Single Cell Sequencing Facilitates Genome-enabled Biology in Uncultured Fungi and Resolves Deep Branches on the Fungal Tree of Life

by

Kevin Riley Amses

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Ecology and Evolutionary Biology)
in The University of Michigan
2021

Doctoral Committee:

Professor Timothy James, Chair
Professor Gyorgyi Csankovszki
Associate Professor Stephen Smith
Professor Jianzhi Zhang

Kevin R. Amses

amsesk@umich.edu

ORCID ID: 0000-0002-4470-104X

Dedication

To both of my parents, Robert Amses and Lisa Piñero.

To my mother for sparking my interest in computers at a young age, thereby initiating the chain of events that would, decades later, bring me full circle back to writing code and running Linux daily. I thought I was here to study fungi.

To my father for teaching me that the secret to writing is not in waiting for profound sentences to flow from your mind out onto paper, but in going back later to take out the trash.

Thank you both for everything. The next 150 pages is for you.

Acknowledgements

There are so many individuals who deserve recognition for the integral roles they played in the completion of this dissertation work. The natures of these roles range from mentorship to technical assistance, to intellectual and emotional support, or to just being there to enjoy some time off.

I want to first acknowledge Dr. Timothy James, my faithful advisor through the work associated with this dissertation. Tim has been a lot of things to me: a boss, a collaborator, a friend, and exactly the critic I needed during some formative times over the past six years. This work would not have possible nor been anywhere near as fun or rewarding without his involvement. His unparalleled insight and support have elevated this dissertation beyond what it would have been without him. He has been an irreplaceable partner in the completion of this work, a partnership that I am going to miss dearly. Thank you.

I want to acknowledge Dr. Alisha Quandt, who was a postdoctoral scholar in the James Lab when I first arrived back in 2015, for the casual but vital mentorship in bioinformatics that she provided to me while we overlapped in Michigan. I arrived as a self-proclaimed organismal biologist and field mycologist but am leaving as a bioinformatician who studies fungi. Neither that transformation nor most of the work associated with this dissertation would have been possible without your support and guidance. You set me off on a trajectory in biology that I never knew that I wanted, but now cannot imagine my career without. Thank you.

I want to acknowledge Dr. William Davis, who was also a postdoctoral scholar in the James Lab who overlapped with me for a few years, for his partnership in finding, identifying, and plucking the spores of uncultured fungi off literal plates with dirt on them. He has taught and shown me

more than I ever wanted to know or see about early diverging fungi and the wild things they do to poor, unsuspecting animals and protozoans. He may never admit it, but squid would never have happened without him, and most of the rest of this dissertation would not have happened without squid, so thank you.

I also want to acknowledge Dr. Buck Castillo, my best friend and most trusted confidant throughout the work of this dissertation and the rest of life surrounding it. I do not know if either of us expected our initial introduction at the UM Biological Station to turn into the friendship that it did, but in looking back it is difficult to picture it any other way. He has been there for me through good times and bad, forced me to take time for myself, and always been willing to crack a beverage to sit and enjoy the moment. Even though we will be separated by the continental United States, I know we have not seen the last of each other. Thank you.

I want to acknowledge all members of the James Lab, past and present, for your time, support, and legacy. This dissertation would not have been possible or as exciting without all of you. Special thanks to Rebecca Clemons, for never seeming annoyed when I asked where something was in the lab for the third time. Another special thanks to Lucas Michelotti, for being a great friend – you were whisked away to Washington State, and then Georgia, far too soon.

I also want to acknowledge Rachel Cable, Dr. Jillian Myers, and Dr. Anat Belasen, as well as Corbin Kuntz, Shawn Colborn, Peter Cerda, Jon Massey, and so many others for their comradery throughout the work of this dissertation. Suffice it to say, it has been a blast.

I want to acknowledge the members of my doctoral committee for your insight and support over the years. Thank you, Dr. Timothy James, Dr. Jianzhi Zhang, Dr. Gyorgyi Csankovszki, and, most recently, Dr. Stephen Smith. Thank you also to Dr. Deborah Goldberg, a previous member on my committee.

I would also like to acknowledge my funding sources, without which this dissertation work would not have been possible. I want to extend a special thanks to the NIH-funded Michigan Predoctoral Training in Genetics fellowship [T32GM007544], which funded me for years 2–3 of

this dissertation. In addition to support in the form of funding, thank you for providing an interdisciplinary environment within which to discuss important topics in biological research. To the administrators and all the friends that I made there, thank you.

Funding sources for individual chapters of this dissertation work follow.

Chapter 2: This work was supported by the National Science Foundation [DEB1441677, DEB1441604] and by the following National Institutes of
Health training grant: Michigan Predoctoral Training in Genetics [T32GM007544].

Chapter 3: This project was funded through UM LSA Department of Ecology and Evolutionary Biology BLOCK Grant awards awarded between 2016 and 2020.

Chapter 4: This project was funded in part by the National Science Foundation, grants DBI-1756202, DBI-1910720, and DEB-1929738. We thank Rebecca Clemons for technical assistance.

Chapter 5: This project was funded through UM LSA Department of Ecology and Evolutionary Biology BLOCK Grant awards awarded between 2016 and 2020.

Table of Contents

List of Tables

List of Figures

List Appendices

Abstract

Microbial life on Earth is the most diverse life on Earth. The magnitude of microbial diversity is obscured by their small statures, relatively short list of defining morphological characteristics, and general recalcitrance to being separated from nature and brought into the laboratory. Most microbes cannot be grown under axenic conditions (i.e., uncultured), a simple reality that impedes their discovery in complex natural systems and downstream studies to understand their biology. A point no less important in the age of genome-enabled biological research, the uncultured status of most microbes precludes sequencing of their genomes via conventional high-throughput sequencing, which requires ample input material. Single cell sequencing offers a viable workaround to this central obstacle by enabling the amplification of genomic DNA from individual cells up to amounts more than sufficient for sequencing. That said, this workaround introduces biases to sequence composition and exacerbates contamination, both of which present unique challenges to downstream genome-scale analyses. Fungi constitute a diverse lineage of heterotrophic eukaryotes that sometimes blur the line between microbial and macroscopic life. Our understanding of fungi is wildly incomplete and biased toward fungi that produce macroscopic forms or those that can be grown under axenic conditions. Even in the age of genome-enabled biological research, most fungi that are microscopic, uncultured, or especially both remain poorly understood. In this dissertation, I use single cell sequencing, sometimes combined with conventional genome sequencing, to address this gap by conducting genome-enabled biological research in uncultured or under-sampled sectors of the fungal tree of life. In Chapter 2, I design and deploy a novel computational approach to filtering the biased and often contaminated sequence data associated with single cell sequencing. I demonstrate its ability to outperform available filtering approaches using genuine and mock datasets. In Chapter 3, I use single cell sequencing of predatory fungi to discover novel endohyphal bacteria colonizing fungi in a phylum where this type of symbiosis was entirely unknown. Genome-scale phylogenetic analyses implicate recent interphylum host switches for bacteria thought to transmit predominantly vertically. The novel bacterial endosymbionts discovered have similar genomes to other endohyphal bacteria but have, in some cases, acquired and retained horizontally transferred genes from animals. In Chapter 4, I use genome-scale data to infer a robustly supported phylogeny of zoosporic fungi. Mapping of genetic traits and ploidy inferred from sequence data suggests that fungal evolution has been driven by gradual loss and that most early diverging lineages have diploid-dominant life cycles. In Chapter 5, I use genome-scale data to resolve a disagreement between classical taxonomy and molecular phylogenetics revolving around the phylogenetic placement of the enigmatic, arthropod-mummifying fungal genus *Neozygites*. Through the development of novel computational methods, genome-scale phylogenetics, and a comparative approach, this dissertation demonstrates the utility of single cell sequencing in closing vast gaps in our understanding of fungi.

Chapter 1: Introduction

1.1. Microbes are more diverse than any of group life of Earth, but this is underappreciated on a broad scale.

Life is complicated. Life is also diverse, and living things differ in how they look, what they do, and where they came from. Microbes constitute the most diverse, abundant, and ancient pool of biodiversity on Earth. They entirely dominate two of the three domains of life and compose the majority of the third, leaving but a small sector of the tree of life to account for the macroscopic multicellular lineages that societal consciousness regards as staggering examples of the diversity of life on Earth. While the diversity of forms of macroscopic life (e.g., animals, plants, etc.) is certainly awe-inspiring, the presumption that they are more diverse than microbial life is plain wrong. Estimates based on mathematical scaling models predict upwards of 1 trillion microbial species on Earth (Locey and Lennon, 2016). Locations like the Great Barrier Reef with its ~3,000 animal species, among other human-declared "hot spots" of biodiversity, pale in comparison and account for an exceedingly small portion of the known and estimated diversity of life on Earth. Why are microbes so underappreciated in a biosphere they dominate?

Microbes are inherently cryptic. Their microscopic stature makes them hard or impossible for humans to visualize without the assistance of technological advances like the microscope. Even under magnification, their stature leaves little space for recognizable morphological characters to present. This constraint breeds the false presumption that microbes are less diverse than their macroscopic relatives. Connections between form, function, and history are often more intuitive through an anthropocentric lens; life forms on our scale make sense to us because we can see pieces of ourselves. For example, the similarities between certain leg bones of terrestrial mammals and the homologous vestigial structures of whales is an intuitive case of structural homology, and in fact true evolutionary homology, that is obvious at human scales (Andrews and Others, 1921). These connections between form and function are rarely as clear in microbial life, and even when they are, they are not always indicative of shared history. Take for example rampant convergent evolution on fruit body form in mushroom-forming fungi that, despite having similar form and function, have emerged independently many times (Binder et al., 2005; Hibbett, 2007; Hibbett et al., 1997). Microbes are several times removed from the morphological selective landscapes that define our view of biodiversity, despite being much more diverse. This has been known in microbial research since its inception and, as such, progress in microbial research has been driven by technological advances that bring microbes onto the plane of human perception.

1.2 Advances in microbial research are driven by methodological innovations that bring microbes closer to the realm of human perception.

The history of microbial research is tightly tied to methodological advances that made such research feasible. To study a microbe it must be physically, or at least conceptually, separated from the complex abiotic and biotic contexts in which it exists. Bacteria were first discovered in the mid-17th century as light microscopes achieved sufficient resolving power (Porter, 1976). Subsequent advances in microscope technology to increase resolving power have and continue to drive microbial research further. Physical separation and cultivation of microbes from nature was not achieved until about two centuries after the discovery of bacteria when, in the mid-19th century, the first successful artificial growth media were formulated (Bonnet et al., 2020). Since then, stepwise refinements and diversification of media recipes have enabled more microbes to be cultivated under axenic conditions. The ability to maintain microbial growth in the vacuum of pure culture paved the way for subsequent experiments to understand the functions of individual members of complex natural systems, from soils to the human body (Novick and Szilard, 1950; van Niel, 1944). We used this information on microbial form and function to design a classification system for them (e.g., Bergey's Manual of Determinative Bacteriology) and add them to our ever-growing concept of the tree of life.

The complementary methodological advances of DNA amplification by polymerase chain reaction (PCR), dideoxynucleotide-based DNA sequencing (i.e., Sanger sequencing), and molecular phylogenetics of the late 20th century provided the first insights into the shortcomings of form and function in accurately describing the evolutionary history of life on Earth (Pace, 1997; Siefert't and Fox, 1998). They allowed us to infer evolutionary history based on DNA sequence characters (e.g., ribosomal DNA), instead of form and function alone. Phylogenetic trees based on these markers have contributed to massive reorganizations of the tree of life, microbes included (Hibbett et al., 2007; Pace, 1997). In general, the utility of short sequences and single markers degrades as the root-to-tip distance of phylogenetic trees increases. To resolve these deep nodes on the tree of life, we needed larger, genome-scale, sequences (Spatafora et al., 2016). Whole genome sequencing was conducted by Sanger sequencing for a few microbes but took years and international teams of scientists (Dujon, 1996; Goffeau et al., 1996).

The subsequent innovation of high-throughput sequencing significantly expedited the path to sequencing whole genomes and, like other advances before it, opened a new frontier in microbial research (Ronaghi et al., 1996; Shendure and Ji, 2008; Tucker et al., 2009). Genome-scale sequence datasets contain an abundance of information with which to better understand the forms, functions, and histories of microbes. To formulate the strongest hypotheses about the evolutionary history of microbes to date, genome-scale sets of phylogenetically informative markers can be compiled from diverse microbes and used to resolve deep nodes in the tree of life (Davis et al., 2019; Parks et al., 2018; Spatafora et al., 2016). Conclusions can be drawn about microbial form and function from analyzing and comparing genome-scale sequence data of diverse microbes (Brun and Silar, 2010; Kohler et al., 2015; Martin et al., 2008). Genome-scale data is vast and complicated but offers unique insight into the lives of microbes. Unfortunately, this kind of insight is not possible for most microbes simply because we do not know how to cultivate them (Amann et al., 1995; Streit and Schmitz, 2004).

1.3 Uncultured microbes remain cryptic in the age of genome-enabled biology because of the importance of axenic cultures in generating genome-scale data.

Uncultured microbes are those that cannot be cultivated under axenic conditions. Whether our inability to cultivate them stems from poorly optimized culture media or growth conditions, the absence of required symbiotic partners in the axenic vacuum, or some other incompatibility, is unclear. Conventional whole genome sequencing of microbes requires the acquisition of sufficient amounts of clean DNA, which is only possible from axenic cultures or macroscopic forms (e.g., fungal fruiting bodies). This central obstacle effectively excludes uncultured microbes from the benefits of genome-enabled biological research.

There are some high-throughput sequencing approaches that circumvent this obstacle, but they pose other obstacles. First, amplicon-based sequencing of complex natural samples can sequence many short pieces of genomic DNA (Claesson et al., 2010; Lazarevic et al., 2009). Amplicon-based sequencing has the potential to detect uncultured microbes in nature, place them in the tree of life, and perhaps hint to their function based on known functions of closely related microbes (Nguyen et al., 2016). However, amplicon-based sequencing introduces bias based on the choice of primers, which are generated from known sequences and can easily select against microbial groups that are poorly known, as are most uncultured microbes (Makiola et al., 2018; Tedersoo et al., 2015; Zhou et al., 2011). More importantly, amplicon-based sequencing cannot generate draft genomes, a necessary starting point for genome-enabled biological research. Second, metagenomics can be used to sequence the entirety of genomic DNA in bulk samples. This is a less biased and more thorough approach, but the resulting data is extremely complex. Metagenome complexity can be algorithmically simplified via segregation into bins that, in theory, each represent one genome from a complex community (Sedlar et al., 2017). However, in practice, bins are likely to contain at least some metagenomes composed of closely related microbes (Yue et al., 2020). This complicates the process of inferring the functions or evolutionary histories of individual microbial species from metagenomes. Further, the complexity of metagenomes necessitates sequencing at extremely high depths to yield genome-scale data for every microbe present in the community, which practically means that most metagenomic datasets are biased against rare microbes (Nelson et al., 2020). Metagenomics does provide an attractive option for reconstructing evolutionary relationships among microbes and the functional potential of complex communities but suffers from reduced resolving power at the species level and requires sequencing depths that can be extraneous. Despite these obstacles and biases, both of these approaches have rapidly increased the detection rate of novel microbes beyond what was possible through axenic culturing and observation, which completely overlooks uncultured microbes (Dick et al., 2009; Li et al., 2016; Tedersoo et al., 2014).

1.4 Single cell genomics facilitates genome-enabled biology in uncultured microbes without sacrificing species-level resolution.

Single-cell genomics (SCG) is a sequencing approach that nonspecifically amplifies genomic DNA from individual cells up to amounts more than sufficient for high-throughput sequencing (Kalisky and Quake, 2011). In the last decade, SCG has gained popularity in multicellular model systems where it can capture cell-to-cell heterogeneity in DNA complement and gene expression, but can also be applied to generate genome-scale sequence data for uncultured microbes (Davis et al., 2019; Kimmerling et al., 2016; Mikhailov et al., 2016; Roy et al., 2014). Cells are lysed under alkaline conditions, the lysate is neutralized, and then DNA is amplified by isothermal multiple displacement amplification (MDA) for 6–8 hours at 30C. The MDA reaction is catalyzed by the DNA polymerase from bacteriophage phi29 and primed by fully factorial sets of DNA hexamer primers (Lovmar and Syvänen, 2006).

SCG enables the sequencing of uncultured microbial genomes because it reduces the threshold for cellular inputs down to individual cells, which can be collected directly from nature without the need for axenic cultivation. Individual cells can be separated from complex samples in a variety of ways, ranging from fluorescent activated cell sorting (FACS) to manual isolation from bulk samples or *in vitro* microcosms (Ahrendt et al., 2018; Davis et al., 2019; Rinke et al., 2014). Although uncultured microbes cannot be cultivated under axenic conditions, they can often be quasi-cultivated in highly mixed *in vitro* microcosms established by depositing small portions of bulk samples onto low nutrient agar. These quasi-cultivation methods have a rich history in biology where they enabled initial descriptions of uncultured microbial species (Davis et al., 2019; Drechsler, 1959; Saxena, 2008; Whisler and Travland, 1974).

Although SCG enables genome sequencing of uncultured microbes, it introduces unique biases derived from the MDA reaction. First, despite the use of fully factorial hexamer primers, amplification bias can be introduced based on the hexamer content and GC content of genomic DNA (Pinard et al., 2006). Second, stochasticity in amplification start positions during the early stages of amplification lead to unequal coverage of the genome in terminal MDA products (e.g., 10^0x – 10^4x coverage); regions amplified early are amplified to higher magnitudes, and vice versa (Davis et al., 2019; Pinard et al., 2006). Both complicate assembly and lead to draft genomes that are often fragmented and incomplete (Davis et al., 2019; Mikhailov et al., 2016; Roy et al., 2014). Finally, SCG is already prone to contamination because of the complex nature of cell sources (e.g., *in vitro* microcosms). MDA exacerbates any natural contamination present in reaction tubes in addition to introducing contamination derived from the reagents it requires (Davis et al., 2019; Rinke et al., 2014). This makes most SCG assemblies mildly to moderately metagenomic (Davis et al., 2019; Mikhailov et al., 2016).

