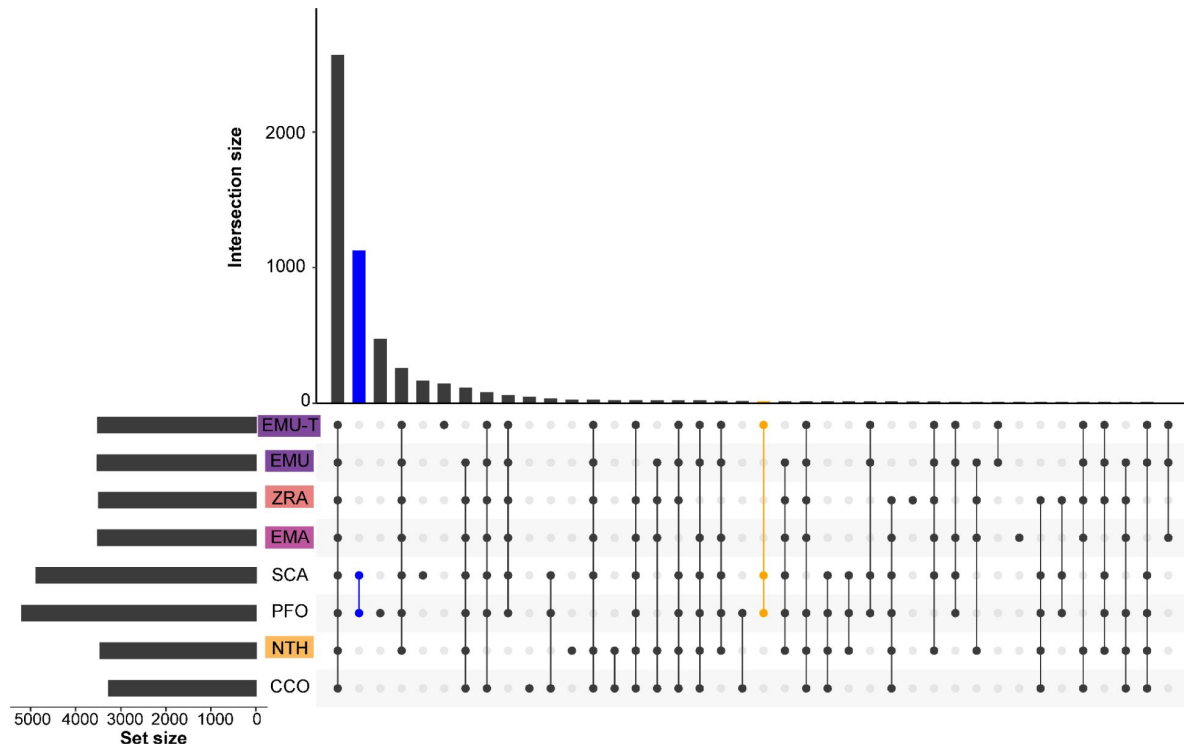


Figure 3-S1.

Pfam UpSet plot analysis including *E. muscae* transcriptome.

UpSet plots displaying Pfam domain intersections including additional predictions from an *E. muscae* transcriptomic dataset (EMU-T; NCBI GSE111046), showing the intersection among included genomes and transcriptomes. This analysis compares the genome predictions (i.e., EMU, EMA, ZRA, CTH and CCO) to the transcriptome predictions (i.e., EMU-T, SCA and PFO). Blue highlights the Pfam domains uniquely shared by SCA and PFO. Orange highlights the Pfam domains uniquely shared by all transcriptomic datasets in this analysis.

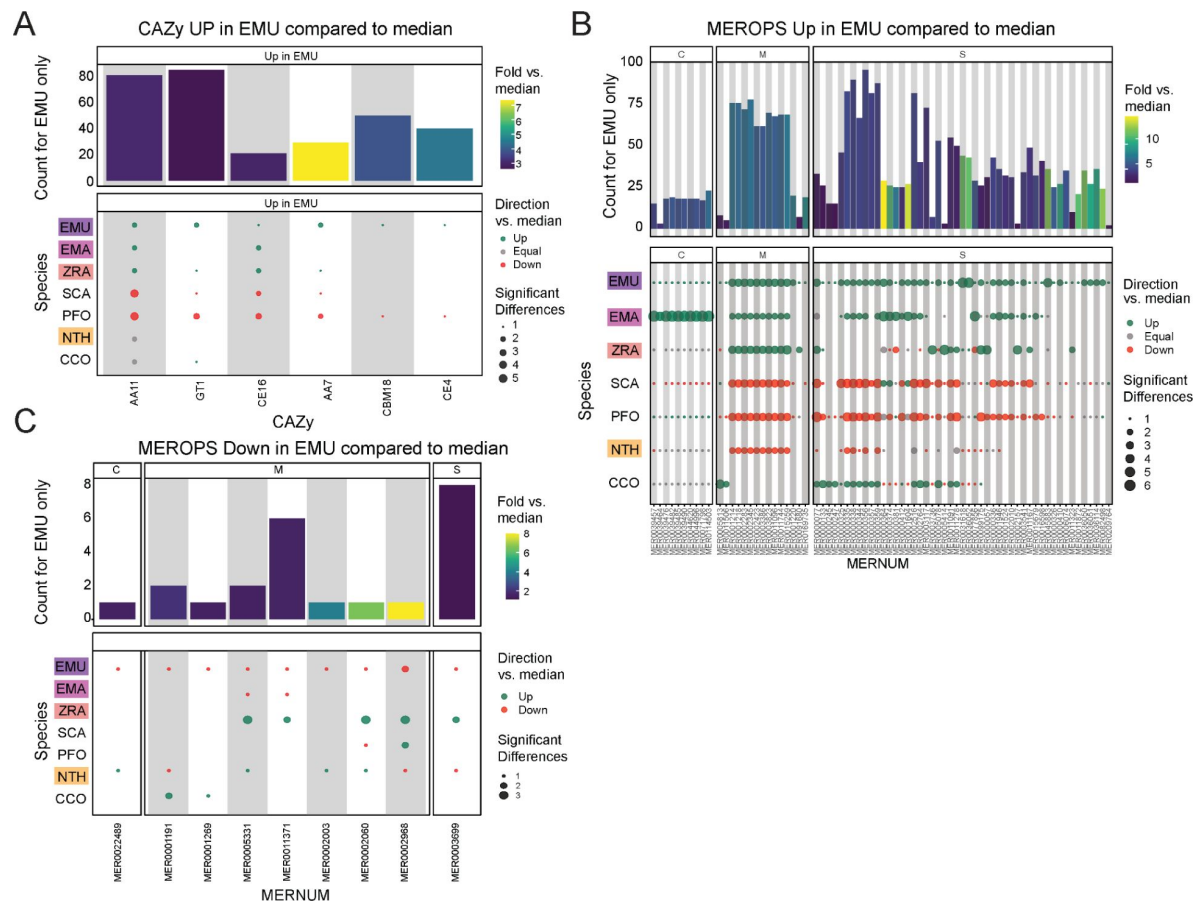


Figure 3-S2.

Additional domain analysis for CAZy and MEROPS databases.

A) CAZy domains significantly overrepresented in *E. muscae* (EMU) compared to other species analyzed (*E. maimaiga* (EMA), *Z. radicans* (ZRA), *S. castreans* (SCA), *P. formicae* (PFO), *N. thromboides* (CTH) and *C. coronatus* (CCO)). Bars represent the counts for *E. muscae* colored by fold-versus-the-median across all genomes. Point size represents the number of significant pairwise comparisons among other genomes and are colored according to whether the value is above, below or equal to the median value across all genomes. B) Plots following a similar format as panel A, displaying MEROPS domains that were found to be overrepresented in EMU by comparison, categorized by MEROPS peptidase category (C: cysteine peptidases, M: metallopeptidases, S: serine peptidases). C) Plot following a similar format to panel B, displaying MEROPS domains that were found to be underrepresented in EMU by comparison.

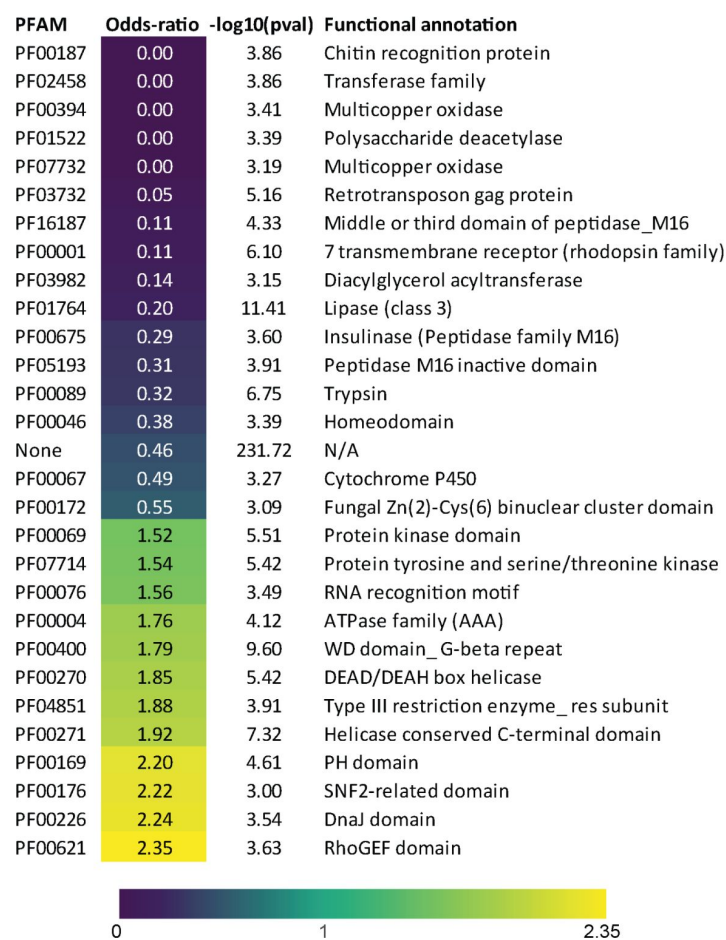


Figure 3-S3.