Since most SCG assemblies are metagenomic, they need to be filtered prior to their use in downstream genome-scale analyses. The retention of contamination-derived sequences can easily lead to misrepresentations of biology. Metagenomic binning algorithms that segregate metagenomic assemblies into bins independent of taxonomy are diverse (Dick et al., 2009; Kang et al., 2019; Kumar et al., 2013; Laczny et al., 2015; Sedlar et al., 2017; Sieber et al., 2018; Wu et al., 2016). Although these algorithms are designed to process metagenomes with levels of complexity that dwarf typical SCG metagenomes, they are poorly optimized to cope with the unique biases that characterize SCG metagenomes. Many lean heavily on coverage to drive binning, a metric that is much less informative for binning SCG metagenomes because of their broad coverage distributions (Laczny et al., 2015; Pinard et al., 2006; Sieber et al., 2018). Because there are no metagenomic binning algorithms designed to specifically address these biases, studies that use SCG to generate genome-scale data for uncultured microbes rely heavily on manual filtration (Gawryluk et al., 2016; Mikhailov et al., 2016).

1.5 Fungi constitute an ancient and diverse lineage of eukaryotic microbes with a high density of "dark matter."

The Kingdom Fungi is a lineage of heterotrophic eukaryotes that diverged from other eukaryotes at least one billion years ago. It is comprised of a diverse assemblage of organisms that assimilate energy and nutrients from their environments in many ways, from parasitic or mutualistic interactions with other organisms to the decay of dead organic matter. In terms of scale, fungi blur the line between microbial and macroscopic life, ranging from unicellular forms (e.g., yeasts) to multicellular webs of cells (i.e., hyphal forms or mycelia) that can grow to span hundreds of hectares over the course of thousands of years (Smith et al., 1992). Further blurring the micro–macro divide, many fungi produce fruiting bodies that are unequivocally macroscopic (e.g., mushrooms). The kingdom is currently divided into 12 phyla with macroscopic forms dominating in the latest-diverging Basidiomycota and Ascomycota, and nearly entirely microscopic forms dominating in all earlier-diverging phyla, with some exceptions (**Figure 1.1**) (James et al., 2020). There are approximately 120,000 species of fungi described to science with estimates of actual fungal diversity suggesting 2-4 million extant species (Hawksworth and Lücking, 2017). If these estimates are accurate, the fungal kingdom is home to a significant amount of "dark matter" fungi, or fungi that are poorly known, if at all. Not knowing about this fungal dark matter precludes our understanding of the kingdom as a whole, including its evolutionary history, the full ecological or functional potential it harbors, and the diversity of its interactions with the rest of the tree of life. Based on the paucity of described species relative to later-diverging clades, it is clear that a significant portion of this fungal dark matter is situated in early-diverging lineages at the base of the fungal tree of life (James et al., 2020).

1.6 Genome-scale sets of phylogenetic markers are required to resolve early diverging fungal lineages.

The fungal tree of life underwent sweeping reorganizations as sequencing technology enabled the use of DNA sequence characters in phylogenetic reconstruction (Dornburg et al., 2017; Spatafora et al., 2016). Like other microbes, fungi lack many of the macroscopic characteristics that have assisted the morphological categorization of plants and animals for centuries. Many of the macro- and micro-morphological characteristics that have historically been available for fungal taxonomy have proven to be poor indicators of shared evolutionary history (e.g., fruiting bodies) (Binder et al., 2005; Hibbett, 2007; Hibbett et al., 1997). In general, rDNA and other markers have been able to resolve well-supported phylogenetic relationships in later-diverging lineages but performed poorly in early-diverging lineages with longer times since divergence (White et al., 2006). In order to resolve these branches of the fungal tree of life, genome-scale sets of markers (i.e., phylogenomics) have quickly become a requirement. Over the past decade, fueled by advances in high-throughput sequencing, increased genome sequencing in these early-diverging lineages has taken major strides toward resolving these foggy parts of the fungal tree of life (Davis et al., 2019; Dornburg et al., 2017; Spatafora et al., 2016).

Accurate inference of evolutionary relationships with molecular phylogenetics has always been dependent on the use of sequences that are conserved across the taxon set (e.g., rDNA). Phylogenomics is no different, which requires orthologous sequences of conserved genes. True orthology is important to avoid the conflicting signal of paralogous sequences that, despite sequence similarity, have different evolutionary histories (Dornburg et al., 2019, 2017; Li et al., 2021). In general practice, sequences are extracted from genome-scale datasets based on matching a sequence model (often a Hidden Markov Model, or HMM) constructed based on sequences in large sequence databases (e.g., NCBI GenBank) (Davis et al., 2019; Li et al., 2021; Spatafora et al., 2016). Due to the overrepresentation of well sampled lineages in these databases, available sets of phylogenomic marker HMMs for fungi are inherently biased toward later-diverging lineages. Their use in phylogenomics of early-diverging lineages, which are often quite divergent, is more likely to extract paralogous or spurious sequences. If care is not taken to remove them, these paralogs insert erroneous phylogenetic signal into phylogenies, confuse relationships between taxa, and potentially produce robustly supported, but incorrect, topologies (James et al., 2020; Prasanna et al., 2019). Ensuring the exclusion of paralogs, model compatibility, and appropriate selection of substitution models, among other concerns, are important to phylogenomic analyses in any sector of the tree of life, but they are especially critical in poorly sampled sectors, such as early diverging fungal lineages (Prasanna et al., 2019).

1.7 Fungi engage in diverse symbiotic interactions with other members of the tree of life, the known diversity of which is constantly expanding.

Fungi do not exist in a vacuum. They are members of complex communities of coexisting organisms that span the tree of life. Endohyphal bacteria (EHB) that colonize the cytosol of fungal cells coexist with fungi in a particularly intimate context (Pawlowska et al., 2018; Torres-Cortés et al., 2015). EHB can be categorized into three major classes (Araldi-brondolo et al., 2017). Class 1 EHB encompasses Mollicutes-related EHB, or MRE, that colonize the cells of plant-associated fungi in the Mucoromycota (e.g., arbuscular mycorrhizal fungi). MRE form a monophyletic clade in the Mollicutes (Naito et al., 2017; Torres-Cortés et al., 2015). MRE are characterized by highly reduced (~400 kbp) genomes, dramatic inter-host genome diversity driven, in part, by horizontal gene transfer (HGT), and predominant vertical transmission between hosts, with some exceptions (Araldi-brondolo et al., 2017; Toomer et al., 2015). As is suggested by their highly reduced genomes, MRE are metabolically dependent on their hosts. The effects of MRE colonization on their host is unclear, and their small, mosaic genomes and uncultured status complicate their demystification (Araldi-brondolo et al., 2017). Class 2 EHB encompasses the *Burkholderia*-related endohyphal bacteria, or BRE, that colonize fungi in the Mucoromycota and Ascomycota (Araldi-brondolo et al., 2017). BRE form a paraphyletic assemblage in the Burkholderiaceae, composed of two monophyletic lineages (Guo et al., 2020). BRE genomes (~1–4 Mbp) are larger than MRE genomes but are reduced relative to their free-living relatives. Like MRE, BRE are dependent on their hosts for basic metabolism and transmission is thought to be predominantly vertical, with some rare examples of horizontal transmission between closely-related hosts (Araldi-brondolo et al., 2017; Mondo et al., 2012). The effects of BRE colonization on their hosts lean toward mutualism where BRE-harboring fungi can receive fitness benefits in bipartite or tripartite symbiotic interactions (e.g., with plants) (Partida-Martinez and Hertweck, 2005). That said, these effects can be context dependent (Araldi-brondolo et al., 2017). Class 3 EHB is a “grab bag” class of facultative opportunists that colonize diverse lineages of fungi where they cause transient infections that are usually detrimental to the host. They do not share a set of unifying traits like MRE and BRE, and are not unified by shared evolutionary history (Araldi-brondolo et al., 2017).

Hosts of obligate EHB (i.e., MRE and BRE) are concentrated in the Mucoromycota, a group of early-diverging fungi. In recent years, Mucoromycota has received significant attention in the form of genome sequencing efforts, which have fueled the detection and characterization of novel EHB (Bianciotto et al., 2003; Torres-Cortés et al., 2015). It is not clear why obligate EHB predominantly associate with fungi in the Mucoromycota, but it could be that Mucoromycota plant associations provide unique niche space within which these bacteria have established (Pawlowska et al., 2018). That said, few members of earlier diverging lineages, where EHB are entirely unknown, have had their genomes sequenced or been screened for EHB. The Zoopagomycotina is an earlier diverging lineage composed of predatory and parasitic fungi that tend to associate with animals, protozoans, or other fungi (Davis et al., 2019; Drechsler, 1959; Spatafora et al., 2016). Despite dramatic differences in ecological strategy relative to the Mucoromycota, these two lineages share traits that could have facilitated the establishment and persistence of EHB colonization, including the absence of regular septa, the hyphal growth form, and use of terrestrial soil habitats. Detection and characterization of EHB associated with fungi in the Zoopagomycotina would dramatically expand the concept of EHB and require a reframing of their function inside the cells of fungal hosts.

1.8 Summary of Dissertation Chapters

My dissertation is divided into four chapters that use SCG, and some conventional whole genome sequencing, to conduct genome-enabled biological research in uncultured and under sampled fungi. Through a combination of computational method development, genome-scale phylogenetics, and genomics, my dissertation investigates the biology of these fungi in the following contexts. Chapter 2 simplifies the path to genome-enabled biology in uncultured fungi by introducing a novel computational approach to filtering SCG metagenomes that yields high fidelity draft genome sequences *in silico*. Chapter 3 uses SCG to discover and characterize novel bacterial endosymbionts of uncultured fungi in the Zoopagomycota. Chapter 4 uses traditional whole genome sequencing and SCG to resolve the evolutionary relationships and nuclear states of an early-diverging group of zoosporic fungi with genome-scale data. Finally, in Chapter 5 I use SCG and some whole genome sequencing and to resolve a conflict between classical taxonomy and modern phylogenetics with genome-scale phylogenetic analyses.

Chapter 2: SCGid, a consensus approach to contig filtering and genome prediction from single-cell sequencing libraries of uncultured eukaryotes.

The genomes of uncultured fungi are difficult or impossible to sequence via conventional whole genome sequencing because of the difficulty in acquiring sufficient input material. SCG provides a viable workaround, but the resulting data is imbued with unique biases derived from MDA. These biases include amplification biases that disrupt equal sequencing depth of template molecules and exacerbate contamination. They complicate *de novo* assembly and downstream filtering (Davis et al., 2019; Pinard et al., 2006). Although methods available for the filtering of metagenomes are abundant, none of them are designed to address the unique biases posed by SCG (Kumar et al., 2013; Laczny et al., 2015; Sedlar et al., 2017; Sieber et al., 2018). In this chapter, I address this gap in the bioinformatics toolkit by developing and benchmarking a computational tool designed specifically for filtering SCG metagenomes.

SCGid is an automated software tool that filters SCG metagenome assemblies based on multiple lines of sequence-based evidence. Through union of its SCG-optimized implementations of existing filtering approaches, *SCGid* brings consensus-based reasoning to SCG metagenome filtering and yields high-fidelity draft genomes primed for downstream genome-enabled analyses. I test *SCGid* on mock and genuine SCG datasets to demonstrate its broad utility in the study of uncultured eukaryotic microbes from across the tree of life.

Chapter 3: Novel obligate endohyphal bacterial symbionts of uncultured predatory fungi revealed by single cell sequencing implicate recent interphylum host switches.

Endohyphal bacterial symbionts (EHB) of fungi are becoming increasingly appreciated. Obligate EHB are phylogenetically restricted to two major bacterial lineages within which they form one or two monophyletic clades, MRE and BRE, respectively, members of which predominantly associate with plant-associated fungi in the Mucoromycota (Araldi-brondolo et al., 2017; Guo et al., 2020; Naito et al., 2017; Pawlowska et al., 2018). Transmission between hosts appears to be predominantly vertical, with some signatures of horizontal transmission between closely related hosts (Mondo et al., 2012; Toomer et al., 2015). The constraints that explain this host range are not clear, but there is evidence that suggests their involvement in plant-fungal interactions (Araldi-brondolo et al., 2017; Partida-Martinez and Hertweck, 2005; Pawlowska et al., 2018). In this chapter I discover and characterize novel MRE and BRE in association fungi in the Zoopagomycota, a phylum in which obligate bacterial endosymbioses have never been detected.

I use SCG to sequence the nearly complete genomes of two novel EHB that colonize animal-associated fungi in the Zoopagomycota. Phylogenomic analyses resolve them as derived members of Mucoromycota-associated lineages of EHB, suggesting that interphylum host switches have occurred in the history of MRE and BRE. Discovery of these EHB in the Zoopagomycota disrupts the concept of obligate EHB as endosymbionts of plant-associated fungi that prevails in the literature, requiring a broadening to include the parasitic, predatory, and plant-indifferent fungi that dominate therein.

Chapter 4: Phylogenomic analysis of zoosporic true fungi suggests most early diverging lineages have diploid-dominant life cycles.

The majority of fungal species diversity is known from only two (i.e., Ascomycota and Basidiomycota, or Dikarya) of the twelve currently recognized phyla (James et al., 2020). This asymmetric sampling of the kingdom influences our view of non-Dikarya fungi and leads to presumptions about their biological characteristics (e.g., aerial spores and haploid-dominant life cycles). It is becoming increasingly appreciated that the evolutionary history of fungi is much deeper than Dikarya, but our understanding of the evolutionary histories and biological characteristics of these early-diverging lineages is incomplete (James et al., 2020; Tedersoo et al., 2018). In this chapter, I take steps toward expanding our understanding of the evolutionary histories and characteristics of the fungal tree of life by focusing on some of its less studied branches.

In collaboration with an interinstitutional team of scientists, I sequence 68 new genomes of early diverging zoosporic fungi and use phylogenomics to infer robustly supported hypotheses about their evolutionary histories. Using careful automated and manual gene tree filtering approaches, I assess the compatibility of each member of a Dikarya-centric phylogenomic marker set to ensure the exclusion of paralogous sequences in phylogenetic reconstructions. Using reimagined computational approaches, I infer the nuclear state of fungi across our 137-taxon dataset *in silico* to investigate the evolution of ploidy in fungi.

Chapter 5: Single cell sequencing and phylogenomics places the enigmatic arthropod-mummifying fungus Neozygites as a distinct lineage in Entomophthorales.

Neozygites is a genus of entomopathogenic fungi that parasitize mites, aphids, and other arthropods (Delalibera and Hajek, 2004; Yaninek et al., 2002). *Neozygites* causes punctuated epizootic outbreaks that can devastate local host populations. Tightly tied to environmental conditions such as temperature and relative humidity, outbreaks are rapidly amplified under optimal conditions and rapidly dissipate following shifts to suboptimal conditions (Delalibera and Hajek, 2004; Steinkraus et al., 2002; Wekesa et al., 2007). Based on characteristic morphological and life history traits, *Neozygites* was taxonomically placed in a its own family in a lineage of other entomopathogens (Butt and Heath, 1988; Butt and Humber, 1989; Keller, 1997). This placement was called into question when molecular phylogenetics based on rDNA markers placed it in a different lineage of fungi that does not parasitize arthropods nor share these characteristic traits (White et al., 2006). Although the long branches *Neozygites* occurred on in these phylogenetic reconstructions has always suggested this new placement could be artifactual, this conflict between classical taxonomy and molecular phylogenetics has yet to be resolved. In this chapter, I resolve this conflict by generating four draft genomes for three species of *Neozygites* and resolving its placement with genome-scale sets of phylogenetic markers.

I use these phylogenomic analyses not only to resolve the placement of *Neozygites* but also to assess its stability toward identifying the source of the signal supporting artifactual placements. To do so, I inferred phylogenomic trees using different sets of markers, tested the impact of reduced taxon sampling within the *Neozygites* clade, and inferred phylogenies based on separate sets of sites with slow and fast relative evolutionary rates.

1.9 Literature Cited

Ahrendt, S.R., Quandt, C.A., Ciobanu, D., Clum, A., Salamov, A., Andreopoulos, B., Cheng, J.-F., Woyke, T., Pelin, A., Henrissat, B., Reynolds, N.K., Benny, G.L., Smith, M.E., James, T.Y., Grigoriev, I.V., 2018. Leveraging single-cell genomics to expand the fungal tree of life. Nature Microbiology 3. https://doi.org/10.1038/s41564-018-0261-0

Amann, R.I., Ludwig, W., Schleifer, K.H., 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. 59, 143–169.

Andrews, R.C., Others, 1921. A remarkable case of external hind limbs in a humpback whale. American Museum novitates; no. 9.

Araldi-brondolo, S.J., Spraker, J., Shaffer, J.P., Woytenko, E.H., Baltrus, D.A., Gallery, R.E., Arnold, A.E., 2017. Bacterial Endosymbionts: Master Modulators of Fungal Phenotypes. The Fungal Kingdom 981–1004.

Bianciotto, V., Lumini, E., Bonfante, P., Vandamme, P., 2003. "Candidatus Glomeribacter gigasporarum" gen. nov., sp. nov., an endosymbiont of arbuscular mycorrhizal fungi. Int. J. Syst. Evol. Microbiol. 53, 121–124.

Binder, M., Hibbett, D.S., Larsson, K., Larsson, E., Langer, E., Langer, G., 2005. The phylogenetic distribution of resupinate forms across the major clades of mushroom-forming fungi (Homobasidiomycetes). System. Biodivers. 3, 113–157.

Bonnet, M., Lagier, J.C., Raoult, D., Khelaifia, S., 2020. Bacterial culture through selective and non-selective conditions: the evolution of culture media in clinical microbiology. New Microbes New Infect 34, 100622.

Brun, S., Silar, P., 2010. Convergent Evolution of Morphogenetic Processes in Fungi, in: Pontarotti, P. (Ed.), Evolutionary Biology – Concepts, Molecular and Morphological Evolution: 13th Meeting 2009. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 317–328.

Butt, T.M., Heath, I.B., 1988. The changing distribution of actin and nuclear behavior during the cell cycle of the mite-pathogenic fungus Neozygites sp. Eur. J. Cell Biol. 46, 499–505.

Butt, T.M., Humber, R.A., 1989. An immunofluorescence study of mitosis in a mite-pathogen,Neozygites sp. (Zygomycetes: Entomophthorales). Protoplasma 151, 115–123.

Claesson, M.J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J.R., Ross, R.P., O'Toole, P.W., 2010. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. Nucleic Acids Res. 38, e200.

Davis, W.J., Amses, K.R., Benny, G.L., Carter-house, D., Chang, Y., Grigoriev, I., Smith, M.E., Spatafora, J.W., Stajich, J.E., James, T.Y., 2019. Genome-scale phylogenetics reveals a monophyletic Zoopagales (Zoopagomycota , Fungi). Mol. Phylogenet. Evol. 133, 152–163.

Delalibera, I., Jr, Hajek, A.E., 2004. Pathogenicity and specificity of Neozygites tanajoae and Neozygites floridana (Zygomycetes: Entomophthorales) isolates pathogenic to the cassava green mite. Biol. Control 30, 608–616.