E. muscae core OG Pfam enrichment.

Enrichment among Pfam annotations for *E. muscae* genes belonging to core OG set. Odds-ratios are colored according to the scale bar below.

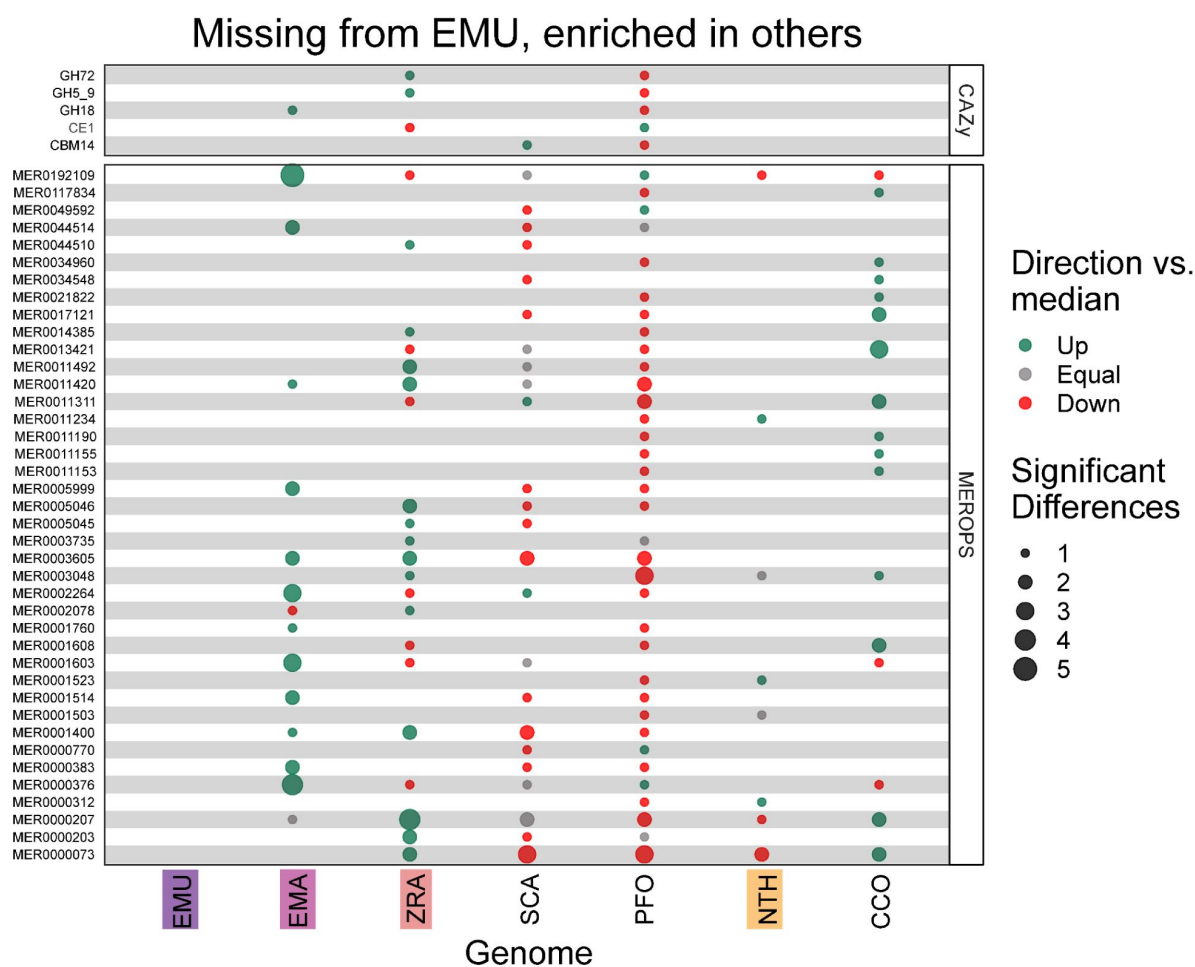


Figure 5-S1.

CAZy and MEROPS domains missing in *E. muscae* and enriched in other fungi.

CAZy and MEROPS domains missing from *E. muscae* (EMU), but significantly underrepresented (red) or overrepresented (green) in other species analyzed (*E. maimaiga* (EMA), *Z. radicans* (ZRA), *S. castrans* (SCA), *P. formicae* (PFO), *N. thromboides* (CTH) and *C. coronatus* (CCO)).

Conflicts of interest

JES was a paid consultant for Zymergen, Sincarne, and Michroma and is a CIFAR fellow in the program Fungal Kingdom: Threats and Opportunities.

Acknowledgements

We are grateful to The Bauer Core Facility at Harvard University for performing library preparation and sequencing of *E. muscae* genomic DNA and for the Bioinformatics team at The FAS Informatics Group at Harvard University for their support in processing and analyzing the resultant data. We also thank the Joint Genome Institute (JGI) for making available the assembly and annotation for the *Entomophaga maimaiga* genome produced under proposal 10.46936/10.25585/60001019 and available at <https://mycocosm.jgi.doe.gov/Entmai1>. JGI (<https://ror.org/04xm1d337>) is a Department of Energy User Facility supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231. Work to develop the *Entomophaga maimaiga* genome was also supported by the Genomics Facility at Cornell University and Dr. J. Romero-Severson of Notre Dame University. Finally, kudos to Ryan Bracewell for identifying drosophilids infected with *E. muscae* (SoCal cadaver 1 and LTE cadaver) and providing these specimens for this study.

References

- Armenteros JJ Almagro, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. (2019) **SignalP 5.0 improves signal peptide predictions using deep neural networks** *Nat Biotechnol* **37**:420–423 <https://doi.org/10.1038/s41587-019-0036-z>
- Amses KR *et al.* (2022) **Diploid-dominant life cycles characterize the early evolution of Fungi** *Proc Natl Acad Sci U S A* **119** <https://doi.org/10.1073/pnas.2116841119>
- Araújo JPM, Hughes DP (2016) **Diversity of entomopathogenic fungi: Which groups conquered the insect body?** *Adv Genet* **94**:1–39 <https://doi.org/10.1016/bs.adgen.2016.01.001>
- Arnesen JA, Małagocka J, Gryganskyi A, Grigoriev IV, Voigt K, Stajich JE, De Fine Licht HH. (2018) **Early Diverging Insect-Pathogenic Fungi of the Order Entomophthorales Possess Diverse and Unique Subtilisin-Like Serine Proteases** *G* **3**:3311–3319 <https://doi.org/10.1534/g3.118.200656>
- Ashburner M *et al.* (2000) **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium** *Nat Genet* **25**:25–29 <https://doi.org/10.1038/75556>
- Auguie B (2017) **gridExtra: Miscellaneous Functions for "Grid" Graphics** *CRAN*
- Baccile JA *et al.* (2016) **Plant-like biosynthesis of isoquinoline alkaloids in *Aspergillus fumigatus*** *Nat Chem Biol* **12**:419–424 <https://doi.org/10.1038/nchembio.2061>
- Bao W, Kojima KK, Kohany O (2015) **Rebase Update, a database of repetitive elements in eukaryotic genomes** *Mob DNA* **6** <https://doi.org/10.1186/s13100-015-0041-9>

- Belyayev A (2014) **Bursts of transposable elements as an evolutionary driving force** *J Evol Biol* **27**:2573–2584 <https://doi.org/10.1111/jeb.12513>
- Bewick AJ, Hofmeister BT, Powers RA, Mondo SJ, Grigoriev IV, James TY, Stajich JE, Schmitz RJ (2019) **Diversity of cytosine methylation across the fungal tree of life** *Nat Ecol Evol* **3**:479–490 <https://doi.org/10.1038/s41559-019-0810-9>
- Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema MH, Weber T. (2021) **antiSMASH 6.0: improving cluster detection and comparison capabilities** *Nucleic Acids Res* **49**:W29–W35 <https://doi.org/10.1093/nar/gkab335>
- Blum M *et al.* (2021) **The InterPro protein families and domains database: 20 years on** *Nucleic Acids Res* **49**:D344–D354 <https://doi.org/10.1093/nar/gkaa977>
- Boer P (2008) **Observations of summit disease in *Formica rufa* Linnaeus, 1761 (Hymenoptera: Formicidae).** *Myrmecological News* **11**
- Boyce GR *et al.* (2019) **Psychoactive plant- and mushroom-associated alkaloids from two behavior modifying cicada pathogens** *Fungal Ecol* **41**:147–164 <https://doi.org/10.1016/j.funeco.2019.06.002>
- Bronski M, Elya C, Stajich JE, Eisen MB (2018) **Run: SRR18312935 - Illumina HiSeq 4000 paired end sequencing; 10X DNA Sequencing of *Entomophthora muscae* strain UCB** *European Nucleotide Archive*
- Buchfink B, Reuter K, Drost H-G (2021) **Sensitive protein alignments at tree-of-life scale using DIAMOND** *Nat Methods* **18**:366–368 <https://doi.org/10.1038/s41592-021-01101-x>
- Bushley KE, Turgeon BG (2010) **Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships** *BMC Evol Biol* **10** <https://doi.org/10.1186/1471-2148-10-26>
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) **BLAST+: architecture and applications** *BMC Bioinformatics* **10** <https://doi.org/10.1186/1471-2105-10-421>
- Cambareri EB, Singer MJ, Selker EU (1991) **Recurrence of repeat-induced point mutation (RIP) in *Neurospora crassa*** *Genetics* **127**:699–710 <https://doi.org/10.1093/genetics/127.4.699>
- Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J (2021) **eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale** *Mol Biol Evol* **38**:5825–5829 <https://doi.org/10.1093/molbev/msab293>
- Cascant-Lopez E, Crosthwaite SK, Johnson LJ, Harrison RJ (2020) **No evidence that homologs of key circadian clock genes direct circadian programs of development or mRNA abundance in *Verticillium dahliae*** *Front Microbiol* **11** <https://doi.org/10.3389/fmicb.2020.01977>
- Cen K, Li B, Lu Y, Zhang S, Wang C (2017) **Divergent LysM effectors contribute to the virulence of *Beauveria bassiana* by evasion of insect immune defenses** *PLoS Pathog* **13** <https://doi.org/10.1371/journal.ppat.1006604>
- Cha J, Zhou M, Liu Y (2015) **Mechanism of the *Neurospora* circadian clock, a FREQUENCY-centric view** *Biochemistry* **54**:150–156 <https://doi.org/10.1021/bi5005624>

- Chang Y *et al.* (2022) **Evolution of zygomycete secretomes and the origins of terrestrial fungal ecologies** *iScience* **104840** <https://doi.org/10.1016/j.isci.2022.104840>
- Cooley JR, Marshall DC, Hill KBR (2018) **A specialized fungal parasite (*Massospora cicadina*) hijacks the sexual signals of periodical cicadas (Hemiptera: Cicadidae: Magicicada)** *Sci Rep* **8** <https://doi.org/10.1038/s41598-018-19813-0>
- de Bekker C, Beckerson WC, Elya C. (2021) **Mechanisms behind the Madness: How Do Zombie-Making Fungal Entomopathogens Affect Host Behavior To Increase Transmission?** *MBio* **12**:e01872–21 <https://doi.org/10.1128/mBio.01872-21>
- de Bekker C, Das B. (2022) **Hijacking time: How Ophiocordyceps fungi could be using ant host clocks to manipulate behavior** *Parasite Immunol* **e 12909** <https://doi.org/10.1111/pim.12909>
- Licht HH De Fine, Jensen AB, Eilenberg J. (2017) **Comparative transcriptomics reveal host-specific nucleotide variation in entomophthoralean fungi** *Mol Ecol* **26**:2092–2110 <https://doi.org/10.1111/mec.13863>
- Drula E, Garron M-L, Dogan S, Lombard V, Henrissat B, Terrapon N (2022) **The carbohydrate-active enzyme database: functions and literature** *Nucleic Acids Res* **50**:D571–D577 <https://doi.org/10.1093/nar/gkab1045>
- Dustan AG (1924) **Studies on a new species of *Empusa* parasitic on the green apple bug (*Lygus communis* var. *novascotiensis* Knight) in the Annapolis Valley** *Acadian Entomological Society* :14–36
- Eddy SR (2011) **Accelerated Profile HMM Searches** *PLoS Comput Biol* **7** <https://doi.org/10.1371/journal.pcbi.1002195>
- El-Gebali S *et al.* (2019) **The Pfam protein families database in 2019** *Nucleic Acids Res* **47**:D427–D432 <https://doi.org/10.1093/nar/gky995>
- Elya C, De Fine Licht HH (2021) **The genus *Entomophthora*: bringing the insect destroyers into the twenty-first century** *IMA Fungus* **12**:1–31 <https://doi.org/10.1186/s43008-021-00084-w>
- Elya C, Lee E. (2022) **Isolation of high molecular weight genomic DNA from *Entomophthora muscae*** *Protocols.io* <https://doi.org/10.17504/protocols.io.rm7vzyje2lx1/v1>
- Elya C, Lok TC, Spencer QE, McCausland H, Martinez CC, Eisen M (2018) **Robust manipulation of the behavior of *Drosophila melanogaster* by a fungal pathogen in the laboratory** *Elife* **7** <https://doi.org/10.7554/eLife.34414>
- Emms DM, Kelly S (2015) **OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy** *Genome Biol* **16** <https://doi.org/10.1186/s13059-015-0721-2>
- Evans HC (1989) **Mycopathogens of Insects of Epigeal and Aerial Habitats In: Wilding N, Collins NM, Hammond PM Webber JF, editors. *Insect-Fungus Interactions*. London: Academic Press :205–238** <https://doi.org/10.1016/B978-0-12-751800-8.50015-X>

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF (2020) **RepeatModeler2 for automated genomic discovery of transposable element families** *Proc Natl Acad Sci U S A* **117**:9451–9457 <https://doi.org/10.1073/pnas.1921046117>

Freimoser FM, Grundschober A, Aebi M, Tuor U (2000) **In vitro cultivation of the entomopathogenic fungus *Entomophthora thripidum*: isolation, growth requirements, and sporulation** *Mycologia* **92**:208–215 <https://doi.org/10.2307/3761552>

Freitag M, Williams RL, Kothe GO, Selker EU (2002) **A cytosine methyltransferase homologue is essential for repeat-induced point mutation in *Neurospora crassa*** *Proc Natl Acad Sci U S A* **99**:8802–8807 <https://doi.org/10.1073/pnas.132212899>

Galagan JE *et al.* (2003) **The genome sequence of the filamentous fungus *Neurospora crassa*** *Nature* **422**:859–868 <https://doi.org/10.1038/nature01554>

Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV (2021) **COG database update: focus on microbial diversity, model organisms, and widespread pathogens** *Nucleic Acids Res* **49**:D274–D281 <https://doi.org/10.1093/nar/gkaa1018>

Garnier S, Ross N, Rudis R, Camargo AP, Sciaini M, Scherer C (2021) **viridis - Colorblind-Friendly Color Maps for R** *CRAN*

Consortium Gene Ontology *et al.* (2023) **The Gene Ontology knowledgebase in 2023** *Genetics* <https://doi.org/10.1093/genetics/iyad031>

Grundschober A, Tuor U, Aebi M (1998) **In vitro Cultivation and Sporulation of *Neozygites parvispora* (Zygomycetes: Entomophthorales)** *Syst Appl Microbiol* **21**:461–469 [https://doi.org/10.1016/S0723-2020\(98\)80057-X](https://doi.org/10.1016/S0723-2020(98)80057-X)

Gryganskyi AP, Humber RA, Stajich JE, Mullens B, Anishchenko IM, Vilgalys R (2013) **Sequential utilization of hosts from different fly families by genetically distinct, sympatric populations within the *Entomophthora muscae* species complex** *PLoS One* **8** <https://doi.org/10.1371/journal.pone.0071168>

Gryganskyi AP, Muszewska A (2014) **Whole Genome Sequencing and the Zygomycota** *Fungal Genom Biol* **4** <https://doi.org/10.4172/2165-8056.1000e116>

Gryganskyi AP *et al.* (2022) **The early terrestrial fungal lineage of *Conidiobolus*-transition from saprotroph to parasitic lifestyle** *J Fungi (Basel)* **8** <https://doi.org/10.3390/jof8080789>

Gupta YK *et al.* (1835) **2023. Major proliferation of transposable elements shaped the genome of the soybean rust pathogen *Phakopsora pachyrhizi*** *Nat Commun* **14** <https://doi.org/10.1038/s41467-023-37551-4>

Haas BJ *et al.* (2013) **De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis** *Nat Protoc* **8**:1494–1512 <https://doi.org/10.1038/nprot.2013.084>

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR (2008) **Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments** *Genome Biol* **9** <https://doi.org/10.1186/gb-2008-9-1-r7>