Dick, G.J., Andersson, A.F., Baker, B.J., Simmons, S.L., Thomas, B.C., Yelton, A.P., Banfield, J.F., 2009. Community-wide analysis of microbial genome sequence signatures. Genome Biol. 10. https://doi.org/10.1186/gb-2009-10-8-r85

Dornburg, A., Su, Z., Townsend, J.P., 2019. Optimal Rates for Phylogenetic Inference and Experimental Design in the Era of Genome-Scale Data Sets. Systematic Biology. https://doi.org/10.1093/sysbio/syy047

Dornburg, A., Townsend, J.P., Wang, Z., 2017. Chapter One - Maximizing Power in Phylogenetics and Phylogenomics: A Perspective Illuminated by Fungal Big Data, in: Townsend, J.P., Wang, Z. (Eds.), Advances in Genetics. Academic Press, pp. 1–47.

Drechsler, C., 1959. Several Zoopagaceae Subsisting on a Nematode and on Some Terricolous Amoebae. Mycologia 51, 787–823.

Dujon, B., 1996. The yeast genome project: what did we learn? Trends Genet. 12, 263–270.

Gawryluk, R.M.R., del Campo, J., Okamoto, N., Strassert, J.F.H., Lukeš, J., Richards, T.A., Worden, A.Z., Santoro, A.E., Keeling, P.J., 2016. Morphological Identification and Single-Cell Genomics of Marine Diplonemids. Curr. Biol. 26, 3053–3059.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S.G., 1996. Life with 6000 genes. Science 274, 546, 563–7.

Guo, Y., Takashima, Y., Sato, Y., Narisawa, K., Ohta, H., Nishizawa, T., 2020. Mycoavidus sp. Strain B2-EB: Comparative Genomics Reveals Minimal Genomic Features Required by a Cultivable Burkholderiaceae-Related Endofungal Bacterium. Appl. Environ. Microbiol. 86. https://doi.org/10.1128/AEM.01018-20

Hawksworth, D.L., Lücking, R., 2017. Fungal diversity revisited: 2.2 to 3.8 million species, in: The Fungal Kingdom. ASM Press, Washington, DC, USA, pp. 79–95.

Hibbett, D.S., 2007. After the gold rush, or before the flood? Evolutionary morphology of mushroom-forming fungi (Agaricomycetes) in the early 21st century. Mycol. Res. 111, 1001–1018.

Hibbett, D.S., Binder, M., Bischoff, J.F., Blackwell, M., Cannon, P.F., Eriksson, O.E., Huhndorf, S., James, T., Kirk, P.M., Lücking, R., Thorsten Lumbsch, H., Lutzoni, F., Matheny, P.B., McLaughlin, D.J., Powell, M.J., Redhead, S., Schoch, C.L., Spatafora, J.W., Stalpers, J.A., Vilgalys, R., Aime, M.C., Aptroot, A., Bauer, R., Begerow, D., Benny, G.L., Castlebury, L.A., Crous, P.W., Dai, Y.C., Gams, W., Geiser, D.M., Griffith, G.W., Gueidan, C., Hawksworth, D.L., Hestmark, G., Hosaka, K., Humber, R.A., Hyde, K.D., Ironside, J.E., Kõljalg, U., Kurtzman, C.P., Larsson, K.H., Lichtwardt, R., Longcore, J., Miadlikowska, J., Miller, A., Moncalvo, J.M., Mozley-Standridge, S., Oberwinkler, F., Parmasto, E., Reeb, V., Rogers, J.D., Roux, C., Ryvarden, L., Sampaio, J.P., Schüßler, A., Sugiyama, J., Thorn, R.G., Tibell, L., Untereiner, W.A., Walker, C., Wang, Z., Weir, A., Weiss, M., White, M.M., Winka, K., Yao, Y.J., Zhang, N., 2007. A higher-level phylogenetic classification of the Fungi. Mycol. Res. 111, 509–547.

Hibbett, D.S., Pine, E.M., Langer, E., Langer, G., Donoghue, M.J., 1997. Evolution of gilled mushrooms and puffballs inferred from ribosomal DNA sequences. Proc. Natl. Acad. Sci. U. S. A. 94, 12002–12006.

James, T.Y., Stajich, J.E., Hittinger, C.T., Rokas, A., 2020. Toward a Fully Resolved Fungal Tree of Life. Annu. Rev. Microbiol. 74, 291–313.

Kalisky, T., Quake, S.R., 2011. Single-cell genomics. Nat. Methods 8, 311–314.

Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., Wang, Z., 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 7, e7359.

Keller, S., 1997. The genus Neozygites (Zygomycetes , Entomophthorales) with special reference to species found in tropical regions. Sydowia 49, 118–146.

Kimmerling, R.J., Lee Szeto, G., Li, J.W., Genshaft, A.S., Kazer, S.W., Payer, K.R., de Riba

Borrajo, J., Blainey, P.C., Irvine, D.J., Shalek, A.K., Manalis, S.R., 2016. A microfluidic platform enabling single-cell RNA-seq of multigenerational lineages. Nat. Commun. 7, 10220.

Kohler, A., Kuo, A., Nagy, L.G., Morin, E., Barry, K.W., Buscot, F., Canbäck, B., Choi, C., Cichocki, N., Clum, A., Colpaert, J., Copeland, A., Costa, M.D., Doré, J., Floudas, D., Gay, G., Girlanda, M., Henrissat, B., Herrmann, S., Hess, J., Högberg, N., Johansson, T., Khouja, H.-R., Labutti, K., Lahrmann, U., Levasseur, A., Lindquist, E.A., Lipzen, A., Marmeisse, R., Martino, E., Murat, C., Ngan, C.Y., Nehls, U., Plett, J.M., Pringle, A., Ohm, R.A., Perotto, S., Peter, M., Riley, R., Rineau, F., Ruytinx, J., Salamov, A., Shah, F., Sun, H., Tarkka, M., Tritt, A., Veneault-fourrey, C., Zuccaro, A., Genomics, M., Consortium, I., Tunlid, A., Grigoriev, I.V., Hibbett, D.S., Martin, F., 2015. Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualists. Nature Publishing Group 47, 410–415.

Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M., Blaxter, M., 2013. Blobology : exploring raw genome data for contaminants , symbionts , and parasites using taxon-annotated GC-coverage plots. Front. Genet. 4, 1–12.

Laczny, C.C., Sternal, T., Plugaru, V., Gawron, P., Atashpendar, A., Margossian, H.H., Coronado, S., Maaten, L.V.D., Vlassis, N., Wilmes, P., 2015. VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. Microbiome 3, 1–7.

Lazarevic, V., Whiteson, K., Huse, S., Hernandez, D., Farinelli, L., Osterås, M., Schrenzel, J., François, P., 2009. Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. J. Microbiol. Methods 79, 266–271.

Li, M., Jain, S., Dick, G.J., 2016. Genomic and Transcriptomic Resolution of Organic Matter Utilization Among Deep-Sea Bacteria in Guaymas Basin Hydrothermal Plumes. Front. Microbiol. 7, 1125.

Li, Y., Steenwyk, J.L., Chang, Y., Wang, Y., James, T.Y., Stajich, J.E., Spatafora, J.W., Groenewald, M., Dunn, C.W., Hittinger, C.T., Shen, X.-X., Rokas, A., 2021. A genome-scale phylogeny of the kingdom Fungi. Curr. Biol. 31, 1653-1665.e5.

Locey, K.J., Lennon, J.T., 2016. Scaling laws predict global microbial diversity. Proc. Natl. Acad. Sci. U. S. A. 113, 5970–5975.

Lovmar, L., Syvänen, A.-C., 2006. Multiple displacement amplification to create a long-lasting source of DNA for genetic studies. Hum. Mutat. 27, 603–614.

Makiola, A., Dickie, I.A., Holdaway, R.J., Wood, J.R., Orwin, K.H., Lee, C.K., Glare, T.R., 2018. Biases in the metabarcoding of plant pathogens using rust fungi as a model system. Microbiologyopen e780.

Martin, F., Aerts, A., Ahrén, D., Brun, A., Danchin, E.G.J., Duchaussoy, F., Gibon, J., Kohler, A., Lindquist, E., Pereda, V., Salamov, A., Shapiro, H.J., Wuyts, J., Blaudez, D., Buée, M., Brokstein, P., Canbäck, B., Cohen, D., Courty, P.E., Coutinho, P.M., Delaruelle, C., Detter, J.C., Deveau, A., DiFazio, S., Duplessis, S., Fraissinet-Tachet, L., Lucic, E., Frey-Klett, P., Fourrey, C., Feussner, I., Gay, G., Grimwood, J., Hoegger, P.J., Jain, P., Kilaru, S., Labbé, J., Lin, Y.C., Legué, V., Le Tacon, F., Marmeisse, R., Melayah, D., Montanini, B., Muratet, M., Nehls, U., Niculita-Hirzel, H., Oudot-Le Secq, M.P., Peter, M., Quesneville, H., Rajashekar, B., Reich, M., Rouhier, N., Schmutz, J., Yin, T., Chalot, M., Henrissat, B., Kües, U., Lucas, S., Van de Peer, Y., Podila, G.K., Polle, A., Pukkila, P.J., Richardson, P.M., Rouzé, P., Sanders, I.R., Stajich, J.E., Tunlid, A., Tuskan, G., Grigoriev, I.V., 2008. The genome of Laccaria bicolor provides insights into mycorrhizal symbiosis. Nature 452, 88–92.

Mikhailov, K.V., Simdyanov, T.G., Aleoshin, V.V., Belozersky, A.N., 2016. Genomic survey of a hyperparasitic microsporidian Amphiamblys sp. (Metchnikovellidae). Genome Biol. Evol. 9, 454–467.

Mondo, S.J., Toomer, K.H., Morton, J.B., Lekberg, Y., Pawlowska, T.E., 2012. Evolutionary stability in a 400-million-year-old heritable facultative mutualism. Evolution 66, 2564–2576.

Naito, M., Desirò, A., González, J.B., Tao, G., Morton, J.B., Bonfante, P., Pawlowska, T.E., 2017. 'Candidatus Moeniiplasma glomeromycotorum', an endobacterium of arbuscular mycorrhizal fungi. Int. J. Syst. Evol. Microbiol. 67, 1177–1184.

Nelson, W.C., Anderson, L.N., Wu, R., McDermott, J.E., Bell, S.L., Jumpponen, A., Fansler, S.J., Tyrrell, K.J., Farris, Y., Hofmockel, K.S., Jansson, J.K., 2020. Terabase Metagenome Sequencing of Grassland Soil Microbiomes. Microbiol Resour Announc 9, e00718-20.

Nguyen, N.H., Song, Z., Bates, S.T., Branco, S., Tedersoo, L., Menke, J., Schilling, J.S., Kennedy, P.G., 2016. FUNGuild: An open annotation tool for parsing fungal community datasets by ecological guild. Fungal Ecol. 20, 241–248.

Novick, A., Szilard, L., 1950. Experiments with the Chemostat on spontaneous mutations of bacteria. Proc. Natl. Acad. Sci. U. S. A. 36, 708–719.

Pace, N.R., 1997. A molecular view of microbial diversity and the biosphere. Science 276, 734–740.

Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., Hugenholtz, P., 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. 36, 996–1004.

Partida-Martinez, L.P., Hertweck, C., 2005. Pathogenic fungus harbours endosymbiotic bacteria for toxin production. Nature 437, 884–888.

Pawlowska, T.E., Gaspar, M.L., Lastovetsky, O.A., Mondo, S.J., Real-Ramirez, I., Shakya, E., Bonfante, P., 2018. Biology of Fungi and Their Bacterial Endosymbionts. Annu. Rev. Phytopathol. 56, 289–309.

Pinard, R., de Winter, A., Sarkis, G.J., Gerstein, M.B., Tartaro, K.R., Plant, R.N., Egholm, M., Rothberg, J.M., Leamon, J.H., 2006. Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. BMC Genomics 7, 216.

Porter, J.R., 1976. Antony van Leeuwenhoek: tercentenary of his discovery of bacteria. Bacteriol. Rev. 40, 260–269.

Prasanna, A.N., Gerber, D., Kijpornyongpan, T., Aime, M.C., Doyle, V.P., Nagy, L.G., 2019. Model Choice, Missing Data, and Taxon Sampling Impact Phylogenomic Inference of Deep Basidiomycota Relationships. Syst. Biol. https://doi.org/10.1093/sysbio/syz029

Rinke, C., Lee, J., Nath, N., Goudeau, D., Thompson, B., Poulton, N., Dmitrieff, E., Malmstrom, R., Stepanauskas, R., Woyke, T., 2014. Obtaining genomes from uncultivated environmental microorganisms using FACS – based single-cell genomics. Nat. Protoc. 9, 1038–1048.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., Nyrén, P., 1996. Real-time DNA sequencing using detection of pyrophosphate release. Anal. Biochem. 242, 84–89.

Roy, R.S., Price, D.C., Schliep, A., Cai, G., Korobeynikov, A., Yoon, H.S., Yang, E.C., Bhattacharya, D., 2014. Single cell genome analysis of an uncultured heterotrophic stramenopile. Sci. Rep. 4, 1–8.

Saxena, G., 2008. Observations on the occurrence of nematophagous fungi in Scotland. Appl. Soil Ecol. 39, 352–357.

Sedlar, K., Kupkova, K., Provaznik, I., 2017. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. Comput. Struct. Biotechnol. J. 15, 48–55.

Shendure, J., Ji, H., 2008. Next-generation DNA sequencing. Nat. Biotechnol. 26, 1135–1145.

Sieber, C.M.K., Probst, A.J., Sharrar, A., Thomas, B.C., Hess, M., Tringe, S.G., Banfield, J.F., 2018. Recovery of genomes from metagenomes via a depreplication, aggregation and scoring str. Nature Microbiology. https://doi.org/10.1038/s41564-018-0171-1

Siefert't, J.L., Fox, G.E., 1998. Phylogenetic mapping of bacterial morphology. Microbiology 144 (Pt 10), 2803–2808.

Smith, M.L., Bruhn, J.N., Anderson, J.B., 1992. The fungus Armillaria bulbosa is among the largest and oldest living organisms. Nature 356, 428–431.

Spatafora, J.W., Chang, Y., Benny, G.L., Lazarus, K., Smith, M.E., Berbee, M.L., Bonito, G., Corradi, N., Grigoriev, I., Gryganskyi, A., James, T.Y., O'Donnell, K., Roberson, R.W., Taylor, T.N., Uehling, J., Vilgalys, R., White, M.M., Stajich, J.E., 2016. A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data. Mycologia 108, 1028–1046.

Steinkraus, D.C., Boys, G.O., Rosenheim, J.A., 2002. Classical biological control of Aphis gossypii (Homoptera : Aphididae) with Neozygites fresenii (Entomophthorales : Neozygitaceae) in California cotton. Biol. Control 25, 297–304.

Streit, W.R., Schmitz, R.A., 2004. Metagenomics--the key to the uncultured microbes. Curr. Opin. Microbiol. 7, 492–498.

Tedersoo, L., Anslan, S., Bahram, M., Põlme, S., Riit, T., Liiv, I., Kõljalg, U., Kisand, V., Nilsson, H., Hildebrand, F., Bork, P., Abarenkov, K., 2015. Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. MycoKeys 10, 1–43.

Tedersoo, L., Bahram, M., Põlme, S., Kõljalg, U., 2014. Global diversity and geography of soil fungi.

Tedersoo, L., Sánchez-Ramírez, S., Kõljalg, U., Bahram, M., Döring, M., Schigel, D., May, T., Ryberg, M., Abarenkov, K., 2018. High-level classification of the Fungi and a tool for evolutionary ecological analyses. Fungal Divers. 90, 135–159.

Toomer, K.H., Chen, X., Naito, M., Mondo, S.J., den Bakker, H.C., VanKuren, N.W., Lekberg, Y., Morton, J.B., Pawlowska, T.E., 2015. Molecular evolution patterns reveal life history features of mycoplasma-related endobacteria associated with arbuscular mycorrhizal fungi. Mol. Ecol. 24, 3485–3500.

Torres-Cortés, G., Ghignone, S., Bonfante, P., Schüßler, A., 2015. Mosaic genome of endobacteria in arbuscular mycorrhizal fungi: Transkingdom gene transfer in an ancient mycoplasma-fungus association. Proc. Natl. Acad. Sci. U. S. A. 112, 7785–7790.

Tucker, T., Marra, M., Friedman, J.M., 2009. Massively parallel sequencing: the next big thing in genetic medicine. Am. J. Hum. Genet. 85, 142–154.

van Niel, C.B., 1944. THE CULTURE, GENERAL PHYSIOLOGY, MORPHOLOGY, AND CLASSIFICATION OF THE NON-SULFUR PURPLE AND BROWN BACTERIA. Bacteriol. Rev. 8, 1–118.

Wekesa, V.W., Moraes, G.J., Knapp, M., Jr, I.D., 2007. Interactions of two natural enemies of Tetranychus evansi , the fungal pathogen Neozygites floridana (Zygomycetes : Entomophthorales) and the predatory mite , Phytoseiulus longipes (Acari : Phytoseiidae). Biol. Control 41, 408–414.

Whisler, H.C., Travland, L.B., 1974. The rotifer trap of Zoophagus. Arch. Microbiol. 101, 95–107.

White, M.M., James, T.Y., Donnell, K.O., Cafaro, M.J., Tanabe, Y., Sugiyama, J., James, T.Y., Carolina, N., Donnell, K.O., 2006. Phylogeny of the Zygomycota based on nuclear ribosomal sequence data. Mycologia 98, 872–884.

Wu, Y.-W., Simmons, B.A., Singer, S.W., 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32, 605–607.

Yaninek, J.S., Moraes, G.J.D.E., Oduor, G.I., 2002. Host specificity of the cassava green mite pathogen Neozygites floridana 61–66.

Yue, Y., Huang, H., Qi, Z., Dou, H.-M., Liu, X.-Y., Han, T.-F., Chen, Y., Song, X.-J., Zhang, Y.-H., Tu, J., 2020. Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. BMC Bioinformatics 21, 334.

Zhou, J., Wu, L., Deng, Y., Zhi, X., Jiang, Y.-H., Tu, Q., Xie, J., Van Nostrand, J.D., He, Z., Yang, Y., 2011. Reproducibility and quantitation of amplicon sequencing-based detection. ISME J. 5, 1303–1313.