- Hajek AE, Harris DC (2023) **Diurnal patterns and conidial dynamics of *Batkoa* major, a generalist entomophthoralean pathogen** *Fungal Ecol* **65** <https://doi.org/10.1016/j.funeco.2023.101278>
- Hajek AE, Humber RA, Elkinton JS, May B, Walsh SR, Silver JC (1990) **Allozyme and restriction fragment length polymorphism analyses confirm *Entomophaga maimaiga* responsible for 1989 epizootics in North American gypsy moth populations** *Proc Natl Acad Sci U S A* **87**:6979–6982 <https://doi.org/10.1073/pnas.87.18.6979>
- Hajek AE, Papierok B, Eilenberg J (2012) **Chapter IX - Methods for study of the Entomophthorales In: Lacey LA, editor *Manual of Techniques in Invertebrate Pathology* (Second Edition). San Diego: Academic Press :285–316** <https://doi.org/10.1016/B978-0-12-386899-2.00009-9>
- Holdom DG (1983) **In vitro culture of the aphid pathogenic fungus *Entomophthora planchoniana* Cornu (zygomycetes: Entomophthorales)** *Aust J Entomol* **22**:188–188 <https://doi.org/10.1111/j.1440-6055.1983.tb01870.x>
- Hood ME, Katawczik M, Giraud T (2005) **Repeat-induced point mutation and the population structure of transposable elements in *Microbotryum violaceum*** *Genetics* **170**:1081–1089 <https://doi.org/10.1534/genetics.105.042564>
- Hua L, Feng M-G (2003) **New use of broomcorn millets for production of granular cultures of aphid-pathogenic fungus *Pandora neoaphidis* for high sporulation potential and infectivity to *Myzus persicae*** *FEMS Microbiol Lett* **227**:311–317 [https://doi.org/10.1016/S0378-1097\(03\)00711-0](https://doi.org/10.1016/S0378-1097(03)00711-0)
- Huang L, Zhang H, Wu P, Entwistle S, Li X, Yohe T, Yi H, Yang Z, Yin Y (2018) **dbCAN-seq: a database of carbohydrate-active enzyme (CAZyme) sequence and annotation** *Nucleic Acids Res* **46**:D516–D521 <https://doi.org/10.1093/nar/gkx894>
- Huerta-Cepas J *et al.* (2019) **eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses** *Nucleic Acids Res* **47**:D309–D314 <https://doi.org/10.1093/nar/gky1085>
- Humber RA (2016) **Entomophthoromycota: A New Overview of Some of the Oldest Terrestrial Fungi** *Biology of Microfungi*. Springer *Cham* :127–145 https://doi.org/10.1007/978-3-319-29137-6_6
- Jackman SD *et al.* (2018) **Tigmint: correcting assembly errors using linked reads from large molecules** *BMC Bioinformatics* **19** <https://doi.org/10.1186/s12859-018-2425-6>
- James TY *et al.* (2006) **Reconstructing the early evolution of Fungi using a six-gene phylogeny** *Nature* **443**:818–822 <https://doi.org/10.1038/nature05110>
- Jensen AB, Thomsen L, Eilenberg J (2006) **Value of host range, morphological, and genetic characteristics within the *Entomophthora muscae* species complex** *Mycol Res* **110**:941–950 <https://doi.org/10.1016/j.mycres.2006.06.003>
- Jones P *et al.* (2014) **InterProScan 5: genome-scale protein function classification** *Bioinformatics* **30**:1236–1240 <https://doi.org/10.1093/bioinformatics/btu031>

- Käll L, Krogh A, Sonnhammer ELL (2004) **A combined transmembrane topology and signal peptide prediction method** *J Mol Biol* **338**:1027–1036 <https://doi.org/10.1016/j.jmb.2004.03.016>
- Katoh K, Standley DM (2013) **MAFFT multiple sequence alignment software version 7: improvements in performance and usability** *Mol Biol Evol* **30**:772–780 <https://doi.org/10.1093/molbev/mst010>
- Keller S (2002) **The genus *Entomophthora* (Zygomycetes Entomophthorales) with a description of five new species.** *Sydowia* **54**:157–197
- Kershaw MJ, Moorhouse ER, Bateman R, Reynolds SE, Charnley AK (1999) **The role of destruxins in the pathogenicity of *Metarhizium anisopliae* for three species of insect** *Invertebr Pathol* **74**:213–223 <https://doi.org/10.1006/jipa.1999.4884>
- Khan AY, Suresh Kumar G (2015) **Natural isoquinoline alkaloids: binding aspects to functional proteins, serum albumins, hemoglobin, and lysozyme** *Biophys Rev* **7**:407–420 <https://doi.org/10.1007/s12551-015-0183-5>
- Kim D, Paggi JM, Park C, Bennett C, Salzberg SL (2019) **Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype** *Nat Biotechnol* **37**:907–915 <https://doi.org/10.1038/s41587-019-0201-4>
- Kirchner J, Sandmeyer S (1993) **Proteolytic processing of Ty3 proteins is required for transposition** *J Virol* **67**:19–28 <https://doi.org/10.1128/JVI.67.1.19-28.1993>
- Koczyk G, Pawłowska J, Muszewska A (2021) **Terpenoid Biosynthesis Dominates among Secondary Metabolite Clusters in *Mucoromycotina* Genomes** *J Fungi (Basel)* **7** <https://doi.org/10.3390/jof7040285>
- Kolmogorov M (2021) **Flye: De novo assembler for single molecule sequencing reads using repeat graphs** *GitHub*
- Kolmogorov M, Yuan J, Lin Y, Pevzner PA (2019) **Assembly of long, error-prone reads using repeat graphs** *Nat Biotechnol* **37**:540–546 <https://doi.org/10.1038/s41587-019-0072-8>
- Korf I (2004) **Gene finding in novel genomes** *BMC Bioinformatics* **5** <https://doi.org/10.1186/1471-2105-5-59>
- Krasnoff SB, Watson DW, Gibson DM, Kwan EC. (1995) **Behavioral effects of the entomopathogenic fungus, *Entomophthora muscae* on its host *Musca domestica*: Postural changes in dying hosts and gated pattern of mortality** *J Insect Physiol* **41**:895–903 [https://doi.org/10.1016/0022-1910\(95\)00026-Q](https://doi.org/10.1016/0022-1910(95)00026-Q)
- Krassowski M, Arts M, Lager C (2022) **Max krassowski/complex-upset: v1.3.5** <https://doi.org/10.5281/zenodo.7314197>
- Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) **MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms** *Mol Biol Evol* **35**:1547–1549 <https://doi.org/10.1093/molbev/msy096>
- Kutchan TM, Dittrich H (1995) **Characterization and mechanism of the berberine bridge enzyme, a covalently flavinylated oxidase of benzophenanthridine alkaloid biosynthesis in plants** *J Biol Chem* **270**:24475–24481 <https://doi.org/10.1074/jbc.270.41.24475>

- Landan G, Graur D (2007) **Local reliability measures from sets of co-optimal multiple sequence alignments** *Biocomputing 2008 World Scientific* :15–24 https://doi.org/10.1142/9789812776136_0003
- Lewis ZA, Honda S, Khlafallah TK, Jeffress JK, Freitag M, Mohn F, Schübeler D, Selker EU (2009) **Relics of repeat-induced point mutation direct heterochromatin formation in *Neurospora crassa*** *Genome Res* **19**:427–437 <https://doi.org/10.1101/gr.086231.108>
- Lovett B, Macias A, Stajich JE, Cooley J, Eilenberg J, de Fine Licht HH, Kasson MT. (2020) **Behavioral betrayal: How select fungal parasites enlist living insects to do their bidding** *PLoS Pathog* **16** <https://doi.org/10.1371/journal.ppat.1008598>
- Lovett B, St Leger RJ (2017) **The Insect Pathogens** *Microbiol Spectr* **5** <https://doi.org/10.1128/microbiolspec.FUNK-0001-2016>
- Lovett B, St Leger RJ, De Fine Licht HH. (2020) **Going gentle into that pathogen-induced goodnight** *J Invertebr Pathol* **107398** <https://doi.org/10.1016/j.jip.2020.107398>
- Lowe TM, Eddy SR (1997) **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence** *Nucleic Acids Res* **25**:955–964 <https://doi.org/10.1093/nar/25.5.955>
- Macias AM, Geiser DM, Stajich JE, Łukasik P, Veloso C, Bublitz DC, Berger MC, Boyce GR, Hodge K, Kasson MT (2020) **Evolutionary relationships among *Massospora* spp (Entomophthorales), obligate pathogens of cicadas.** *Mycologia* :1–15 <https://doi.org/10.1080/00275514.2020.1742033>
- MacLeod DM, Müller-Kögler E (1973) **Entomogenous fungi: Entomophthora species with pear-shaped to almost spherical conidia (Entomophthorales: Entomophthoraceae)** *Mycologia* **65**:823–893
- MacLeod DM, Müller-Kögler E, Wilding N (1976) **Entomophthora species with E. muscae-like conidia** *Mycologia* **68**:1–29 <https://doi.org/10.2307/3758894>
- Majoros WH, Pertea M, Salzberg SL (2004) **TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders** *Bioinformatics* **20**:2878–2879 <https://doi.org/10.1093/bioinformatics/bth315>
- Małagocka J, Grell MN, Lange L, Eilenberg J, Jensen AB (2015) **Transcriptome of an entomophthoralean fungus (*Pandora formicae*) shows molecular machinery adjusted for successful host exploitation and transmission** *J Invertebr Pathol* **128**:47–56 <https://doi.org/10.1016/j.jip.2015.05.001>
- Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM (2021) **BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes** *Mol Biol Evol* **38**:4647–4654 <https://doi.org/10.1093/molbev/msab199>
- Marçais G, Kingsford C (2011) **A fast, lock-free approach for efficient parallel counting of occurrences of k-mers** *Bioinformatics* **27**:764–770 <https://doi.org/10.1093/bioinformatics/btr011>
- Marikovsky PI (1962) **On some features of behavior of the ants *Formica rufa* L. infected with fungous disease** *Insectes Soc* **9**:173–179 <https://doi.org/10.1007/BF02224263>