1.10 Figures

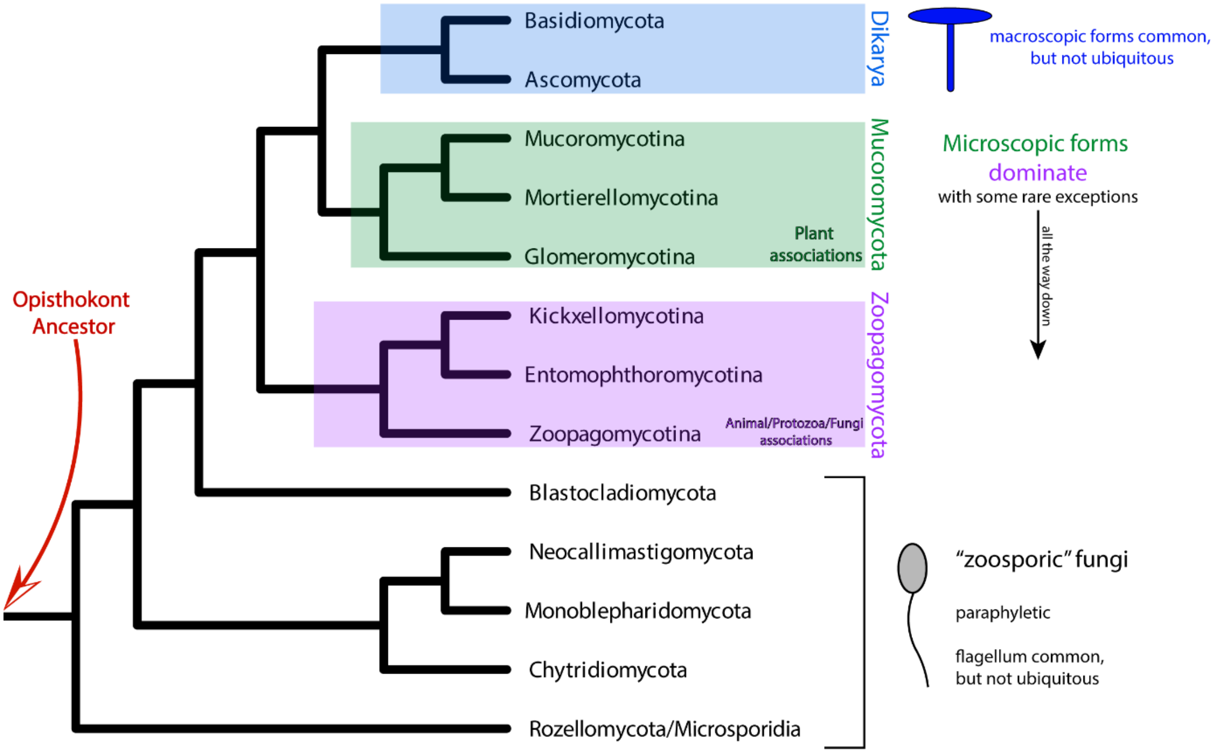


Figure 1.1. Artistic rendition of some of the major superphylum- (i.e., Dikarya), phylum- (*mycota) and subphylum-level (*mycotina) divisions of the Kingdom Fungi as it is currently understood. Recently recognized phyla Entorrhizomycota and Aphelidimycota are not shown. Macroscopic fungal forms dominate in Dikarya, with microscopic forms being the overarching norm in earlier-diverging lineages, with a few rare examples of macroscopic forms (e.g., *Endogone*). When symbiotic interactions occur, fungi in the Mucoromycota tends to associate with plants while fungi in the Zoopagomycota tend to associate with animals, protozoans, or other fungi. Zoosporic fungi comprise a paraphyletic, artificial assemblage. The root of the cladogram continues backwards in time until it connects with the Opisthokont MRCA of Animalia and Fungi.

Chapter 2: SCGid, a consensus approach to contig filtering and genome prediction from single-cell sequencing libraries of uncultured eukaryotes

2.1 Abstract

Whole-genome sequencing of uncultured eukaryotic genomes is complicated by difficulties in acquiring sufficient amounts of tissue. Single-cell genomics (SCG) by multiple displacement amplification provides a technical workaround, yielding whole-genome libraries which can be assembled de novo. Downsides of multiple displacement amplification include coverage biases and exacerbation of contamination. These factors affect assembly continuity and fidelity, complicating discrimination of genomes from contamination and noise by available tools. Uncultured eukaryotes and their relatives are often underrepresented in large sequence data repositories, further impairing identification and separation. We compare the ability of filtering approaches to remove contamination and resolve eukaryotic draft genomes from SCG metagenomes, finding significant variation in outcomes. To address these inconsistencies, we introduce a consensus approach that is codified in the *SCGid* software package. *SCGid* parallelly filters assemblies using different approaches, yielding three intermediate drafts from which consensus is drawn. Using genuine and mock SCG metagenomes, we show that our approach corrects for variation among draft genomes predicted by individual approaches and outperforms them in recapitulating published drafts in a fast and repeatable way, providing a useful alternative to available methods and manual curation. The *SCGid* package is implemented in python and R. Source code is available at http://www.github.com/amsesk/*SCGid* under the GNU GPL 3.0 license.

2.2 Introduction

Contamination is an ever-present concern in the preparation of high-throughput sequencing libraries. Certain methods of sample preparation are more susceptible to contamination, whether it is from the laboratory or from the environment. Approaches that involve a non-specific amplification step, such as the multiple displacement amplification (MDA) associated with single-cell genomics (SCG), are especially prone to contamination from these sources. As DNA is amplified non-specifically, even small amounts of contamination, including that derived from the MDA reagents themselves, can lead to significant dilution of target molecules (Gawad et al., 2016; Rinke et al., 2014). Perhaps best known for its applications in model systems where it can capture cell-to-cell heterogeneity in molecular processes, SCG has also been leveraged toward generating genome-scale data for groups of uncultured bacteria, archaea, fungi and protozoans (Ahrendt et al., 2018; Davis et al., 2019; Gawryluk et al., 2016; Mikhailov et al., 2016; Rinke et al., 2013; Roy et al., 2014).

Uncultured microbes are those that cannot or have not been successfully grown axenically in pure laboratory cultures. Their study necessitates that tissues be collected directly from the environment or from highly mixed in vitro microcosms. Collecting ample material that is reasonably pure and yields sufficient quantities of DNA to serve as input for whole-genome sequencing is often a near insurmountable obstacle. While SCG techniques circumvent this obstacle through non-specific DNA amplification, the data they yield poses a unique set of bioinformatic challenges: (i) SCG is highly subject to contamination, making most, if not all, SCG-derived genomes of uncultured microbes mildly to moderately metagenomic (Davis et al., 2019; Gawryluk et al., 2016; Mikhailov et al., 2016; Roy et al., 2014); (ii) despite the capacity for fully factorial priming, the replication enzymes involved in MDA introduce amplification biases that eventually manifest as read libraries that do not accurately represent the starting population of template molecules, making coverage statistics less reliable (Gawad et al., 2016; Pinard et al., 2006) and, (iii) uncultured organisms are often underrepresented in sequence databases, complicating taxonomic delineation from contaminants. Taken together, all of these factors make identification of the target genome from noise a major obstacle.

2.3 Approaches to isolating genomes from metagenomes

Methods for extracting individual genomes from metagenomic data are diverse. Utilizing features inherent to or derivative of nucleotide sequences, these approaches cluster contigs independent of any taxonomy assigned by BLAST searches of large sequence repositories (i.e., taxonomy-independent binning) (Sedlar et al., 2017). Common features include the relationship between GC-content and coverage, *k*mer frequencies and relative synonymous codon usage (RSCU) (Dick et al., 2009; Kumar et al., 2013; Laczny et al., 2015; McInerney, 1998; Mikhailov et al., 2016; Sedlar et al., 2017; Wu et al., 2016). Despite being clustered independent of taxonomy, identification, selection and verification of clusters is almost always informed by assigned taxonomy (Dick et al., 2009; Kumar et al., 2013; Laetsch et al., 2017; Mikhailov et al., 2016).

The relationship between GC-content and coverage

GC-coverage-taxonomy (GCT) plots graph contigs as points in two dimensions, allowing visualization and separation of metagenomic assemblies into clusters based on the GC-content and sequencing depth (i.e., coverage) of their constituent contigs (e.g., Figure 2.1) (Kumar et al., 2013; Laetsch et al., 2017). Since GC-content varies in between organisms and per-organism genome coverage is correlated with the relative abundance of fragments of its DNA in the sequencing library, these clusters can correspond to the genomes of individual organisms. Points are annotated with taxonomic information from nucleotide BLAST searches of large sequencing databases to determine the taxonomic affinities of clusters. Resolution depends on the complexity of the metagenome, the quality of the annotations and the phylogenetic distances between constituent genomes. GCT plots quickly visualize the 'metagenomic-ness' of assemblies and can be used to determine GC and coverage cutoffs for extracting particular clusters for independent processing and analysis (Kumar et al., 2013; Laetsch et al., 2017).

kmer frequencies

Separation of individual genomes from metagenomic backgrounds by *k*mer frequencies hinges on the assumption that the frequencies of specific oligonucleotide sequences of length *k* are internally consistent across each genome. Under this assumption, the frequencies of any particular *k*mer on assembled contigs that originate from the same genome will be similar, distributed around the frequency of that *k*mer in the entire genome. While *k*mers cannot be homogeneously distributed within genomes, *k*mer frequencies can be used to cluster metagenome assemblies and separate sets of contigs belonging to individual genomes (Dick et al., 2009). One approach applies unsupervised machine learning to cluster a matrix of the relative frequencies of all informative *k*mers across a contig (its *k*mer profile) to generate emergent self-organizing maps (ESOMs) that visualize this *n*-dimensional data (Dick et al., 2009; Ultsch and Moerchen, 2005). This approach yields a 2D topology that visualizes the boundaries between clusters, and theoretically, individual genomes (e.g., Figure 2.1). Taxonomic annotations can be overlaid the final topology to predict the identity of clusters, which can then be carved out of the larger map by eye and analyzed independently (Dick et al., 2009).

Relative synonymous codon usage

RSCU measurements are numerical representations of codon bias, describing the preferential use of different codons coding for the same amino acid (i.e., synonymous codons) in protein coding nucleotide sequences (CDS) (McInerney, 1998; Mikhailov et al., 2016). As codon bias is often species-specific, RSCU profiles represent another feature by which assembled contigs can be clustered and separated. Following protein annotation, coding portions of contigs are concatenated into a single joint CDS sequence for each contig, upon which whole-contig RSCU profiles are calculated. RSCU values for each of the 59 codons with alternative synonymous codons that are not STOP codons are calculated across the entire concatenate according to the generalized expression considering codon i…

$$RSCU_i = \frac{X_i}{\frac{1}{n}\sum_{i=1}^{n} X_i}$$

…where n is the number of codons synonymous to i and X_i is the number of occurrences of i in the concatenate (McInerney, 1998). These profiles are subsequently used to generate an RSCU distance matrix based on the generalized distance measure…

$$D_{jk} = \sum_{i=1}^{n} \frac{|RSCU_{ji} - RSCU_{ki|}}{n}$$

…where $RSCU_{ji}$ is the RSCU of codon i on CDS concatenate j, $RSCU_{ki}$ is the RSCU of codon i on concatenate k and n is the total number of synonymous codons in the concatenate (McInerney, 1998). Hierarchical clustering of RSCU matrices exposes clusters of contigs with similar profiles that can be assigned taxonomy by BLAST searches (e.g., Figure 2.1). Clusters of contigs with known or inferred origin can be used as training sets in subsequent rounds of clustering by different features to retrieve the short, protein-less contigs that could not be included in the initial clustering (Mikhailov et al., 2016).

2.4 Obstacles to filtering single-cell eukaryotic metagenomes

Methods currently available for separating metagenomes address some of the issues associated with filtering SCG assemblies, but there remain gaps in their ability to do so. While useful features for clustering are necessarily present in sequences from across the tree of life, the collection of tools that use them is generally skewed toward prokaryotes (Sieber et al., 2018; Wu et al., 2016). This limits the pool of available options when filtering SCG assemblies of uncultured eukaryotes. Tools that lean on contig coverage for clustering (Kumar et al., 2013; Wu et al., 2016) have considerably less utility with SCG because of the biased sequencing depth that characterizes *de novo* assemblies (Davis et al., 2019; Pinard et al., 2006). This bias tends to lead to *de novo* assemblies that are highly fragmented, introducing significant variance in contig-level sequence features (e.g., *k*mer frequencies) used for clustering and negatively affecting filtering outcomes (Davis et al., 2019). Moreover, as the target organisms of SCG and their relatives are usually uncultured, contigs belonging to their genomes rarely share sufficient sequence similarity with those contained in public sequence repositories. This impairs the ability of BLAST searches to assign taxonomy for the vast majority of contigs, making annotation difficult.

Despite these obstacles, filtering SCG metagenomes of uncultured eukaryotes with available tools can yield draft genomes predicted to be nearly complete (Davis et al., 2019; Mikhailov et al., 2016). However, there often remains uncertainty in the fidelity of filtered drafts because verification by unified taxonomy or coverage information is difficult or impossible. Downstream analyses of these drafts are imbued with similar uncertainty when the inclusion or exclusion of a contig could arbitrarily introduce false negatives for genome functionality or attribute functionality that is derived from a contaminant.

2.5 *SCGid*: a consensus-based filtering tool for SCG of uncultured eukaryotes

To address this uncertainty and fully investigate the efficacy of different approaches in filtering de novo assemblies of single-cell sequencing libraries, we implemented three in *SCGid*, an automated filtering tool for SCG assemblies. *SCGid* filters assemblies separately using each approach described above, generating three intermediate drafts. A final consensus draft is generated by majority rule at the overlaps of different approaches, where inclusion of a contig is dependent on its inclusion in two of the three intermediate drafts (Figure 2.1). Each filtering approach, including consensus, is invoked separately from the CLI. Module-specific implementations, discussed in the coming sections, introduce novel code automating all but a single step of the tripartite pipeline. Automation is enabled by *SCGid*'s requirement of *a priori* specifications of 'target' taxa. This duality of 'target' and 'nontarget' taxonomic annotations is hereafter referred to as such. To reduce computational time spent assigning taxonomy, *SCGid* uses the Uniprot *swissprot* database (SPDB) for protein sequences and the full NCBI *nt* database for nucleotide sequences (NCBI Resource Coordinators, 2017; The UniProt Consortium, 2017). To increase coverage of non-model lineages, utilities are included to supplement the SPDB with additional protein sequences.

GCT plots (SCGid gc-cov)

SCGid plots BLAST-annotated *AUGUSTUS*-predicted (Stanke and Morgenstern, 2005) proteins as points in GC-coverage space and draws a total of 13 separate flexible selection windows (FSWs) around them (Figure 2.2A). The 2D bounds of windows are calculated with respect to proteins that had a significant hit (e-value ≤ 1e-5, by default) in the SPDB. These bounds are used downstream to make inclusion decisions on contigs that either contain no proteins or contain proteins with no significant hit (i.e., unclassified contigs). All contigs identified as target by virtue of the sum and strength of their protein hits are *ad hoc* included in the GC-coverage-based filtered draft, by default. The flexibility of FSWs provides a unique SCG optimization as it allows for wide GC and coverage distributions, artifacts of highly fragmented assemblies and MDA amplification, respectively.

The bounds of FSWs are calculated through two sequential rounds of 1D expansion, one along each axis (e.g., round 1 along GC, round 2 along coverage). Beginning at the mean value of target points on that axis, expansion outward is incremental, proceeding to the limits of annotated points (Figure 2.2B). The proportions of target and nontarget proteins inside versus outside the bounds are computed at each step, $P_{tar} = tar_{inside}/tar_{total}$ and $P_{ntar}=ntar_{inside}/ntar_{total}$, and used to calculate a trade-off value defined as $D_{tradeoff} = P_{tar}\ (P_{tar}–P_{ntar})$ (Figure 2.2B,F). At the end of each round of expansion, bounds are set where $D_{tradeoff}$ is maximized (Figure 2.2F). The second round of expansion is identical to the first except that all points outside the bounds set in round 1 are ignored (Figure 2.2C,D). The end product is a 2D FSW with cutoffs on both the GC and coverage axes (Figure 2.2G, crosshatched region).

Accounting for all thirteen FSWs, to cope with dataset-specific distributional differences in GC-Coverage space, *SCGid* draws FSWs for all factorial combinations of first axis analyzed (GC or coverage) and three expansion types (unbounded = 0, coupled = 1 or uncoupled = 2) (Figure 2.2E). Unbounded (0) expansion rounds do not compute P_{tar}, P_{ntar} or $D_{tradeoff}$ at all, merely setting the bounds at the limits of annotated points along that axis (gc0 or co0) (Figure 2.2E). Coupled (1) expansion rounds compute P_{tar}, P_{ntar} or $D_{tradeoff}$ once per step for the positive and negative directions taken together (gc1 or co1) (Figure 2.2E). Uncoupled (2) expansion rounds compute P_{tar}, P_{ntar}, and $D_{tradeoff}$ twice per step for the positive and negative directions separately, allowing for unequal bound divergence from the mean (gc2 or co2) (Figure 2.2A–G). From this set of FSWs (Figure 2.2E), an optimal window is chosen that maximizes P_{tar}, but minimizes P_{ntar} at or below a set stringency level, *s* (i.e., $P_{ntar} \leq$ s). As stated above, cutoffs defined by the optimal window determine the inclusion or exclusion of unclassified contigs (Figure 2.2G, blue points in crosshatched region). All contigs identified as target are included, by default.

Emergent self-organizing maps (SCGid kmers)

SCGid provides automated preparation of all the files required to train and generate an ESOM topology using outside scripts and Databionics ESOM Tools (Dick et al., 2009; Ultsch and Moerchen, 2005). *SCGid* introduces an automated annotation pipeline that links contigs with their best BLAST hit in the NCBI *nt* database, coloring them according to user-defined taxonomic levels. The task of sectioning-out a target cluster from the topology (using Databionics ESOM tools) relies on the user. An automated algorithm has not yet been implemented in *SCGid* and mouse-sectioning by human eye is standard practice (Dick et al., 2009; Ultsch and Moerchen, 2005). Following sectioning and export, *SCGid* pulls the contigs belonging to the target class, yielding the ESOM-filtered draft assembly.

Relative synonymous codon usage (SCGid codons)

SCGid implements RSCU-based metagenome filtering in line with the concepts and applications described above (McInerney, 1998; Mikhailov et al., 2016). CDS sequences are pulled from

AUGUSTUS models and joined into a single CDS concatenate for each contig. Short concatenates are discarded (<3000 bp, by default). RSCU profiles are calculated for large concatenates and used to compute an RSCU distance matrix (McInerney, 1998). A neighbor-joining tree is computed from this matrix, the tips of which are assigned taxonomy. The tree is iteratively searched, and all sufficiently sized clades (≥30 tips, by default) are binned by shared node architecture to avoid duplication. Clades in bins are ranked by the target-nontarget ratio of their descendant tips; ties are resolved by maximizing clade size. The highest-ranking clades from each bin are compared and the best clade, presumed to originate from the target genome, is nominated as a training set to collect small protein-less contigs from the rest of the metagenome. Clustering is done in *ClaMs* (Pati et al., 2011), a *k*mer-based (k = 2, by default) binning algorithm that assesses contig similarity to the trainset (Pearson's distance ≤ 0.1) and bins them accordingly (Mikhailov et al., 2016).

2.6 Validation

2.6.1 Methods

To assess the performance of our filtering implementations and the ability of consensus to resolve inconsistencies between them, we ran *SCGid* on two mock and three elsewhere-published SCG datasets (Davis et al., 2019; Mikhailov et al., 2016; Roy et al., 2014).

Dataset selection

We generated two mock-MDA read libraries from the *Saccharomyces cerevisiae* S288C reference genome. We selected three studies where MDA was used to prepare sequencing libraries and that had the goal of generating draft genome sequences for one or more uncultured eukaryotes. To sample from a range of eukaryotic lineages, we selected two studies targeting fungi, one microsporidian (PRJNA321520) and five zoopagalean fungi (PRJNA451036), and one targeting a stramenopile (PRJNA244411) (Davis et al., 2019; Mikhailov et al., 2016; Roy et al., 2014). The studies' filtering methods involved various tools and levels of scrutiny, sometimes implementing similar approaches to those implemented in *SCGid*, other times relying solely on taxonomy-dependent approaches.

Mock dataset preparation

We artificially contaminated the *S. cerevisiae* S288C reference genome with the green alga *Chlamydomonas reinhardtii* CC-503 cw92 mt+, and the bacteria *Bacillus cereus* ATCC 14579, *Cellulomonas* sp. FA1 GY42 and *Pseudomonas putida* KT2440 (Belda et al., 2016; Cohen et al., 2015; Fisk et al., 2006; Ivanova et al., 2003; Merchant et al., 2007). To test *SCGid* on both eukaryotic and prokaryotic contamination, we generated two mock-MDA read libraries *in silico* that were either contaminated with just the bacteria (mockB) or the bacteria and *C. reinhardtii* (mockBE). To simulate biased and unequal coverage across the metagenome, we used bounded Brownian motion to generate unique discrete probability mass functions for each chromosome or contig that modulated the likelihood of each nucleotide being sampled as a start point for a 500 bp fragment (e.g., Figure A1). We sampled fragment start locations from these distributions and read 150 bp from both ends (i.e., paired-end), sampling to a mean expected coverage of 80× without simulating sequencing errors. In this way, we simulated the output of sequencing an MDA-derived library from three or four cells on the Illumina NextSeq platform. The mock metagenomes were assembled using *SPAdes* v3.9.0 (Bankevich et al., 2012), yielding initial assemblies of 58.48 Mbp on 3,102 contigs (coverage range: 2.45–17,369.63x, mean = 60.04x) and 127.80 Mbp on 31,781 contigs (coverage range: 1.172–10,261, mean = 134.26x) for mockB and mockBE, respectively, confirming that our fabricated SCG metagenomes were MDA-like (i.e., fragmented with wide coverage distributions). All contigs <200 bp were trimmed from initial assemblies prior to filtering. To simulate under-representation of *S. cerevisiae* during filtering, we manually purged the SPDB of all entries corresponding to the Saccharomycotina.

Genuine SCG dataset preparation

Since initial unfiltered assemblies are not usually made publicly available upon publication, we independently processed and assembled libraries of raw paired-end reads deposited in NCBI SRA according to the methods and parameters outlined by the authors (Mikhailov et al., 2016; Roy et al., 2014). Since we authored the study for the five zoopagalean fungi featured here, we worked directly with our initial assemblies (Davis et al., 2019).

Analysis of filtering outcomes

We compared filtering outcomes to each other, to their corresponding consensus draft, and to the published assembly. Comparisons were made on the basis of cumulative assembly size, number of contigs and CEGMA/BUSCO completeness (Parra et al., 2007; Waterhouse et al., 2017). Where informative, we made whole-genome alignments in *MUMmer* v3.23 (Kurtz et al., 2004) to quantify and visualize the proportion of the published assembly that was recapitulated in the *SCGid* consensus draft. For the mock datasets, we split the read libraries based on origin and mapped them to each assembly with *BWA-MEM* (Li, 2013) to quantify the respective contribution of yeast or contamination to filtered draft size.

2.6.2 Results

Automated filtering with *SCGid* yielded three filtered drafts and one consensus-filtered draft for each organism. In total, we generated 36 filtered assemblies for 9 target organisms. The filtered drafts predicted by separate approaches were often different, distinct in number of contigs, cumulative sequence length and predicted completeness (Figure 2.3). Filtering with the consensus method applied by *SCGid* averaged sometimes dramatic variation where present, yielding conservative filtered drafts at the overlaps of different approaches and sometimes improving completeness (Figure 2.3, pink bars; Table A1). In general, *SCGid* consensus recapitulated the sequence content and genome size of reference genomes and published drafts (Figure 2.3 bottom, dashed bars; Table A1).

Mock Saccharomyces cerevisiae SCG metagenomes

Automated filtering of the two mock SCG metagenomes yielded *S. cerevisiae* consensus drafts that nearly recapitulated the size of the 12.16 Mbp S288C reference genome: 11.76 Mbp on 436 contigs and 11.47 Mbp on 280 contigs for mockB and mockBE, respectively. Individual filtering approaches commonly yielded different drafts compared to the reference or even the draft produced by that approach on the other mock (Figure 2.3; Table A1). GC-coverage (i.e., *SCGid gc-cov*) either over- or under-filtered (mockB: 10.01 Mbp on 390 contigs, mockBE: 14.7 Mbp on 432 contigs), *k*mer frequencies (i.e., *SCGid kmers*) over-filtered in both cases, dramatically so for mockBE (mockB: 11.47 Mbp on 396 contigs; mockBE: 5.4 Mbp on 158 contigs) and RSCU (i.e., *SCGid codons*) under-filtered both metagenomes, generating similarly sized drafts (mockB:

16.72 Mbp on 529 contigs; mockBE: 16.83 Mbp on 398 contigs). Consensus outperformed all three on the basis of closest cumulative sequence length.

SCGid consensus drafts were mostly composed of yeast sequence data with relatively small fractions of contamination. With only bacterial contamination included (i.e., mockB), *SCGid kmers* produced the best draft, with a 99.21–0.15% ratio of mapped reads originating from yeast versus contamination, compared to 98.69–2.62% for consensus (*SCGid gc-cov*: 82.78–2.62%; *SCGid codons*: 98.68–34.46%). With bacterial and eukaryotic contamination included (mockBE), consensus outperformed individual approaches with a 98.04–1.38% ratio (*SCGid gc-cov*: 98.32–9.08%; *SCGid kmers*: 48.04–2.8%; *SCGid codons*: 98.94–5.70%). Taken together, these results underpin the uncertainty in filtering SCG metagenomes using any one approach and demonstrate the benefits of consensus.

Five zoopagalean fungi

As we noted in the original publication, manually-applied consensus averaged variation among separate filtering approaches and reduced uncertainty in the final drafts (Davis et al., 2019). Compared to those consensus drafts, automated *SCGid* filtering tended to increase assembly size and predicted completeness (Figure 2.3; Table A1). The filtered assemblies of *Zoopage sp.* (Zsp) and *Zoophagus insidians* (Zi) were significantly increased in size from 13.92 Mbp on 1,958 contigs to 17.84 Mbp on 2,892 contigs (*SCGid gc-cov*: 20.71 Mbp, 3,809 contigs; *SCGid kmers*: 13.29 Mbp, 2,056 contigs; *SCGid codons*: 48.01 Mbp, 5,358 contigs) and from 21.01 Mbp on 2,432 contigs to 31.01 Mbp on 5,839 contigs (*SCGid gc-cov*: 24.37 Mbp, 3,360 contigs; *SCGid kmers*: 15.83 Mbp, 6,055 contigs; *SCGid codons*: 128.10 Mbp, 20,013 contigs), respectively. Those of *Acaulopage tetraceros* (At) and *Cochlonema odontosperma* (Co) were only marginally increased from 10.20 Mbp on 472 contigs to 11.20 Mbp on 525 contigs (*SCGid gc-cov*: 11.45 Mbp, 539 contigs; *SCGid kmers*: 11.20 Mbp, 523 contigs; *SCGid codons*: 19.10 Mbp, 597 contigs) and 16.84 Mbp on 1,819 contigs to 18.05 Mbp on 2,274 contigs (*SCGid gc-cov*: 17.81 Mbp, 2,108 contigs; *SCGid kmers*: 18.26 Mbp, 2,399 contigs; *SCGid codons*: 17.84 Mbp, 2,670 contigs), respectively. Finally, the *Stylopage hadra* (Sh) assembly decreased in size from 55.96 Mbp on 20,112 contigs to 53.01 Mbp on 18,082 contigs (*SCGid gc-cov*: 44.96 Mbp, 13,902 contigs; *SCGid kmers*: 57.76 Mbp, 21,459 contigs; *SCGid codons*: 42.77 Mbp, 11,592 contigs).

Increases in assembly size were often accompanied by boosts in predicted completeness. Predicted completeness of At and Zsp were greatly increased from 83.06% to 90.32% and 71.77% to 78.63%, respectively. Co and Zi only saw marginal boosts from 89.52% to 89.92% and 90.73% to 91.13%, respectively. Consistent with a decrease in assembly size, predicted completeness of Sh was marginally decreased from 77.42% to 77.02% (Figure 2.3; Table A1).

Amphiamblys sp.

SCGid yielded a consensus draft of 6.09 Mbp on 1,464 contigs, compared to the published assembly of 5.62 Mbp on 1,727 contigs (Mikhailov et al., 2016). The *SCGid* consensus draft was more similar in size to the published draft than those of separate approaches (*SCGid gc-cov*: 13.08 Mbp, 17,469 contigs; *SCGid kmers*: 8.60 Mbp, 1,987 contigs; *SCGid codons*: 10.69 Mbp, 3,628 contigs; Figure 2.3, Table A1).

In the original publication, completeness was estimated at ~90% with a custom microsporidian database of core eukaryotic genes in *BUSCO* v1.1b (Mikhailov et al., 2016; Sima et al., 2015). Unable to directly replicate the unpublished custom database, we instead compared completeness of both assemblies using the *fungi_odb9* database in *BUSCO* v3.0.2 (Waterhouse et al., 2017). Of the 290 core fungal genes in *fungi_odb9*, the *SCGid* assembly contained 205 complete copies (70.69%) while the original published assembly contained only 193 complete copies (66.55%), equating to a 4.14% completeness advantage in favor of the *SCGid* assembly (Figure 2.3, Table A1).

Whole-genome alignment detected ~740 contigs with cumulative length 0.508 Mbp in the published assembly that was unaccounted for in the *SCGid*-filtered assembly and ~880 contigs with cumulative length 1.59 Mbp in the *SCGid*-filtered assembly that was unaccounted for in the published draft (Figure A2). These values indicate that the unaligned contigs were generally quite short. To confirm that alignments were not being made too liberally, we measured sequence similarity between the two drafts (Figure A3). When ordered by decreasing contig size, there is a general trend of decreased sequence identity toward the end of the published draft that we explain as variability in initial assemblies.

MAST-4 type stramenopile

The automated *SCGid* run yielded a final consensus draft of 13.08 Mbp on 3,298 contigs compared to the published draft of 16.93 Mbp on 4,611 contigs (Roy et al., 2014). The *SCGid* consensus draft was most similar in size to the published draft (*SCGid gc-cov*: 12.98 Mbp, 4,647 contigs; *SCGid kmers*: 12.71 Mbp, 2,128 contigs; *SCGid codons*: 12.88 Mbp, 3,195 contigs; Figure 2.3; Table A1). Predictions of genome completeness using the *eukaryota_odb9* database (303 core eukaryotic genes) favored the published draft with 102 complete copies (33.66%) compared to 83 complete copies (27.39%) in the *SCGid* consensus draft. Whole-genome alignment with *MUMmer* identified 1,803 contigs with a cumulative sequence length of 1.61 Mbp in the published draft that were unaccounted for in the *SCGid* consensus draft. The *SCGid*-predicted genome draft contained 172 contigs with a cumulative sequence length of 0.081 Mbp that were unaccounted for in the published draft.

2.7 Discussion

We demonstrate that the outcomes of filtering SCG metagenomes can vary dramatically with the particular approach taken. *SCGid* is a consensus filtering tool designed to address this problem. It brings automation to the process of filtering SCG metagenomes, offering an alternative to the time-consuming manual curation or strict BLAST-based filtering that are typical of most SCG projects to date. It is a fast and informative tool that quickly characterizes the landscape of SCG metagenomes and produces filtered drafts at the interstices of three different approaches.

We go on to show that *SCGid* successfully filters both genuine and fabricated SCG metagenomes. We demonstrate *SCGid*'s ability to recover the well-known *S. cerevisiae* S288C reference genome from a significantly muddled background using databases simulating its novelty. We benchmark *SCGid* against filtering approaches used in the literature, where it recapitulates final genome size, content, and completeness. For five zoopagalean fungi, *SCGid* generally predicted larger filtered drafts than those we previously published (Davis et al., 2019). Compared to the published *Amphiamblys* sp. assembly, *SCGid* yielded a similarly sized draft that corresponds well to the published draft (Mikhailov et al., 2016). While *SCGid* generated a smaller draft for the MAST-4-like stramenopile, it is not evident that any filtering was conducted in the original publication, indicating that perhaps *SCGid* filtered out previously-overlooked contamination (Roy et al., 2014). In terms of predicted completeness, *SCGid*-filtered drafts landed on both sides of the line, overall tending to increase completeness: an average +2.91% for five zoopagalean fungi, +4.14% for microsporidian Amphiamblys sp. and -6.27% for a MAST-4-like stramenopile.

SCGid's consensus approach blends the outcomes of the filtering approaches it employs, leading to conservatism in contig inclusion decisions. We view this is as a beneficial trait as it protects against the over-inclusion of sequence data, the converse of which can lead to misrepresentations of biology as inferred from genome annotation and pollute public repositories with misidentified sequence data. While there is the potential for contigs that belong to be excluded by consensus, the majority of contigs that are selected against are either non-coding or of unknown function and do not usually contribute to predicted genome function or completeness. Further, consensus offers protection against the unstable behavior of individual approaches confronted with different metagenomic backgrounds. Given the fundamental reliance of these filtering approaches on sequence data, it is not surprising that decreasing phylogenetic distance between contaminants and target can obscure filtering outcomes. In filtering mock *S. cerevisiae* S288C SCG metagenomes, two of the three filtering approaches (*SCGid gc-cov* and *SCGid kmers*) yielded very different outcomes dependent on the inclusion of algal contamination. Encouragingly, despite over- or under-filtered intermediate drafts, the consensus outcome was similar to that reached from a solely bacterial background. We noted similar successful removal of rotifer contamination from the genuine *Zoophagus insidians* (Zi) SCG metagenome (Davis et al., 2019). Taken together, these examples demonstrate moderate resilience of *SCGid*'s consensus approach to both bacterial and eukaryotic contamination.

SCGid can yield draft genomes ready for downstream analyses or partial solutions in need of further manual curation (Davis et al., 2019). This depends on the robustness of at least two of its integrated filtering approaches and the nature of planned downstream analyses. *SCGid* was conceived with these outcomes in mind. As such, it comes with a highly customizable set of options and utilities to augment the ways in which filtering decisions are made. *SCGid* can be iteratively rerun with different settings fast as it recycles the results of long-running steps. While the first run on an assembly can take 1–2 days, alternative filtered drafts can be produced by additional runs within minutes. An iterative *SCGid* workflow combined with tweaks to module and database configurations leads to increasingly refined filtering outcomes. By virtue of its consensus approach, *SCGid* has the potential to grow through the addition of novel filters leveraging variation in intergenic distance, intron length, etc.

SCG, despite its biases, weaknesses to contamination and inherent noise, generates genome-level sequence data for microbes that are inaccessible via standard approaches. This data contains fewer constituent genomes at higher coverage than analogous high-complexity metagenomes, but their identities are shrouded by unique biases. Where the goal of metagenomic binning may be the separation of many genomes, the goal of SCG is the separation of one or a few genomes from background contamination, endosymbionts, and noise. This sets SCG apart from metagenomics and in turn sets *SCGid* apart from other tools. *SCGid* takes prior expectations of taxonomy into account, using it as a central driver of filtering outcomes. *SCGid* is not intended for use in determining community composition or isolating hundreds of genomes from soil samples, but for filtering the genomes of the uncultured targets of sequencing efforts where whole community sequencing and brute-force metagenomics is unfeasible or extraneous. *SCGid* is made for SCG and is capable of mitigating its downsides in a fast, automated, and repeatable way. As such, it wields potential to unlock genome-enabled biology for the innumerable uncultured eukaryotes that depend on SCG for the acquisition of genome-scale data.

2.8 Literature Cited

Ahrendt, S.R., Quandt, C.A., Ciobanu, D., Clum, A., Salamov, A., Andreopoulos, B., Cheng, J.-F., Woyke, T., Pelin, A., Henrissat, B., Reynolds, N.K., Benny, G.L., Smith, M.E., James, T.Y., Grigoriev, I.V., 2018. Leveraging single-cell genomics to expand the fungal tree of life. Nature Microbiology 3. https://doi.org/10.1038/s41564-018-0261-0

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477.

Belda, E., Van Heck, R.G.A., Fraser, C., Klenk, H.-P., Sekowska, A., Vallenet, D., Martins, V.A.P., 2016. The revisited genome of Pseudomonas putida KT2440 enlightens its value as a robust metabolic chassis. Environ. Microbiol. 18, 3403–3424.

Cohen, M.F., Hu, P., Nguyen, M.V., Kamennaya, N., Brown, N., Woyke, T., Kyrpides, N., Holman, H.-Y., Torok, T., 2015. Genome Sequence of the Alkaline-Tolerant Cellulomonas sp. Strain FA1. Genome Announc. 3, e00646-15.

Davis, W.J., Amses, K.R., Benny, G.L., Carter-house, D., Chang, Y., Grigoriev, I., Smith, M.E., Spatafora, J.W., Stajich, J.E., James, T.Y., 2019. Genome-scale phylogenetics reveals a monophyletic Zoopagales (Zoopagomycota , Fungi). Mol. Phylogenet. Evol. 133, 152–163.