- Milner RJ, Holdom DG, Glare TR (1984) **Diurnal patterns of mortality in aphids infected by entomophthoran fungi** *Entomol Exp Appl* **36**:37–42 <https://doi.org/10.1111/j.1570-7458.1984.tb03404.x>
- Mistry J *et al.* (2021) **Pfam: The protein families database in 2021** *Nucleic Acids Res* **49**:D412–D419 <https://doi.org/10.1093/nar/gkaa913>
- Mohanta TK, Bae H (2015) **The diversity of fungal genome** *Biol Proced Online* **17** <https://doi.org/10.1186/s12575-015-0020-z>
- Montenegro-Montero A, Canessa P, Larrondo LF (2015) **Around the Fungal Clock: Recent Advances in the Molecular Study of Circadian Clocks in Neurospora and Other Fungi** *Adv Genet* **92**:107–184 <https://doi.org/10.1016/bs.adgen.2015.09.003>
- Mullens BA (1986) **A Method for Infecting Large Numbers of *Musca domestica* (Diptera: Muscidae) with *Entomophthora muscae* (Entomophthorales: Entomophthoraceae)** *J Med Entomol* **23**:457–458 <https://doi.org/10.1093/jmedent/23.4.457>
- Mullens BA, Rodriguez JL, Meyer JA (1987) **An Epizootiological Study of *Entomophthora muscae* in Muscoid Fly Populations on Southern California Poultry Facilities, with Emphasis on *Musca domestica*** *Hilgardia* **55** <https://doi.org/10.3733/hilg.v55n03p041>
- Muszewska A, Hoffman-Sommer M, Grynberg M (2011) **LTR retrotransposons in fungi** *PLoS One* **6** <https://doi.org/10.1371/journal.pone.0029425>
- Muszewska A, Taylor JW, Szczesny P, Grynberg M (2011) **Independent subtilases expansions in fungi associated with animals** *Mol Biol Evol* **28**:3395–3404 <https://doi.org/10.1093/molbev/msr176>
- Nakazawa M. (2022) **fmsb: Functions for Medical Statistics Book with some Demographic Data** *CRAN*
- Naundrup A, Bohman B, Kwadha CA, Jensen AB, Becher PG, De Fine Licht HH. (2022) **Pathogenic fungus uses volatiles to entice male flies into fatal matings with infected female cadavers** *ISME J* **16**:2388–2397 <https://doi.org/10.1038/s41396-022-01284-x>
- Nicolás FE, Garre V (2016) **RNA Interference in Fungi: Retention and Loss** *Microbiol Spectr* **4** <https://doi.org/10.1128/microbiolspec.FUNK-0008-2016>
- Nielsen C, Hajek AE (2006) **Diurnal pattern of death and sporulation in *Entomophaga maimaiga*-infected *Lymantria dispar*** *Entomol Exp Appl* **118**:237–243 <https://doi.org/10.1111/j.1570-7458.2006.00382.x>
- Nie Y, Yu D-S, Wang C-F, Liu X-Y, Huang B (2020) **A taxonomic revision of the genus *Conidiobolus* (Ancylistaceae Entomophthorales): four clades including three new genera.** *MycoKeys* **66**:55–81 <https://doi.org/10.3897/mycokeys.66.46575>
- Oxford Nanopore Technologies Ltd. (2021) **medaka: Sequence correction provided by ONT Research** *GitHub*
- Palmer JM, Stajich J (2020) **Funannotate v1.8.1: Eukaryotic genome annotation** *Zenodo* <https://doi.org/10.5281/zenodo.4054262>

Paysan-Lafosse T *et al.* (2023) **InterPro in 2022** *Nucleic Acids Res* **51**:D418–D427 <https://doi.org/10.1093/nar/gkac993>

Pell JK, Eilenberg J, Hajek AE, Steinkraus DC (2001) **Biology, Ecology and Pest Management Potential of Entomophthorales** In: Butt TM, Jackson C, Magan N, editors. *Fungi as Biocontrol Agent* :71–153

Pickford R, Riegert PW (1964) **The fungous disease caused by *Entomophthora grylli* Fres , and its effects on grasshopper populations in Saskatchewan in 1963** <https://doi.org/10.4039/Ent961158-9>

Price MN, Dehal PS, Arkin AP (2009) **FastTree: computing large minimum evolution trees with profiles instead of a distance matrix** *Mol Biol Evol* **26**:1641–1650 <https://doi.org/10.1093/molbev/msp077>

Rambaut A. (2017) **FigTree-version 1.4.3, a graphical viewer of phylogenetic trees** *Molecular Evolution, Phylogenetics and Epidemiology*

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA (2018) **Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7** *Syst Biol* **67**:901–904 <https://doi.org/10.1093/sysbio/syy032>

Ranallo-Benavidez TR, Jaron KS, Schatz MC (2020) **GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes** *Nat Commun* **11** <https://doi.org/10.1038/s41467-020-14998-3>

Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD (2018) **The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database** *Nucleic Acids Res* **46**:D624–D632 <https://doi.org/10.1093/nar/gkx1134>

Core Team R (2021) **R: A language and environment for statistical computing**

Robinson D, Hayes A, Couch S. (2023) **broom: Convert Statistical Objects into Tidy Tibbles** *Comprehensive R Archive Network (CRAN)*

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. (2012) **MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space** *Syst Biol* **61**:539–542 <https://doi.org/10.1093/sysbio/sys029>

Roy HE, Steinkraus DC, Eilenberg J, Hajek AE, Pell JK (2006) **Bizarre interactions and endgames: entomopathogenic fungi and their arthropod hosts** *Annu Rev Entomol* **51**:331–357 <https://doi.org/10.1146/annurev.ento.51.110104.150941>

RStudio Team (2022) **RStudio: Integrated Development Environment for R**

Sacco NE, Hajek AE. (2023) **Diversity and Breadth of Host Specificity among Arthropod Pathogens in the Entomophthoromycotina** *Microorganisms* **11** <https://doi.org/10.3390/microorganisms11071658>

Salichos L, Rokas A (2010) **The diversity and evolution of circadian clock proteins in fungi** *Mycologia* **102**:269–278 <https://doi.org/10.3852/09-073>

- Scheinin I, Kalimeri M, Jagerroos V, Parkkinen J, Tikkanen E, Würtz P, Kangas A. (2020) **ggforestplot: An R package to create forestplots of measures of effects with their confidence intervals** *Github*
- Sela I, Ashkenazy H, Katoh K, Pupko T (2015) **GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters** *Nucleic Acids Res* **43**:W7–14 <https://doi.org/10.1093/nar/gkv318>
- Selker EU (2002) **Repeat-induced gene silencing in fungi** *Adv Genet* **46**:439–450 [https://doi.org/10.1016/s0065-2660\(02\)46016-6](https://doi.org/10.1016/s0065-2660(02)46016-6)
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs** *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btv351>
- Slater GSC, Birney E (2005) **Automated generation of heuristics for biological sequence comparison** *BMC Bioinformatics* **6** <https://doi.org/10.1186/1471-2105-6-31>
- Smit AFA, Hubley R, Green P (2022) **RepeatMasker Open-4.0 (4-1-1)** *Institute for Systems Biology*
- Spatafora JW *et al.* (2016) **A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data** *Mycologia* **108**:1028–1046 <https://doi.org/10.3852/16-042>
- Stajich JE (2023) **PHYling for Phylogenomic reconstruction from genomes** *Zenodo* <https://doi.org/10.5281/zenodo.1257001>
- Stajich JE., Lovett B, Ettinger CL., Carter-House DA., Kurbessoian T, Kasson MT (2022) **An Improved 1.5-Gigabase Draft Assembly of Massospora cicadina (Zoopagomycota), an Obligate Fungal Parasite of 13- and 17-Year Cicadas** *Microbiology Resource Announcements* **11**:e00367–22 <https://doi.org/10.1128/mra.00367-22>
- Stajich JE, Lovett B, Lee E, Macias AM, Hajek AE, de Bivort BL, Kasson MT, de Fine Licht HH, Elya C (2023) **Genome annotation of Entomophthora muscae strain UCB** *Zenodo* <https://doi.org/10.5281/ZENODO.8339744>
- Stajich JE, Lovett B, Lee E, Macias AM, Hajek AE, de Bivort BL, Kasson MT, de Fine Licht HH, Elya C. (2023) **Entomophthora muscae comparative genomics** *Zenodo* <https://doi.org/10.5281/ZENODO.8339767>
- Stamatakis A. (2014) **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies** *Bioinformatics* **30**:1312–1313 <https://doi.org/10.1093/bioinformatics/btu033>
- Stanke M, Diekhans M, Baertsch R, Haussler D (2008) **Using native and syntenically mapped cDNA alignments to improve de novo gene finding** *Bioinformatics* **24**:637–644 <https://doi.org/10.1093/bioinformatics/btn013>
- Stecher G, Tamura K, Kumar S (2020) **Molecular Evolutionary Genetics Analysis (MEGA) for macOS** *Mol Biol Evol* **37**:1237–1239 <https://doi.org/10.1093/molbev/msz312>
- Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) **The COG database: a tool for genome-scale analysis of protein functions and evolution** *Nucleic Acids Res* **28**:33–36 <https://doi.org/10.1093/nar/28.1.33>

- Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M (2008) **Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training** *Genome Res* **18**:1979–1990 <https://doi.org/10.1101/gr.081612.108>
- Testa AC, Hane JK, Ellwood SR, Oliver RP (2015) **CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts** *BMC Genomics* **16** <https://doi.org/10.1186/s12864-015-1344-4>
- The UniProt Consortium, et al. (2020) **UniProt: the universal protein knowledgebase in 2021** *Nucleic Acids Res* **49**:D480–D489 <https://doi.org/10.1093/nar/gkaa1100>
- Thompson SN (2003) **Trehalose – The Insect “Blood” Sugar** *Advances in Insect Physiology Academic Press* :205–285 [https://doi.org/10.1016/S0065-2806\(03\)31004-5](https://doi.org/10.1016/S0065-2806(03)31004-5)
- Tobias PA et al. (2021) **Austropuccinia psidii, causing myrtle rust, has a gigabase-sized genome shaped by transposable elements** *G3* <https://doi.org/10.1093/g3journal/jkaa015>
- van Wyk S, Wingfield BD, De Vos L, van der Merwe NA, Steenkamp ET. (2020) **Genome-Wide Analyses of Repeat-Induced Point Mutations in the Ascomycota** *Front Microbiol* **11** <https://doi.org/10.3389/fmicb.2020.622368>
- Vasimuddin M, Misra S, Li H, Aluru S (2019) **Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems** *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* :314–324 <https://doi.org/10.1109/IPDPS.2019.00041>
- Voigt K, Wolf T, Ochsenreiter K, Nagy G, Kaerger K, Shelest E, Papp T., Hoffmeister D (2016) **Genetic and Metabolic Aspects of Primary and Secondary Metabolism of the Zygomycetes** *Biochemistry and Molecular Biology* :361–385 https://doi.org/10.1007/978-3-319-27790-5_15
- Wang C, St Leger RJ (2006) **A collagenous protective coat enables Metarhizium anisopliae to evade insect immune responses** *Proc Natl Acad Sci U S A* **103**:6647–6652 <https://doi.org/10.1073/pnas.0601951103>
- Wang Y et al. (2023) **Divergent evolution of early terrestrial fungi reveals the evolution of Mucormycosis pathogenicity factors** *Genome Biol Evol* <https://doi.org/10.1093/gbe/evad046>
- Wei KHC, Aldaimalani R, Mai D, Zinshteyn D, Satyaki PRV, Blumenstiel J, Kelleher ES, Brooks EC (2022) **Rethinking the “gypsy” retrotransposon: A roadmap for community-driven reconsideration of problematic gene names** *OSF Preprints* <https://doi.org/10.31219/osf.io/fma57>
- Wickham H et al. (2019) **Welcome to the tidyverse** *Journal of Open Source Software* **4** <https://doi.org/10.21105/joss.01686>
- Wolf YI, Koonin EV (2013) **Genome reduction as the dominant mode of evolution** *Bioessays* **35**:829–837 <https://doi.org/10.1002/bies.201300037>
- Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ (2016) **GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality** *Methods Mol Biol* **1418**:283–334 https://doi.org/10.1007/978-1-4939-3578-9_15
- Yeo S, Coombe L, Warren RL, Chu J, Biro I (2018) **ARCS: scaffolding genome drafts with linked reads** *Bioinformatics* **34**:725–731 <https://doi.org/10.1093/bioinformatics/btx675>

Yildirim G, Malar C M, Kokkoris V, Corradi N (2020) **Parasexual and Sexual Reproduction in Arbuscular Mycorrhizal Fungi: Room for Both** *Trends Microbiol* **28**:517–519 <https://doi.org/10.1016/j.tim.2020.03.013>

Zhang S, Wang Y, Wu H, Li N, Jiang J, Guo Y, Feng Y, Xiao L (2019) **Characterization of a Species-Specific Insulinase-Like Protease in *Cryptosporidium parvum*** *Front Microbiol* **10** <https://doi.org/10.3389/fmicb.2019.00354>

Article and author information

Jason E. Stajich

Department of Microbiology and Plant Pathology, University of California-Riverside, Riverside, CA United States

For correspondence: jason.stajich@ucr.edu

ORCID iD: [0000-0002-7591-0020](https://orcid.org/0000-0002-7591-0020)

Brian Lovett

Emerging Pests and Pathogens Research Unit, USDA-ARS, Ithaca, NY, United States

ORCID iD: [0000-0002-5721-7695](https://orcid.org/0000-0002-5721-7695)

Emily Lee

Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, United States

Angie M. Macias

Division of Plant and Soil Sciences, West Virginia University, Morgantown, WV, United States

ORCID iD: [0000-0002-2975-686X](https://orcid.org/0000-0002-2975-686X)

Ann E. Hajek

Department of Entomology, Cornell University, Ithaca, NY, United States

ORCID iD: [0000-0001-5740-4717](https://orcid.org/0000-0001-5740-4717)

Benjamin L. de Bivort

Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, United States

ORCID iD: [0000-0001-6165-7696](https://orcid.org/0000-0001-6165-7696)

Matt T. Kason

Division of Plant and Soil Sciences, West Virginia University, Morgantown, WV, United States

ORCID iD: [0000-0001-5602-7278](https://orcid.org/0000-0001-5602-7278)

Henrik H. De Fine Licht

Section for Organismal Biology, Department of Plant and Environmental Sciences, University of Copenhagen, Copenhagen, Denmark

ORCID iD: [0000-0003-3326-5729](https://orcid.org/0000-0003-3326-5729)

Carolyn Elya

Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, United States, Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, United States

For correspondence: cnelya@g.harvard.edu

ORCID iD: [0000-0002-9634-0303](https://orcid.org/0000-0002-9634-0303)

Copyright

© 2023, Stajich et al.

This article is distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use and redistribution provided that the original author and source are credited.

Editors

Reviewing Editor

Detlef Weigel

Max Planck Institute for Biology Tübingen, Tübingen, Germany

Senior Editor

Detlef Weigel

Max Planck Institute for Biology Tübingen, Tübingen, Germany

Reviewer #1 (Public Review):

Summary:

The authors present a detailed study of a nearly complete *Entomophthora muscae* genome assembly and annotation, along with comparative analyses among related and non-related entomopathogenic fungi. The genome is one of the largest fungal genomes sequenced, and the authors document the proliferation and evolution of transposons and presence/absence of related genetic machinery to explore how this may have occurred. There has also been an expansion in gene number, which appears to contain many "novel" genes unique to *E. muscae*. Functionally, the authors were interested in CAZymes, proteases, circadian clock related genes (due to entomopathogenicity/ host manipulation), other insect pathogen specific genes, and secondary metabolites. There are many interesting findings including expansions in trahalases, unique insulinase and another peptidase, and some evidence for RIP in Entomophthoralean fungi. The authors performed a separate study examining *E. muscae* species complex and related strains. Specifically, morphological traits were measured for strains and then compared to the 28S+ITS-based phylogeny, showing little informativeness of these morpho characters with high levels of overlap.

This work represents a big leap forward in genomics of non-Dikarya fungi and large fungal genomes. Most of the gene homologs have been studied in species that diverged hundreds of millions of years ago, and therefore using standard comparative genomic approaches are not trivial and still relatively little is known. This paper provides many new hypotheses and potential avenues of research about fungal genome size expansion, entomopathogenesis in zygomycetes, and cellular functions like RIP and circadian mechanisms.

Strengths:

There are many strengths to this study. It represents a massive amount of work and a very thorough functional analysis of the gene content in these fungi (which are largely unsequenced and definitely understudied). Too often comparative genomic work will focus on one aspect and leave the reader wondering about all the other ways genome(s) are unique or different from others. This study really dove in and explored the relevant aspects of the *E. muscae* genome.

The authors used both a priori and emergent properties to shape their analyses (by searching for specific genes of interest and by analyzing genes underrepresented, expanded, or unique to their chosen taxa), enabling a detailed review of the genomic architecture and content. Specifically, I'm impressed by the analysis of missing genes (pFAMs) in *E. muscae*, none of which are enriched in relatives, suggesting this fungus is really different not by gene loss, but by its gene expansions.

Analyzing species-level boundaries and the data underlying those (genetic or morphological) is not something frequently presented in comparative genomic studies, however, here it is a welcome addition as the target species of the study is part of a species complex where morphology can be misleading and genetic data is infrequently collected in conjunction with the morphological data.

Weaknesses:

The conclusions of this paper are well supported, and I think the clarifications and improvements made to the manuscript in the revision process have greatly improved the paper.

<https://doi.org/10.7554/eLife.92863.2.sa0>

Author response:

The following is the authors' response to the original reviews.

Public Reviews:

Reviewer #1 (Public Review):

Summary:

The authors present a detailed study of a nearly complete Entomophthora muscae genome assembly and annotation, along with comparative analyses among related and non-related entomopathogenic fungi. The genome is one of the largest fungal genomes sequenced, and the authors document the proliferation and evolution of transposons and the presence/absence of related genetic machinery to explore how this may have occurred. There has also been an expansion in gene number, which appears to contain many "novel" genes unique to E. muscae. Functionally, the authors were interested in CAZymes, proteases, circadian clock related genes (due to entomopathogenicity/ host manipulation), other insect pathogenspecific genes, and secondary metabolites. There are many interesting findings including expansions in trahalases, unique insulinase, and another peptidase, and some evidence for RIP in Entomophthoralean fungi. The authors performed a separate study examining E. muscae species complex and related strains. Specifically, morphological traits were measured for strains and then compared to the 28S+ITSbased phylogeny, showing little informativeness of these morpho characters with high levels of overlap.

This work represents a big leap forward in the genomics of non-Dikarya fungi and large fungal genomes. Most of the gene homologs have been studied in species that diverged hundreds of millions of years ago, and therefore using standard comparative genomic approaches is not trivial and still relatively little is known. This paper provides many new hypotheses and potential avenues of research about fungal genome size expansion, entomopathogenesis in zygomycetes, and cellular functions like RIP and circadian mechanisms.

Strengths:

*There are many strengths to this study. It represents a massive amount of work and a very thorough functional analysis of the gene content in these fungi (which are largely unsequenced and definitely understudied). Too often comparative genomic work will focus on one aspect and leave the reader wondering about all the other ways genome(s) are unique or different from others. This study really dove in and explored the relevant aspects of the *E. muscae* genome.*

*The authors used both a priori and emergent properties to shape their analyses (by searching for specific genes of interest and by analyzing genes underrepresented, expanded, or unique to their chosen taxa), enabling a detailed review of the genomic architecture and content. Specifically, I'm impressed by the analysis of missing genes (pFAMs) in *E. muscae*, none of which are enriched in relatives, suggesting this fungus is really different not by gene loss, but by its gene expansions.*

Analyzing species-level boundaries and the data underlying those (genetic or morphological) is not something frequently presented in comparative genomic studies, however, here it is a welcome addition as the target species of the study is part of a species complex where morphology can be misleading and genetic data is infrequently collected in conjunction with the morphological data.

Thank you for your careful reading of our work. We're glad that you identified these areas as strengths.

Weaknesses:

The conclusions of this paper are mostly well supported by data, but a few points should be clarified.

*In the analysis of Orthogroups (OGs), the claim in the text is that *E. muscae* "has genes in multi-species OGs no more frequently than *Enotomophaga maimaiga*. (Fig. 3F)" I don't see that in 3F. But maybe I'm really missing something.*

Thank you for catching this. You were, in fact, not missing anything at all. There was a mismatch between the data plotted in F and G and how the caption described these data. We very much apologize for the confusion that this must have caused. We have corrected these plots and also made changes to improve interpretability (see below).

*Also related, based on what is written in the text of the OG section, I think portions of Figure 3G are incorrect/ duplicated. First, a general question, related to the first two portions of the graph. How do "Genes assigned to an OG" and "Genes not assigned to an OG" not equal 100% for each species? The graph as currently visualized does not show that. Then I think the bars in portion 3 "Genes in species-specific OG" are wrong (because in the text it says "*N. thromboides* had just 16.3%" species-specific OGs, but the graph clearly shows that bar at around 50%. I think portion 3 is just a duplicate of the bars in portion 4 - they look exactly the same - and in addition, as stated in the text portion 4*

"Potentially species-specific genes" should be the simple addition of the bars in portion 2 and portion 3 for each species.

As mentioned above, we sincerely regret the error made in the plot and for the confusion that this caused. F now reflects the percentage of orthogroups (OGs) that possess at least one representative from the indicated species (left) and the percentage of OGs that are species-specific (only possess genes from one species; right). The latter is a subset of the former. G now reflects the percentage of annotated genes that were assigned an OG, per species, as well as the inverse of this - genes that were not assigned to any OG. These should, and now do, sum to 100%. The "Within species-specific OG" data summed with the "Not assigned OG" data yields the "Potentially species-specific data" in the rightmost column.

In the introduction, there is a name for the phenomenon of "clinging to or biting the tops of plants," it's called summit disease. And just for some context for the readers, summit disease is well-documented in many of these taxa in the older literature, but it is often ignored in modern studies - even though it is a fascinating effect seen in many insect hosts, caused by many, many fungi, nematodes (!), etc. This phenomenon has evolved many times. Nice discussions of this in Evans 1989 and Roy et al. 2006 (both of whom cite much of the older literature).

You're right. We have now clarified that this behavior is called "summit disease" and referenced the suggested articles, along with a more recent review.

Reviewer #2 (Public Review):

*In their study, Stajich and co-authors present a new 1.03 Gb genome assembly for an isolate of the fungal insect parasite *Entomophthora muscae* (Entomophthoromycota phylum, isolated from *Drosophila hydei*). Many species of the Entomophthoromycota phylum are specialised insect pathogens with relatively large genomes for fungi, with interesting yet largely unexplored biology. The authors compare their new *E. muscae* assembly to those of other species in the Entomophthorales order and also more generally to other fungi. For that, they first focus on repetitive DNA (transposons) and show that Ty3 LTRs are highly abundant in the *E. muscae* genome and contribute to ~40% of the species' genome, a feature that is shared by closely related species in the Entomophthorales. Next, the authors describe the major differences in protein content between species in the genus, focusing on functional domains, namely protein families (pfam), carbohydrate-active enzymes, and peptidases. They highlight several protein families that are overrepresented/underrepresented in the *E. muscae* genome and other*

*Entomophthorales genomes. The authors also highlight differences in components of the circadian rhythm, which might be relevant to the biology of these insect-infecting fungi. To gain further insights into *E. muscae* specificities, the authors identify orthologous proteins among four Entomophthorales species. Consistently with a larger genome and protein set in *E. muscae*, they find that 21% of the 17,111 orthogroups are specific to the species. To finish, the authors examine the consistency between methods for species delineation in the genus using molecular (ITS + 28S) or morphological data (# of nuclei per conidia + conidia size) and highlight major incongruences between the two.*

Although most of the methods applied in the frame of this study are appropriate with the scripts made available, I believe there are some major discrepancies in the datasets that are compared which could undermine most of the results/conclusions. More precisely, most of the results are based on the comparison of protein family content between four Entomophthorales species. As the authors mention on page 5, genome (transcriptome) assembly and further annotation procedures can strongly influence gene discovery. Here, the authors re-annotated two assemblies using their own methods and recovered

between 30 and 60% more genes than in the original dataset, but if I understand it correctly, they perform all downstream comparative analyses using the original annotations. Given the focus on E. muscae and the small sample size (four genomes compared), I believe performing the comparisons on the newly annotated assemblies would be more rigorous for making any claim on gene family variation.

Thank you for this comment. While we did compare gene model predictions for two of these assemblies to assess if this difference could account for discrepancies in gene counts, completely reannotating all non-E. muscae datasets was outside of the scope of this study. In our opinion, the total number of predicted genes in a genome is not a best representation of differences since splitting or fusing gene models can inflate seeming differences; the orthology and domain counts are a more accurate assessment of the content. It's possible that annotation differences may have inflated some gene family counts, however we will note that similar domain trends were observed between the closest species to E. muscae, Entomophaga maimaiga, suggesting that these differences were not sufficient to prevent us from detecting real biological signals. We look forward to continued improvement of our genome through additional sequencing and more clarity on total gene content of E. muscae.

The authors also investigate the putative impact of repeat-induced point mutation on the architecture of the large Entomophthorales genomes (for three of the eight species in Figure 1) and report low RIP-like dinucleotide signatures despite the presence of RID1 (a gene involved in the RIP process in Neurospora crassa) and RNAi machinery. They base their analysis on the presence of specific PFAM domains across the proteome of the three Entomophthorales species. In the case of RID1, the authors searched for a DNA methyltransferase domain (PF00145), however other proteins than RID1 bear such functional domain (DNMT family) so that in the current analysis it is impossible to say if the authors are actually looking at RID1 homologs (probably not, RID1 is monophyletic to the Ascomycota I believe). Similar comments apply to the analysis of components of the RNAi machinery. A more reliable alternative to the PFAM analysis would be to work with full protein sequences in addition to the functional domains.

While we understand this concern regarding domain vs. full length protein, the advantage of the domain search is that HMM-based searches are sensitive to detecting more distantly related homologs. Entomophthoralean fungi are distantly related from the ascomycetes in which these mechanisms have been characterized, so we chose a broader search approach that may identify proteins with similar domain structure, but are not necessarily homologs. These searches are presented in the manuscript as preliminary, but worth further investigation. However, our RID-based analysis did not identify convincing homologs for RID1 in entomophthoralean fungi included in our investigation, and we reported low homology (i.e., 12-14%) among our orthogroup of interest and RID1. We have further edited this section to clarify our understanding that these candidates are not RID1 homologs. We had hoped to avoid this implication, but we felt this investigation and null result were worth reporting.