Dick, G.J., Andersson, A.F., Baker, B.J., Simmons, S.L., Thomas, B.C., Yelton, A.P., Banfield, J.F., 2009. Community-wide analysis of microbial genome sequence signatures. Genome Biol. 10. https://doi.org/10.1186/gb-2009-10-8-r85

Fisk, D.G., Ball, C.A., Dolinski, K., Engel, S.R., Hong, E.L., Issel-Tarver, L., Schwartz, K., Sethuraman, A., Botstein, D., Cherry, J.M., Project, S.G.D., 2006. Saccharomyces cerevisiae S288C genome annotation: a working hypothesis. Yeast 23, 857–865.

Gawad, C., Koh, W., Quake, S.R., 2016. Single-cell genome sequencing : current state of the science. Nat. Rev. Genet. 17, 175–188.

Gawryluk, R.M.R., del Campo, J., Okamoto, N., Strassert, J.F.H., Lukeš, J., Richards, T.A., Worden, A.Z., Santoro, A.E., Keeling, P.J., 2016. Morphological Identification and Single-Cell Genomics of Marine Diplonemids. Curr. Biol. 26, 3053–3059.

Ivanova, N., Sorokin, A., Anderson, I., Galleron, N., Candelon, B., Kapatral, V., Bhattacharyya, A., Reznik, G., Mikhailova, N., Lapidus, A., Chu, L., Mazur, M., Goltsman, E., Larsen, N., D’Souza, M., Walunas, T., Grechkin, Y., Pusch, G., Haselkorn, R., Fonstein, M., Dusko Ehrlich, S., Overbeek, R., Kyrpides, N., 2003. Genome sequence of Bacillus cereus and comparative analysis with Bacillus anthracis. Nature 423, 87.

Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M., Blaxter, M., 2013. Blobology : exploring raw genome data for contaminants , symbionts , and parasites using taxon-annotated GC-coverage plots. Front. Genet. 4, 1–12.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S.L., 2004. Versatile and open software for comparing large genomes. Genome Biol. 5.

Laczny, C.C., Sternal, T., Plugaru, V., Gawron, P., Atashpendar, A., Margossian, H.H., Coronado, S., Maaten, L.V.D., Vlassis, N., Wilmes, P., 2015. VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. Microbiome 3, 1–7.

Laetsch, D.R., Blaxter, M.L., Leggett, R.M., 2017. BlobTools : Interrogation of genome assemblies. F1000Res. 6, 1–16.

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

McInerney, J.O., 1998. GCUA: general codon usage analysis. Bioinformatics 14, 372–373.

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., Maréchal-Drouard, L., Marshall, W.F., Qu, L.H., Nelson, D.R., Sanderfoot, A.A., Spalding, M.H., Kapitonov, V.V., Ren, Q., Ferris, P., Lindquist, E., Shapiro, H., Lucas, S.M., Grimwood, J., Schmutz, J., Grigoriev, I.V., Rokhsar, D.S., Grossman, A.R., Cardol, P., Cerutti, H., Chanfreau, G., Chen, C.L., Cognat, V., Croft, M.T., Dent, R., Dutcher, S., Fernández, E., Fukuzawa, H., González-Ballester, D., González-Halphen, D., Hallmann, A., Hanikenne, M., Hippler, M., Inwood, W., Jabbari, K., Kalanon, M., Kuras, R., Lefebvre, P.A., Lemaire, S.D., Lobanov, A.V., Lohr, M., Manuell, A., Meier, I., Mets, L., Mittag, M., Mittelmeier, T., Moroney, J.V., Moseley, J., Napoli, C., Nedelcu, A.M., Niyogi, K., Novoselov, S.V., Paulsen, I.T., Pazour, G., Purton, S., Ral, J.P., Riaño-Pachón, D.M., Riekhof, W., Rymarquis, L., Schroda, M., Stern, D., Umen, J., Willows, R., Wilson, N., Zimmer, S.L., Allmer, J., Balk, J., Bisova, K., Chen, C.J., Elias, M., Gendler, K., Hauser, C., Lamb, M.R., Ledford, H., Long, J.C., Minagawa, J., Page, M.D., Pan, J., Pootakham, W., Roje, S., Rose, A., Stahlberg, E., Terauchi, A.M., Yang, P., Ball, S., Bowler, C., Dieckmann, C.L., Gladyshev, V.N., Green, P., Jorgensen, R., Mayfield, S., Mueller-Roeber, B., Rajamani, S., Sayre, R.T., Brokstein, P., Dubchak, I., Goodstein, D., Hornick, L., Huang, Y.W., Jhaveri, J., Luo, Y., Martínez, D., Ngau, W.C.A., Otillar, B., Poliakov, A., Porter, A., Szajkowski, L., Werner, G., Zhou, K., 2007. The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science 318, 245–251.

Mikhailov, K.V., Simdyanov, T.G., Aleoshin, V.V., Belozersky, A.N., 2016. Genomic survey of a hyperparasitic microsporidian Amphiamblys sp. (Metchnikovellidae). Genome Biol. Evol. 9, 454–467.

NCBI Resource Coordinators, 2017. Database Resources of the National Center for Biotechnology Information 45, 12–17.

Parra, G., Bradnam, K., Korf, I., 2007. CEGMA : a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23, 1061–1067.

Pati, A., Heath, L.S., Krypides, N.C., Ivanova, N., 2011. ClaMS: A Classifier for Metagenomic Sequences. Stand. Genomic Sci. 5, 248–253.

Pinard, R., de Winter, A., Sarkis, G.J., Gerstein, M.B., Tartaro, K.R., Plant, R.N., Egholm, M., Rothberg, J.M., Leamon, J.H., 2006. Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. BMC Genomics 7, 216.

Rinke, C., Lee, J., Nath, N., Goudeau, D., Thompson, B., Poulton, N., Dmitrieff, E., Malmstrom, R., Stepanauskas, R., Woyke, T., 2014. Obtaining genomes from uncultivated environmental microorganisms using FACS – based single-cell genomics. Nat. Protoc. 9, 1038–1048.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., Dodsworth, J.A., Hedlund, B.P., Tsiamis, G., Sievert, S.M., Liu, W.-T., Eisen, J.A., Hallam, S.J., Kyrpides, N.C., Stepanauskas, R., Rubin, E.M., Hugenholtz, P., Woyke, T., 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499, 431–437.

Roy, R.S., Price, D.C., Schliep, A., Cai, G., Korobeynikov, A., Yoon, H.S., Yang, E.C., Bhattacharya, D., 2014. Single cell genome analysis of an uncultured heterotrophic stramenopile. Sci. Rep. 4, 1–8.

Sedlar, K., Kupkova, K., Provaznik, I., 2017. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. Comput. Struct. Biotechnol. J. 15, 48–55.

Sieber, C.M.K., Probst, A.J., Sharrar, A., Thomas, B.C., Hess, M., Tringe, S.G., Banfield, J.F., 2018. Recovery of genomes from metagenomes via a depreplication, aggregation and scoring str. Nature Microbiology. https://doi.org/10.1038/s41564-018-0171-1

Sima, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO : assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212.

Stanke, M., Morgenstern, B., 2005. AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 33, 465–467.

The UniProt Consortium, 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45, D158–D169.

Ultsch, A., Moerchen, F., 2005. ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM. Technical Report Dept. of Mathematics and Computer Science, University of Marburg, Germany 46.

Waterhouse, R.M., Seppey, M., Sim, F.A., Ioannidis, P., 2017. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics Letter Fast Track. Molecular Biology and Evolution 35, 543–548.

Wu, Y.-W., Simmons, B.A., Singer, S.W., 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32, 605–607.

2.9 Figures

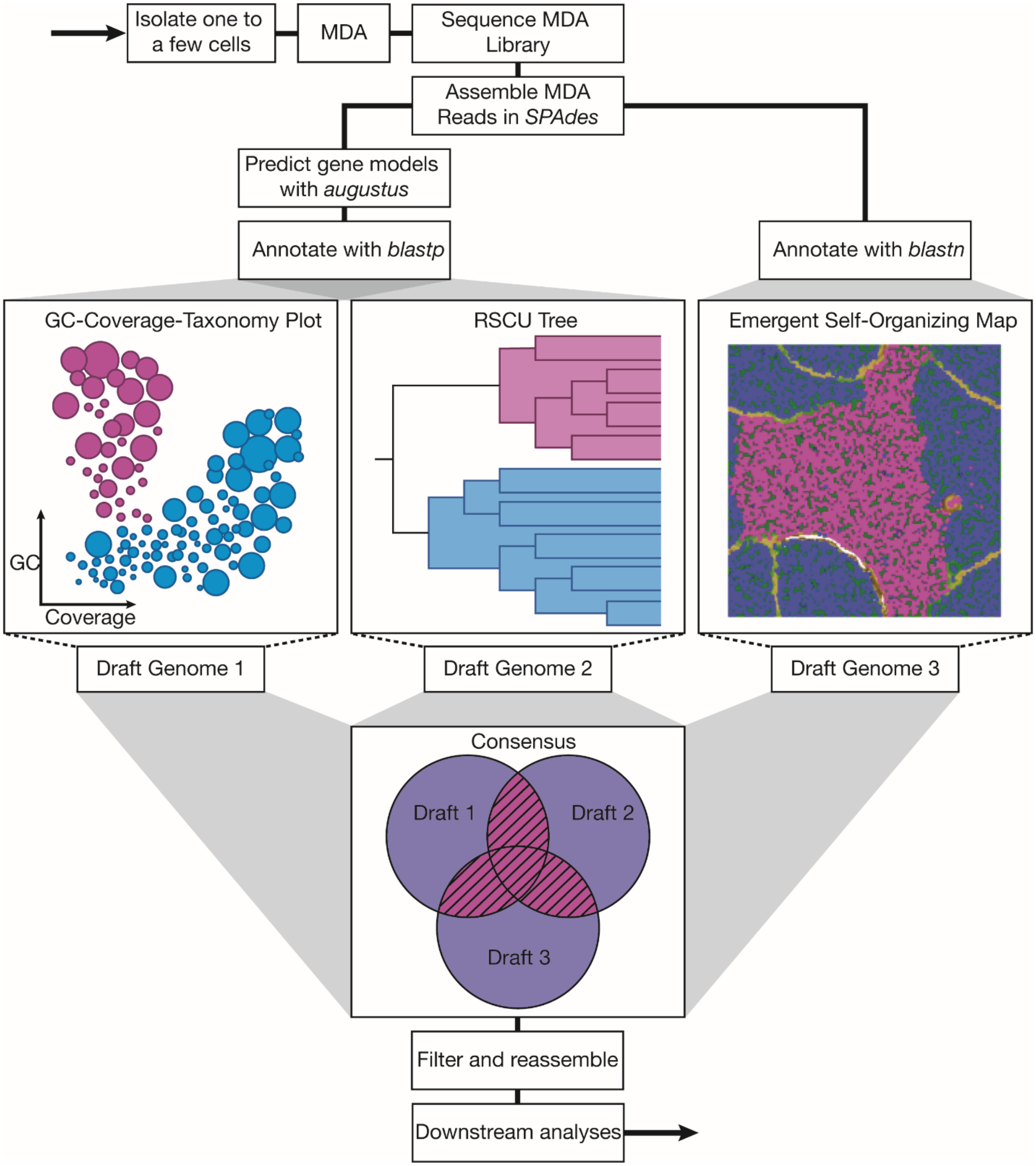


Figure 2.1. Flow chart showing overview of the automated SCGid workflow, from isolation of one to a few cells of an uncultured eukaryote to a consensus-filtered assembly. The initial SCG assembly is annotated with predicted protein models and taxonomy based on BLAST searches. Three draft genomes are independently predicted by separate binning methods, representing each method's inference of whether each contig belongs in the target genome (purple) or not (blue). Consensus takes these three draft genomes and identifies their overlaps, generating a final filtered draft assembly by majority rule that is at the interstices of the three independent methods, averaging over variation and reinforcing confidence in the final position of contigs. Parameters affecting filtering decisions at each step are highly customizable and the SCGid workflow is built to be run iteratively.

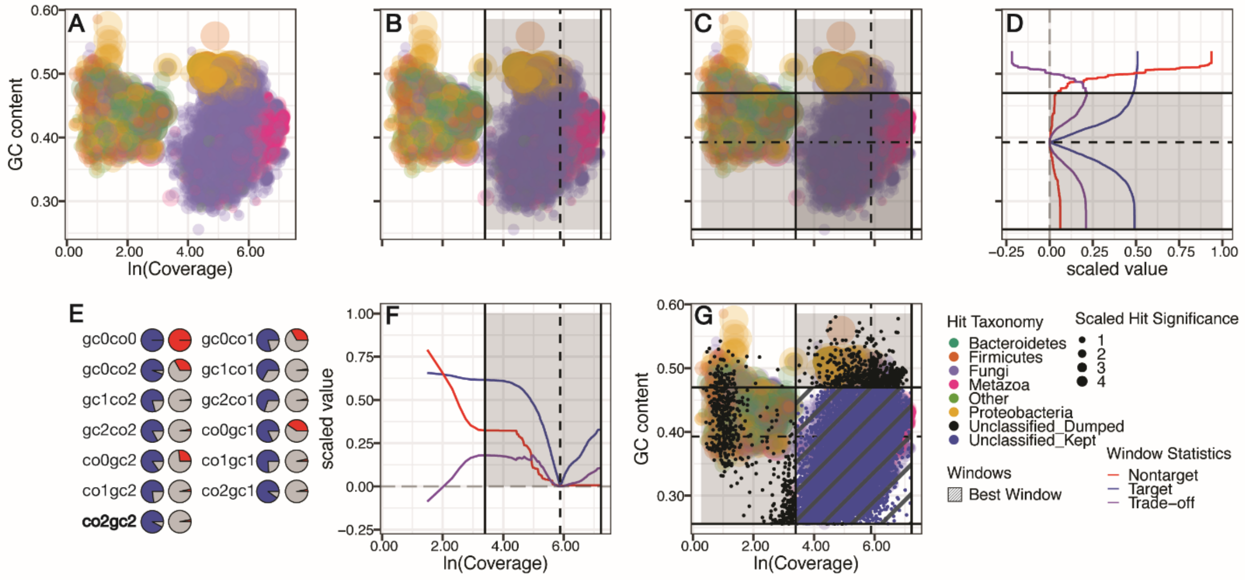


Figure 2.2. Plots visualizing the process of 2D GC-coverage window expansion over SCG data for the zoopagalean fungus *Stylopage hadra* (Sh). (A) GCT plot generated by SCGid, points are AUGUSTUS-predicted proteins plotted by GC and ln(coverage) of the containing contig, colors represent phylum-level taxonomic classification and size represents strength of the hit. (B, F) First round of window expansion along the coverage axis using method 'co2', window arms (grey box in B, F) originate from the target mean (dashed line in B, F). The final window arms are defined by maximization of a trade-off value (purple line in F) which balances the proportions of target (blue line in F) and nontarget (red line in F) in the window as it expands. (C, D) Second round of window expansion, along the GC axis using method 'gc2'. (E) All 13 window expansion methods and associated P_{tar} (blue) and P_{ntar} (red) values for final windows; note that only method 'co2gc2' is shown in (A–D), (F) and (G). The optimal window (crosshatched box in G) is defined by maximization of P_{tar} below set P_{ntar} stringency threshold. (G) GCT plot (from A) now overlaid with all unclassified contigs (black and blue points) showing optimal final window (crosshatched box). Unclassified contigs falling within the optimal final window (blue) are included in the final genome while the rest (black) are discarded.

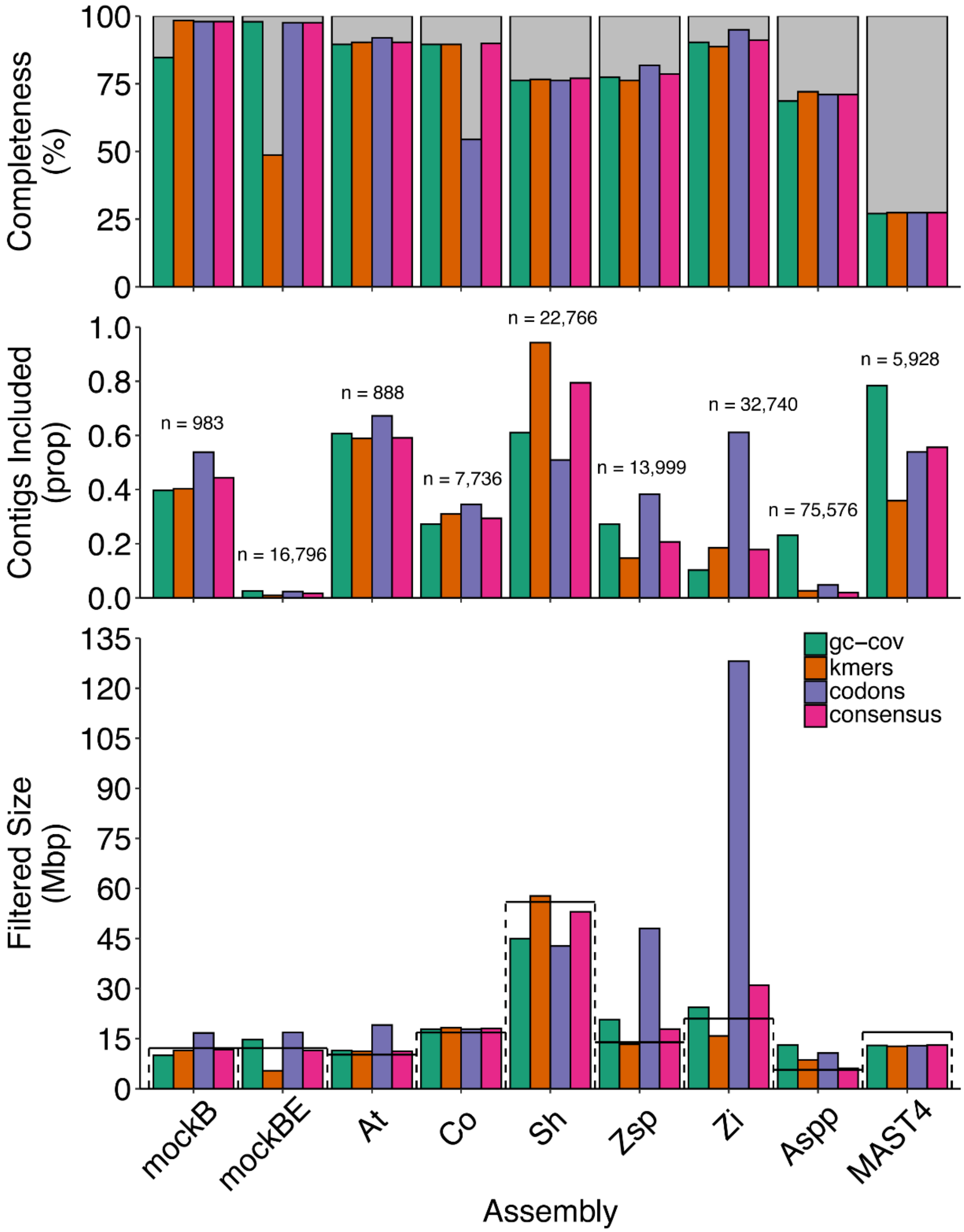


Figure 2.3. Set of grouped bar charts showing variation in filtering outcomes of the three different filtering approaches implemented in SCGid (green, orange and purple bars) and the averaging effect of consensus (pink bars). Filtered assemblies were often different in terms of cumulative filtered assembly size (bottom), proportion of initial assembly contigs persisting into the filtered draft (middle) and predicted genome completeness (top). Cumulative assembly sizes of the references are shown as dashed bars in the lower pane. The total number of contigs in each initial assembly is shown above bars in middle pane. Abbreviations are as follows: mockB, mock with bacterial contamination only; mockBE, mock with bacterial and eukaryotic contamination; At, *Acaulopage tetraceros*; Co, *Cochlonema odontosperma*; Sh, *Stylopage hadra*; Zsp, *Zoopage* sp.; Zi, *Zoophagus insidians*; Aspp, *Amphiamblys sp.* and MAST4, MAST4-like stramenopile.

Chapter 3: Novel obligate endohyphal bacterial symbionts of uncultured predatory fungi revealed by single cell sequencing implicate recent interphylum host switches.

3.1 Abstract

Fungi do not exist in a vacuum. They engage in diverse symbioses with diverse life forms, from harsh parasitic interactions with insects to mycorrhizal mutualisms with plants. One particularly intimate symbiosis that fungi are involved in finds bacteria colonizing the intracellular space of fungal cells. While many examples of intracellular colonization of fungi by bacteria are transient and spell bad news for the fungus, two lineages of bacteria, the *Burkholderia*-related and Mollicutes-related endobacteria, have made an apparently irreversible habit of it and appear to sometimes be mutualistic partners. These obligate endohyphal bacteria (EHB) are predominantly known to colonize the cells of plant-associated fungi in the Mucoromycota. This narrow host range has led to the framing of these EHB in the contexts of the fungal-plant interactions their hosts are involved in. Obligate EHB have not been recognized from earlier-diverging fungal lineages, but these lineages are poorly sampled and rarely screened for intracellular symbionts. In this study, we use single cell sequencing to detect and characterize novel EHB colonizing the cells of predatory fungi in the Zoopagomycota, an early diverging lineage where fungal-plant interactions are exceedingly rare. Using genome-scale phylogenetic and comparative analyses, we show that these novel EHB are members of EHB lineages known from plant-associated fungi in the Mucoromycota. Our phylogenetic reconstructions place these novel EHB nested within, not ancestral to, these lineages, implicating interphylum host switches in the history of these obligate EHB. This result requires a dramatic broadening of the concept of EHB in fungi and a reframing of their potential function that grows to include fungal-animal interactions.

3.2 Introduction

Symbioses are sustained and intimate interactions between two or more living organisms. The nature of symbiotic interactions (i.e., mutualistic, commensal, parasitic) can vary dependent of the evolutionary backgrounds of symbiotic partners as well as the biotic or abiotic environmental contexts within which interactions occur. Symbioses that involve the colonization of the intracellular space (i.e., the cytosol) of one organism by another (i.e., endosymbiosis) fall into a particularly intimate and invasive category of symbioses, the outcomes of which can vary with respect to the effect on the host.

Endosymbiotic interactions between bacteria and fungi, where bacterial cells colonize the cytosol of fungal cells, are becoming increasing appreciated (Araldi-brondolo et al., 2017; Pawlowska et al., 2018). The bacterial partners in bacterial-fungal endosymbiotic interactions, known collectively as endohyphal bacteria (EHB), constitute an artificial assemblage of bacteria that have been detected within fungal hyphae. According to a classification system established by Araldi-brondolo and colleagues, EHB can be divided into three major classes (i.e., EHB Class 1, 2, and 3) based on the phylogenetic affinities and functional classifications of their hosts as well as the genomic, evolutionary, and ecological characteristics of the EHB themselves (Araldi-brondolo et al., 2017). The Class 1 and Class 2 subdivisions of EHB refer to groups of endohyphal bacteria that are phylogenetically restricted to the Mollicutes (Mollicutes-related EHB, or MRE) or Burkholderiaceae (*Burkholderia*-related EHB, or BRE), respectively. EHB belonging to these classes are dependent on their hosts for basic metabolism, a dependence that is evidenced by reduced genomes that lack core metabolic genes and their resistance to axenic cultivation (Araldi-brondolo et al., 2017; Guo et al., 2020; Pawlowska et al., 2018; Torres-Cortés et al., 2015). That is, MRE and BRE are obligate or near-obligate in their association with fungi.

MRE and BRE have so far only been discovered in fungi affiliated with the phyla Ascomycota, Basidiomycota, and Mucoromycota (Araldi-brondolo et al., 2017). These intracellular symbiotic bacteria are heritable, and the current paradigm is that vertical transmission is the predominant mode of transmission from present to future hosts via fungal reproductive propagules (Araldi-brondolo et al., 2017; Pawlowska et al., 2018). A history of vertical transmission is evidenced in analytical comparisons of the phylogenies of fungal hosts and bacterial symbionts that suggest codiversification, versus independent lineage diversification (Mondo et al., 2012; Toomer et al., 2015). That said, there is evidence suggesting that horizontal transmission also plays an important role in the evolutionary history of MRE and, to a lesser degree, BRE (Mondo et al., 2012; Toomer et al., 2015). In MRE, the signal of codivergence with fungal hosts is supported in the recent past, with independent divergence better-supported at deeper nodes, suggesting that MRE engage in both vertical and horizontal transmission between hosts (Toomer et al., 2015). Horizontal transmission of MRE has been used to explain high observed interhost variability among symbionts at marker loci (e.g., 16S rDNA) as well as the apparent stability of MRE associations over evolutionary time despite their presumed parasitic nature (Toomer et al., 2015). In BRE associated with arbuscular mycorrhizal fungi, there are instances where recombination or host-switching via horizontal transmission of bacterial endosymbionts disrupts patterns of codiversification (Mondo et al., 2012). However, codivergence with hosts and predominant vertical transmission appear to be more important in the history of BRE (Araldi-brondolo et al., 2017; Mondo et al., 2012).

The distribution of these types of bacterial-fungal endosymbioses within the fungal kingdom is asymmetric, with a notable concentration in the Mucoromycota, a phylum where plant-associated and saprotrophic "molds" dominate (Araldi-brondolo et al., 2017). Although MRE and BRE are predominantly known from plant-associated mucoromycotan fungi (e.g., arbuscular mycorrhizal fungi, or AMF), they are also known to associate with later-diverging fungi in the Ascomycota and Basidiomycota (i.e., Dikarya) (Araldi-brondolo et al., 2017). MRE and BRE endosymbioses are entirely unknown from fungal lineages that diverged prior to the Mucoromycota (e.g., Zoopagomycota, Chytridiomycota, etc.) (Araldi-brondolo et al., 2017; Pawlowska et al., 2018).

The causal explanation of the asymmetric distribution of MRE/BRE endosymbioses in Fungi is not immediately apparent. Dikarya is the best-sampled lineage of fungi and so it seems unlikely that sampling biases can explain the paucity of associations known there. Perhaps intracellular colonization of fungi by bacteria is rare. However, the capacity of MRE and BRE symbionts to transfer horizontally between host fungi, to varying degrees, suggests that colonization events are not sufficiently rare to explain their asymmetric distribution (Mondo et al., 2012; Toomer et al., 2015). It could be that MRE/BRE colonization events are relatively common but tend to be evolutionarily unstable. This either requires that some unique characteristics of mucoromycotan fungi (e.g., intracellular environmental conditions, overlaps in environmental niche, etc.) facilitate stabilization of endosymbioses or that harboring endosymbiotic bacteria imparts context-dependent fitness benefits to the host (e.g., in the context of fungal-plant interactions). If fitness benefits are furnished to the host, it would imply that these EHB are mutualists, and while this may represent our understanding of BRE (in some contexts), it does not represent MRE, which are thought to lean parasitic (Araldi-brondolo et al., 2017). Even then, it is unlikely that the ancestors of extant BRE colonizing fungal cells were mutualistic, more likely being driven there by subsequent coevolutionary pressures.

Cellular organization of the fungal mycelium underwent a dramatic shift from coenocytic hyphae (i.e., without septa between cells) to septate hyphae. This transition sets Dikarya, where septate hyphae is the norm, apart from earlier-diverging lineages (e.g., Mucoromycota) where coenocytic hyphae are the norm. In these fungi, septa are vegetatively rare and only present in association with reproductive structures. For bacterial endosymbionts that rely, at least in part, on vertical transmission to colonize new hosts (e.g., via spores), regular septa impose a structural barrier that could preclude successful transmission and explain why MRE/BRE associations are rare in Dikarya. While presence or absence of regular septa might explain why these types of endosymbioses are abundant in Mucoromycota and rare in Dikarya, it does not explain why fungi from other early-diverging lineages (e.g., Zoopagomycota, Chytridiomycota, etc.), which also lack regular septa, apparently do not harbor MRE- or BRE-type endosymbionts (Araldi-brondolo et al., 2017; Pawlowska et al., 2018). Unlike Dikarya, these lineages are poorly sampled and even more rarely screened for EHB. If regular septa pose a major obstacle to the establishment of stable MRE/BRE endosymbioses in fungi, the coenocytic cellular organization of other early diverging fungal lineages should encourage their formation and stability.

The extant fungi that represent these earlier-diverging lineages are dramatically under sampled both in terms of raw diversity and when considering recent transformational advances in fungal biology (e.g., genomics) (James et al., 2020). Many members of these early-diverging lineages are often uncultured, meaning that they are difficult or impossible to grow under axenic conditions. This simple reality precludes isolation, genome sequencing, imaging, and other vital methods in the detection of EHB in fungi (Moran et al., 2008). Members of the early-diverging phylum Zoopagomycota are non-flagellated parasitic, mutualistic, or predatory fungi that engage in a variety of symbioses with lifeforms from across the tree of life (Spatafora et al., 2016). Divergence of the Zoopagomycota predates the adoption of persistent symbioses between plants and fungi (e.g. arbuscular mycorrhizae) and their symbiotic partners are usually small animals (e.g., insects, nematodes, amoeba), protozoa, or other fungi (Spatafora et al., 2016). Certain lineages within the Zoopagomycota (i.e., Zoopagales) are obligate or near-obligate predators of free-living nematodes, amoebae, and other protozoans, representing an independent derivation of "trapping" relative to the nematode-trapping ascomycetes (Davis et al., 2019; Drechsler, 1959, 1936). The uncultured status of these fungi complicates their study in contexts that have facilitated the discovery of novel EHB in Mucoromycota and related fungi (Bianciotto et al., 2003; Torres-Cortés et al., 2015). Single-cell genomics (SCG) can alleviate some of the restrictions imposed by the uncultured status of these early-diverging fungi by allowing individual cells collected from nature to serve as input to whole genome amplification (WGA) and sequencing (Davis et al., 2019; Mikhailov et al., 2016). Despite the contamination, biases, and noise introduced into sequence data by the WGA, SCG is a viable workaround that facilitates genome-enabled biological research, including the *in silico* detection of associated EHB, in uncultured fungi (Ahrendt et al., 2018; Amses et al., 2020; Davis et al., 2019; Mikhailov et al., 2016).

In our past work using SCG to generate draft genome sequences for uncultured fungi in the Zoopagomycota, we detected the presence of two bacterial symbionts with phylogenetic affinities to the MRE (RhopMRE in *Rhopalomyces* sp.) or BRE (*Mycoavidus* sp. SOG in *Stylopage hadra*) lineages (Davis et al., 2019). This is the first record of MRE and BRE symbionts occurring in the Zoopagomycota or, for the matter, any fungal lineage that diverged prior to the Mucoromycota. In this work, we present a full report of our findings including the genomes of these novel EHB, their relationship to other EHB, and discussion of the implications these findings have for our understanding of the evolution and maintenance of bacterial-fungal endosymbioses.

3.3 Methods

Sample Collection.

The uncultured status of the two zoopagomycotan host fungi, *S. hadra* and *Rhopalomyces* sp., precludes their cultivation under axenic conditions (Barron, 1973; Drechsler, 1936). To circumvent this central obstacle, each host fungus was cultivated in highly mixed *in vitro* microcosms that were generated by depositing the hand-homogenized upper organic layer of forest soils onto ¼x Corn Meal Agar (CMA: 2 g/L Corn meal infusion from solids, 15 g/L Agar) and supplemented with the nematode *Caenorhabditis elegans*, a compatible prey species. Microcosms were incubated in the dark at room temperature for 2–3 weeks and monitored for the growth of *S. hadra* or *Rhopalomyces* sp. on a dissecting microscope at 20–80x magnification. Microcosms that were identified as containing mycelia of either host species were subcultured onto new plates of either ¼x CMA or water agar (15 g/L Agar) and supplemented with autoclave-sterilized soil (25-minute exposure, 15 PSI, 121C), allowing maintenance of quasi-cultures over the course of this study. Forest soils from which we eventually identified and maintained *S. hadra* or *Rhopalomyces* sp. were collected from two locations at the University of Michigan Edwin S. George Reserve in Pinckney, Michigan, United States (CBSP: 42.461038, -84.022486; CB-Mid: 42.461324, -84.024143) as well as private property in Kalamazoo, Michigan, United States (JamHou2: 42.449890, -85.317860).

Single-cell DNA extraction, library preparation, and sequencing.

To generate genome-scale data for these fungi using SCG, one to a few cells of each fungus were collected by hand with UV-sterilized dental files and transferred to 2 µL of sterile PBS. Lysis and non-specific MDA amplification was conducted in line with the Qiagen REPLI-g Single Cell Kit (Cat No. 150343) to yield DNA at sufficient concentrations for Illumina library preparation. Sequencing libraries were prepared with either the Illumina Nextera XT (5 libraries from *S. hadra*) or Illumina Nextera Flex (2 libraries from *Rhopalomyces* sp.) library preparation kits and 150 bp paired-end reads were sequenced on the Illumina NextSeq platform. Additional isolates collected from different locations or at different times in Michigan were screened for *Mycoavidus* sp. colonization via PCR amplification of the 16S rDNA locus from MDA products followed by Sanger sequencing. PCR was conducted using the 16S universal primers BSF-8/20

(5′-AGAGTTTGATCCTGGCTCAG-3) and BSR-926/20 (5′-CCGTCAATTYYTTTRAGTTT-3′).

Single-cell metagenome assembly, filtering, and protein annotation.

We assembled SCG sequencing libraries using SPAdes v3.11.1 (Bankevich et al., 2012) in single-cell mode. Assemblies were identified as metagenomic and filtered into segregate bacterial and fungal bins based on the clustering by GC-content and coverage, tetranucleotide frequencies, and codon bias using SCGid v0.9b (Amses et al., 2020) by targeting contigs identified as Proteobacteria (*S. hadra*) or Mollicutes (*Rhopalomyces sp.*). Contigs in these bins were incorporated into draft genome sequences representative of the population of bacterial endosymbionts present in each sample. We excluded one filtered assembly in which we did not detect *Mycoavidus* sp. SOG and another where contamination by another bacterial genome (Bacteroidetes) reduced our confidence in filtering outcomes. This left us with two single-library assemblies for RhopMRE and three for *Mycoavidus* sp. SOG (SRX5014069, SRX5014895, SRX5014897).

Filtered single library assemblies ranged from 438,870–497,049 bp on 11–85 contigs and 1,771,688–1,875,133 bp on 102–186 contigs for RhopMRE (2 assemblies) and *Mycoavidus* sp. SOG (3 assemblies), respectively. We used *barrnap* v0.9 (Seemann, 2015) to extract 16S rDNA loci *in silico* from our five EHB assemblies. Alignment showed that the two sequences from each RhopMRE assembly were identical (100% identity, 100% query cover) and that 3 sequences from each *Mycoavidus* sp. SOG assembly were identical (100% identity, 100% query cover) (data not shown). Since extracted 16S sequences were identical across assemblies for each EHB (i.e., RhopMRE *Mycoavidus* sp. SOG) and their hosts were identified in microcosms derived from the same soil samples, we presumed the multiple assemblies to be derived from the same bacterial isolate. As such, we decided to coassemble the two RhopMRE and three *Mycoavidus* sp. SOG sequencing libraries together into two draft genome assemblies, one for each EHB isolate. Coassembly and subsequent metagenome filtering was conducted identically to the assembly of single-library drafts. Coassembled drafts were more contiguous in both cases but intermediate in length for RhopMRE and longer for *Mycoavidus* sp. SOG. We identified one contig in the filtered coassembly as entirely fungal (via BLAST protein searches) and excluded it under the assumption that it was filtered incorrectly. Completeness of these filtered drafts was estimated with *BUSCO* v4.1.4 (Seppey et al., 2019) using the *mollicutes_odb10* (151 models) or *burkholderiales_odb10* (688 models) COG databases. Protein coding genes were annotated in genome drafts with *prodigal* (Hyatt et al., 2010) using either translation table 4 (Mollicutes-specific) or 11 (general Bacteria) for RhopMRE and *Mycoavidus* sp. SOG, respectively.

Genome-scale phylogenetic analyses.

Prior to conducting genome-scale phylogenetic analyses to place these two novel EHB in the context of known EHB, we assembled a representative taxon set for each major lineage of Bacteria (i.e., Mollicutes and Burkholderiaceae) by downloading available genome-level protein datasets from NCBI GenBank. In addition, we downloaded datasets corresponding to all MRE and BRE bacteria for which genome-scale data is available. Including our two novel draft genomes, our final datasets consisted of 149 Mollicutes-related taxa (including three outgroup taxa) and 54 Burkholderiaceae-related genomes (with Burkholderiaceae-related *Polynucleobacter* serving as outgroup) (Tables B1 and B2).