Recommendations for the authors:

Reviewer #1 (Recommendations For The Authors):

Specific points:

Results:

"1.03 Gb genome consisting of 7,810 contigs (N50 = 301.1 kb). Additional... resulted in a final contig count of 7,810 (N50 = 329.6 kb)" So you started and ended with the same contig count but a different N50? Is this a typo?

Yes, this was a typo. Thank you for bringing this to our attention.

Figure 1D.

The colors of Complete1x and Complete2x are too similar to tell them apart.

The colors have been made more distinct.

Figure 4B.

I know C. rosea has been found from insects before, but it's mostly a mycoparasite and occasionally an endophyte, and has bioactivity against a lot of things. I just saw that it's listed as an entomopathogen, and I was surprised. Anyway, leave it as is if you want to, but it's definitely better studied and better known (Google Scholar) as a mycoparasite.

Thanks for this comment. For the sake of including a more diverse representation of entomopathogenic fungi, we have opted to leave this as is.

Full references (from the public comment)

Evans, H.C., 1989. Mycopathogens of insects of epigeal and aerial habitats. Insect-fungus interactions, pp.205-238.

Roy, H.E., Steinkraus, D.C., Eilenberg, J., Hajek, A.E. and Pell, J.K., 2006. Bizarre interactions and endgames: entomopathogenic fungi and their arthropod hosts. Annu. Rev. Entomol., 51, pp.331-357.

Reviewer #2 (Recommendations For The Authors):

I believe the manuscript could largely benefit from restructuring the results section to enhance clarity. The results section reads like a lot of descriptive back and forth, so that the reader lacks a clear rationale. The absence of a consistent dataset used for the different comparisons made all along the manuscript makes it hard to follow.

Minor comments:

(No line numbers were available so I refer to page numbers).

p1

- not sure about the use of "allied" to describe other fungal species in the title and after (sister species?).*

We didn't want to use the word sister because not all of these species could be considered sister.

- Genomic defence against transposable elements rather than "anti"?*

We have rephrased to genomic defense.

p3

- Extra parenthesis at Bronski et al.*

This is now corrected.

- *What does newly-available mean here?*

We mean recent. A lot of the datasets we used were very new, and we wanted to emphasize that point.

- *The back and forth between genomes and transcriptomes makes it hard to follow, would clarify from the beginning (in addition to the sequencing method - short vs long-read assemblies as in Figure 1B) or perhaps use a consistent dataset for all subsequent comparative analysis in the Entomophthorales.*

We have denoted our transcriptomic datasets in Fig 1C using parentheses.

p5

- *Perhaps clarify that class II DNA transposons can also "copy" (single-strand excisions can be repaired by the host machinery).*

We have now included mention of “copy” as well as “jump” mechanisms of Class II transposons per your suggestion.

p6

- *"beginning roughly concurrently", not clear what "began".*

This is now corrected.

- *"control" rather than "protect against"?*

We’ve changed “protect against” to “counter”.

- *I believe RIP has only been observed (experimentally) in a handful of fungal species, all from the Ascomycota phylum.*

Hood et al, 2005 found signatures of RIP in anther-smut fungus and Horns et al, 2012, found evidence of hypermutability across repeat elements within several Pucciniales species.

- *"RID1 contains two DNA_methylase domains", RID1 has one methyltransferase domain according to the reference Freitag et al, 2002.*

Thank you for drawing this to our attention. It is true RID1 has one methyltransferase region; however, the sequence deposited by Freitag et al, 2002 (AAM27408) is predicted by HMMer to have two adjacent Pfam DNA_methylase domains (i.e., PF00145). In this exploratory analysis, we tried to leverage this characteristic to identify candidate proteins of interest. We have reworded this section to clarify this.

p8

- *Here and after I would use more informative titles for each paragraph.*

With the exception of the headings for Pfam, CAZy and MEROPs analyses, we believe the other headings are informative. We appreciate this comment, but opt to leave the heading titles as is.

- *I believe presenting the orthology analysis before the more in-depth protein family domain search.*

We leveraged the OG analysis mostly as a way to identify potentially unique genes in *E. muscae*, so we think the current order makes the most sense.

p10

- *Figures 3F and G are confusing. The legend for Figure 3F mentions "OGs with ≥ 2 species" while the figure shows "multi-species OGs", and reads as redundant with the "species-specific" OGs. For the "OGs within species" do I understand it correctly that it represents the number of genes assigned to OGs for each species? If yes, the numbers are in contradiction with Figure 3G. And in Figure 3G shouldn't the sum of "genes assigned in OGs" and "genes not assigned in OGs" add up to 100? I'm probably missing something here, but I would clarify what the different sets of orthogroups are in the figure and in the text (perhaps adopting a pangenome-like nomenclature).*

Thanks for this comment. This legend, unfortunately, reflected an earlier version of the figure and was overlooked prior to submission. We have since amended this and sincerely apologize for the error on our part.

p12

- *The whole first paragraph reads more like it should be part of an introduction/discussion.*

We've moved some of this paragraph to the discussion but left the background information necessary for the reader to understand why we were looking for homologs of *wc* and *frq*.

p13

- *The last paragraph reads like discussion.*

We have revised this paragraph so it now reads: "Because *E. muscae* is an obligate insect-pathogen only living inside live flies, we investigate the presence of canonical entomopathogenic enzymes in the genome. We find that *E. muscae* appear to have an expanded group of acid-trehalases compared to other entomopathogenic and non-entomopathogenic Entomophthorales (Fig. 4A), which correlates with the primary sugar in insect blood (hemolymph) being trehalose (Thompson, 2003). The obligate insectpathogenic lifestyle is also evident when comparing the repertoire of lipases, subtilisin-like serine proteases, trypsins, and chitinases in our focal species versus Zoopagomycota and Ascomycota fungi that are not obligate insect pathogens (Fig. 4B). Sordariomycetes within Ascomycota contains the other major transition to insect-pathogenicity within the kingdom Fungi (Araújo and Hughes, 2016). Based on our comparison of gene numbers, Entomophthorales possess more enzymes suitable for cuticle penetration than Sordariomycetes (Fig. 4B). In contrast, insect-pathogenic fungi within Hypocreales possess a

more diverse secondary metabolite biosynthesis machinery as evidenced by the absence of polyketide synthase (PKS) and indole pathways in Entomophthorales (Fig. 4C).”

p15 and 16

- *This all reads as redundant with the previous protein family domain analysis. I would try to merge them.*

Thank you for this comment, however we have opted to maintain the current structure.

p18

- *In the first sentence, I'm not sure about what was performed here.*

This has been reworded to clarify.

p20

- *Regarding the assembly, do I understand it correctly that a nuclear genome can be partially haploid / diploid?*

Thanks for your comment. The genome itself is, of course, some integer multiple of n, but based on BUSCO scores our assembly doesn't appear to have completely collapsed into a haploid genome. We think it makes more sense here to say “partially haploid” than “partially diploid” so have altered this.

p21

- *RIP has only been observed in a couple of Ascomycetes. RIP-like genomic signatures (GC bias) have been observed elsewhere.*

Hood et al, 2005 found signatures of RIP in anther-smut fungus and Horns et al, 2012, found evidence of hypermutability across repeat elements within several Pucciniales species.

p23

- *Interesting that the peptidase A2B domain is found uniquely in E. muscae genome and is associated with Ty3 activity. Does the domain often overlap with annotated Ty3 in E. muscae genome? Or how come the domain is not present in other sister species with large genomes full of Ty3 transposons? Could it relate to a new active transposon in E. muscae specifically?*

Thanks for this comment. The domain-based analysis was only performed on the predicted transcriptome of the genome assembly, which does not include the repeat elements (e.g., Ty3). It could be that this peptidase reflects a new active transposon that's specific to E. muscae, which would certainly be very interesting. We've now included this idea in the discussion.

p26

- *In the case of fungal genomes, I would not advise masking the assembly for repeated sequences prior to gene annotation (in particular given the current focus on protein family variation).*

Thank you for this comment, however we disagree with this assertion as a typical approach for genome annotation in fungi and eukaryotic genomes is to use soft masking of transposable elements before performing gene prediction to avoid over-prediction. While there could be alternative approaches that compare masked or unmasked. This is a recommended protocol for underlying tools like Augustus (10.1002/cpbi.57) and in general descriptions of genome annotation (10.1002/0471250953.bi0401s52). The false positive rate of genes predicted through TE regions is likely to be more a problem than false negatives of missed genes in our experience. Further it seems appropriate to use consistent approach to annotation throughout when including genomes from other sources (e.g., Joint Genome Institute annotated genomes) which also use a repeat masking approach first before annotation. It seems most appropriate to use consistent methods when generating datasets to be used for comparative analyses. It is outside the scope of this project to reannotate all genomes with and without repeat masking.

p27

- *Interrupted sentence at "Classification of DNA and LTR .. by similarity The".*

This was an unnecessary partial phrase as the information on classification of elements via RepBase was made a few sentences above this.

p28

- *Enriched/depleted rather than "significantly different"?*

Thank you for this comment, however we have opted to maintain the current phrasing.