Genome-scale phylogenetic analyses were conducted in parallel for each of these two major groups of bacteria using appropriate sets of core orthologous genes (COGs) from the *BUSCO odb10* database (Seppey et al., 2019). That is, *mollicutes_odb10* (151 COGs) and *burkholderiales_odb10* (688 COGs) served as the marker sets for genome-scale phylogenetic analyses of Mollicutes and Burkholderiaceae, respectively. Homologous protein sequences were extracted, aligned, and trimmed from the predicted proteomes using a standard approach. Briefly, marker HMMs were downloaded as is from *odb10* and combined with *hmmpress* (Eddy and HMMER development team, 2015). Predicted proteins were searched against this multi-HMM marker file with *hmmsearch*. The resulting domain tables were filtered such that the predicted protein with the most significant hit (as determined by *hmmsearch*) to each marker HMM was selected for inclusion in marker gene alignments. In cases where the most significant hit to the HMM from a proteome was below the *BUSCO* internal score cutoff for that marker, no protein sequence was selected, and gaps were inserted instead. The resulting FASTA files, which contained up to 149 (Mollicutes) or 54 (Burkholderiaceae) sequences, were aligned with *Mafft* v7.310 (Katoh and Standley, 2013). To determine the most applicable substitution model with which to conduct concatenated ML analyses, each marker alignment was run individually through *ModelFinder* in *IQ-TREE* v2.0.5 (Kalyaanamoorthy et al., 2017; Minh et al., 2020). All individual marker gene alignments were then concatenated to yield a single multiple sequence alignment that contained 149 (Mollicutes) and 54 (Burkholderiaceae) sequences, with gap sequences inserted for missing genes. Maximum likelihood concatenated trees were computed in *IQ-TREE* using the most popular best substitution model among individual marker alignments (LG+F+R6 for MRE tree, LG+I+G4 for BRE tree) with 10,000 ultrafast bootstrap replicates.

Detection of horizontally transferred genes.

To identify genes that were good candidates for some past horizontal gene transfer, we conducted protein BLAST searches (max_target_seqs: 100, evalue: 10) for each predicted protein encoded in each novel EHB draft genome against a custom database that incorporated the entirety of the UniRef protein database collapsed at 90% identity (i.e., UniRef90) (Suzek et al., 2015) supplemented with the predicted proteomes of 27 early-diverging fungi from Chytridiomycota, Blastocladiomycota, and Zoopagomycota (Davis et al., 2019). Prior to BLAST searches, we purged the UniRef90 database of all entries corresponding to known MRE or BRE proteomes. We then calculated alien index (AI) scores for the top 100 blast hits to each predicted protein according to the generalized formula $ln((E_{Domestic} + e^{-200}) – (E_{Alien} + e^{-200}))$, where $E_{Domestic}$ is the e-value of the best BLAST hit to the “expected” domain (i.e., Bacteria) and E_{Alien} is the e-value of the best BLAST hit to the “alien” domain (i.e., Eukaryota) (Alexander et al., 2016). Positive AI scores indicate that a protein is more similar to a eukaryotic protein than it is to a bacterial protein, thereby identifying genes that could have been horizontally transferred. We used an AI cutoff of 20, meaning that alien domain hits had to be ~20 orders of magnitude stronger than the best hit to the expected domain to be considered a HGT candidate (Alexander et al., 2016). To confirm that each HGT candidate was not identified as a result of a spurious BLAST hit, we computed ML gene trees for each candidate that included the top 100 best BLAST hits, the best blast hit to each candidate in other EHB, and the candidate sequence itself. Trees were computed in *IQ-TREE* using the best model for each alignment as determined by *ModelFinder* in *IQ-TREE*. Gene trees were visualized and annotated with *ggtree* (Yu et al., 2018, 2017) and *tidyverse* (Wickham et al., 2019) in R. To verify that HGT candidates were not identified on contigs derived from SCG-related contamination, we constructed genetic maps of their parent contigs to ensure that candidates occurred in otherwise-bacterial genomic regions. Genetic maps were visualized with *ggplot2* (Wickham, 2016) in R.

Comparative genomics.

We annotated our EHB assemblies, and all other genome included in this study, with *prodigal*. Translation table code 4 or 11 was used to annotate genomes from the Mollicutes or elsewhere, respectively. We functionally annotated predicted proteomes using *interproscan* v5.47-82.0 (Jones et al., 2014). PFAM domains were extracted from *interproscan* outputs using custom scripts and *tidyverse* in R. Phylogenetically scaled PCA ordinations were computed with *phytools* (Revell, 2012) and visualized with *ggplot2*, in R. Trees were visualized and annotated with *ggtree*.

3.4 Results

BRE colonization of S. hadra detected in multiple strains isolated from the Midwestern United States.

Although we only generated genome-scale data for one *Mycoavidus*-harboring isolate of *S. hadra* from Pickney, Michigan (*S. hadra* SOG: 42.461324, -84.024143), we detected *Mycoavidus* spp. in association with other isolates of *S. hadra* collected from different locations and at different times. Novel EHB were detected via PCR amplification of the 16S rDNA locus from MDA amplification products. One of these isolates was collected outside of Kalamazoo, Michigan (*Mycoavidus* sp. JamHou2: 42.449890, -85.317860) around the same time of our collection and sequencing of *S. hadra* SOG (Figure 3.1C, purple asterisk). The other novel EHB colonizing *S. hadra* (*Mycoavidus* sp. CBSP2: 42.461038, -84.022486) was collected nearby the site where *S. hadra* SOG was collected in Pinckney, Michigan (i.e., CB-Mid: 42.461324, -84.024143), but was collected several years afterwards (Figure 3.1C, red asterisk). Both isolates harbored strains of *Mycoavidus* sp. that were more closely related to *Mycoavidus* sp. SOG than other members of *Mycoavidus* detected colonizing fungi in the *Mortierellomycotina* (Figure 3.1B**)** Together, these spatially and temporally separate detections suggest that BRE colonization of *S. hadra* is common in nature and has probably only gone undetected due to rare screening of uncultured zoopagalean fungi for EHB.

Phylogenetic analyses place endohyphal bacterial symbionts of zoopagalean fungi in well-established groups of EHB.

We generated draft genome sequences for two novel EHB found in association with fungi in the Zoopagomycota. The draft genome of RhopMRE consists of 456,991 bp on 16 contigs, encodes 483 *prodigal*-predicted proteins, and is estimated to be 47.02% complete (*BUSCO, mollicutes_odb10*). This is in line with completeness estimates for the seven publicly available MRE genomes, which we independently estimated as ranging from 31.13%–42.38% complete. This means that despite its short cumulative length, the RhopMRE genome has retained the highest number of Mollicutes COGs of any MRE genome sequenced to date. The draft genome of *Mycoavidus* sp. SOG consists of 1.92 Mbp on 135 contigs, encodes 2,014 *prodigal*-predicted proteins, and is estimated to be 87.94% complete (*BUSCO, burkholderiales_odb10*). This is in line with the estimated completeness of other sequenced BRE genomes, which we measured to range from 23.54%–98.98% complete.

Our rDNA and genome-scale phylogenetic analyses of Mollicutes and Burkholderiaceae place these novel endohyphal symbionts of zoopagalean fungi in well-established groups of EHB (Figure 3.1A,B, Figure 3.2A,B). We resolve the *Mycoplasma*-related endohyphal bacterium present in *Rhoplalomyces* (i.e., RhopMRE) nested within the MRE II clade alongside bacteria known only as endohyphal symbionts of arbuscular mycorrhizal fungi (Glomeromycotina) (Figure 3.1A). We resolve the larger MRE group (i.e., MRE I and MRE II) as sister to a relatively early-diverging clade of Mollicutes that is largely composed of bacteria from the genera *Spiroplasma*, *Mesoplasma*, and *Entomoplasma* (Figure 3.2A). We resolve the *Burkholderia*-related endohyphal symbiont of *S. hadra* (i.e., *Mycoavidus* sp. SOG) in an ancestral position and sister to a clade of BRE known only as endohyphal symbionts of fungi in the Mortierellomycotina (Figure 3.1B, Figure 3.2B). Interestingly, this novel BRE is nested within a larger clade of BRE that is composed of Candidatus Glomeribacter, and *Mycoavidus* rather than ancestral to the entire group, which describes the relationship between Zoopagomycota and Mucoromycota (Spatafora et al., 2016). Considering the entirety of sequenced BRE genomes, we resolve a paraphyletic assemblage of BRE composed of two monophyletic clades, a Ca. Glomeribacter–*Mycoavidus* clade and a *Mycetohabitans* clade, which occupy an ancestral position relative to the latest-diverging clades of Burkholderiaceae or is sister to the *Burkholderia–Paraburkholderia* clade, respectively.

RhopMRE has acquired and retained genes horizontally transferred from fungi, animals, and protozoans.

Our AIS-based approach identified 15 genes in the genome of RhopMRE as good candidates for some past horizontal transfer (i.e., AIS >= 20). The predicted functions of these candidate genes were diverse, including AIG1 (AvrRpr2-Induces gene 1), lectins, proteases, and nucleic acid metabolism-related genes, among others with known and unknown functions (Table B3). Following identification, we examined individual gene trees to exclude HGT candidates with uninformative, incomplete, or artifactual gene trees. This led to the exclusion of 4 candidates. Based on the taxonomy and clustering pattern of tips in annotated gene trees, we inferred the putative origins of HGT candidate genes to be either animal or fungal (Figure 3.3; grey insets). Surprisingly, gene trees suggested that 4/11 HGT candidates are bacterial in origin and were only detected by our approach due to potential horizontal transfer out of, not into, MRE. Realizing this, we were left with a set of 7 HGT candidates encoded in the RhopMRE genome that were of putative fungal, animal, or protozoan origin. To confirm these 7 genes were not encoded on contamination-derived contigs present in our assembly, we constructed genetic maps of their parent contigs. These maps clearly place the 7 HGT candidates in the RhopMRE genome on large contigs surrounded by bacterial genes (Figure 3.3; HGT candidates indicated by black arrows). By virtue of their high AIS scores, phylogenetic clustering with eukaryotes in gene trees, and conspicuous positions on otherwise bacterial contigs, we consider these genes to be true cases of HGT from eukaryotes into the RhopMRE genome.

To assess the prevalence of potential protein homologs of these HGT candidates in other sequenced MRE, we conducted blast searches (evalue: 1e^{-50}) against the seven publicly available MRE genomes and five outgroup taxa. Our searches revealed a patchwork distribution of HGT candidate homologs, where some genes are present in all MRE genomes and others are restricted to particular clades or even individual genomes (Figure 3.4, green and orange columns). HGT candidates with a putative fungal origin evidenced wide and narrow distributions within MRE while those putatively originating in nonfungal eukaryotes were only detected in RhopMRE

(Figure 3.4, green versus orange columns). Although not HGT genes in the context of RhopMRE, we were also interested in the possible origins of genes apparently transferred from MRE to fungi (Figure 3.4, blue columns). While the origin of two of these genes is difficult to identify because of their wide prevalence in MRE and Mollicutes, the unique presence of the other two in the RhopMRE genome suggests RhopMRE as the source of transfer (Figure 3.4, left two blue columns versus right two blue columns).

Given the sole presence of HGT candidates with putative nonfungal eukaryote origin in the RhopMRE genome, we wanted to further confirm their identity as such by ruling out contamination as a potential source. We annotated these two genes as a hypothetical protein from the amoeba *Naegleria gruberi* (37.76% identical, e-value: 1e^{-117}, 99% query coverage) and a mexicain cysteine peptidase (MEROPS class C01) from the mite *Galendromus occidentalis* (57.01% identical, e-value: 1e^{-94}, 89% query coverage), respectively. To confirm that the two parent contigs were unequivocally bacterial, we examined the annotations for the other proteins encoded on these contigs. For the parent contig of the hypothetical amoebae gene, 33/42 encoded proteins were most similar to other bacterial proteins. Of the 9 that were not, one was identified as fungal (1.01e^{-27}), one was identified as Archean (7.93e^{-13}), and the other 7 had no significant hit (Figure 3.3, bottom contig). For the parent contig of the mexicain, 32/38 encoded proteins were most similar to other bacterial proteins. Of the 6 that were not, one was identified as fungal (5.3e^{-48}) and the others had no significant hit (Figure 3.3, third contig down). Based on significant bacterial composition of these contigs, we ruled out contamination as the source of these HGT candidates This signal is particularly interesting given the host ranges of zoopagalean fungi and their relatives, which include nematodes, amoeba, rotifers, and mites (Drechsler, 1959, 1936; Wekesa et al., 2007; Whisler and Travland, 1974). We repeated our HGT-detection analyses for *Mycoavidus* sp. SOG using the same approach as above but did not identify any good candidates for HGT into the *Mycoavidus* sp. SOG genome.

Comparative genomics of novel EHB reveals typical EHB genomes.

To compare the genome of *Mycoavidus* sp. SOG to other sequenced BRE and genome of non-BRE Burkholderiaceae, we annotated PFAM domains in the *prodigal*-predicted proteomes of all 54 taxa included in our phylogenomic reconstructions. We annotated 4,764 unique PFAM domains across our entire dataset, 2,520 of which were represented in BRE proteins, and 126 which were unique to BRE (Figure 3.5A). Interestingly, the predicted proteomes of BRE genera contain different unique PFAM domains relative to each other. We identified 58, 46, and 9 PFAM domains that were unique to *Mycoavidus*, Ca. Glomeribacter, and *Mycetohabitans*, respectively (Figure 3.5A). In general, *Mycoavidus* sp. SOG has a typical BRE genome that is similarly reduced in size relative to other non-MRE Burkholderiaceae (excluding *Polynucleobacter*) (Figure 3.6A). In addition to being shorter than free-living Burkholderiaceae, BRE genomes encode fewer unique PFAM domains, suggesting reduced functionality on the genome scale (Figure 3.6A). We also clustered taxa by their proteome-wide PFAM domain profiles in a phylogenetically scaled PCA (Figure 3.6B). In general, non-BRE Burkholderiaceae genomes were more tightly clustered along the PC axes than BRE genomes. Within BRE, we noted tighter clustering of Ca. *Glomeribacter* (n = 4) and *Mycetohabitans* (n = 2, too few points for ellipse) relative to *Mycoavidus* (n = 4), which shows the largest variance on the PC axes. The top ten most influential PFAM domains driving separation along the PC axes were variable in function, including transposases, a Formylglycine-generating sulfatase, a DNA ligase, and an RNase, among others (Table B4, left).

To compare the genetic repertoire of RhopMRE with previously described MRE and non-MRE Mollicutes, we annotated PFAM domains in the *prodigal*-predicted proteomes of all 149 genomes included in our phylogenomic trees of Mollicutes. We annotated 3,781 unique PFAM domains across our entire MRE dataset, 499 of which were represented in MRE, and 45 of which were unique to MRE genomes (Figure 3.5B). Analogous to BRE, MRE genomes are generally reduced in size and encode fewer unique PFAM domains than non-MRE Mollicutes (Figure 3.6D). Again, we clustered genomes by their proteome-wide PFAM profiles using a phylogenetically scaled PCA (Figure 3.6C). Clustering of MRE relative to non-MRE Mollicutes was less distinct than our Burkholderiaceae ordination (i.e., Figure 3.6B), with MRE overlapping almost entirely with the uncertainty ellipse for non-MRE Mollicutes. The top ten most influential PFAM domains driving separation along the PC axes included ATPases, GTPases, and ribosome-associated proteins, among domains of other functions (Table B4, right).

To investigate the capacity for MRE, including RhopMRE, to carry out autonomous energy production via cellular respiration, we identified PFAM domains associated with cellular respiration enzymes and compared their abundance in *prodigal*-predicted proteomes of RhopMRE, the seven publicly available MRE genomes, three non-MRE Mollicutes, and *Staphylococcus aureus*. Like other MRE genomes, RhopMRE has no occurrences of the ten cellular respiration-related domains that we searched for, suggesting that RhopMRE is dependent on *Rhopalomyces* sp. for important aspects of basic metabolism (Figure B1).

3.5 Discussion

We used SCG to sequence the genomes of two novel EHB detected in association with early-diverging animal-associated fungi of the Zoopagomycota, *S. hadra* and *Rhopalomyces* sp. Our 16S and genome-scale phylogenetic reconstructions place RhopMRE (MRE) and *Mycoavidus* sp. SOG (BRE) nested within canonical MRE and BRE lineages (Figure 3.1A,B; Figure 3.2A,B). Our well-supported phylogenies leave little room for doubt that these novel EHB are close relatives of EHB known predominantly from plant-associated fungi (Bianciotto et al., 2003; Guo et al., 2020; Naito et al., 2017, 2015; Sun et al., 2019; Torres-Cortés et al., 2015). It is important not to underestimate the exacerbating effect that the SCG-associated MDA reaction can have on contamination when inputs are derived from complex environments (Amses et al., 2020; Davis et al., 2019; Gawad et al., 2016; Mikhailov et al., 2016). However, it is unlikely that detection of these novel EHB is due to contamination. First, we detected *Mycoavidus* sp. SOG in multiple cells of *S. hadra* collected from regionally disparate soils in Michigan via 16S amplification and sequencing (Figure 3.1C). Second, we quasi-cultivated *S. hadra* colonized by *Mycoavidus* sp. from multiple locations and several years apart (Figure 3.1C). Third, we *in silico*-isolated nearly complete EHB genomes from multiple SCG sequencing libraries derived from the same soil sample. Finally, SCG metagenomic assemblies were never found to contain sequence data belonging to fungi other than are our targets (i.e., *Rhopalomyces* sp. and *S. hadra*) despite their common presence in our microcosms (data not shown). This final point negates the possibility that these novel EHB were derived from cells of a known host fungus (e.g., *Mortierella*). Based on these multiple lines of evidence, we consider RhopMRE and *Mycoavidus* sp. SOG codified EHB associates of *Rhopalomyces* sp. and *S. hadra*, respectively.

The nested placements that we resolve for RhopMRE and *Mycoavidus sp. SOG* within canonical EHB clades suggest that colonization of these hosts are the results of horizontal transmission and host-switching from mucoromycotan fungi (Figure 3.1A,B; Figure 3.2A,B). Although horizontal transmission of both MRE and BRE endosymbionts between fungal hosts is recognized in the history of EHB, documented cases of such tend to be between fungi in the same family or order (Mondo et al., 2012; Toomer et al., 2015). On an entirely different scale, our results implicate interphylum host switches from the Mucoromycota to the Zoopagomycota. The divergence times between the putative source and destination hosts in our study is significantly more vast than in documented examples of horizontal transmission of obligate EHB (i.e., MRE and BRE) (Mondo et al., 2012; Spatafora et al., 2016; Toomer et al., 2015).

Although inferring the timing and source of these horizontal transmissions is complicated by the absence of records of MRE or BRE endosymbionts colonizing zoopagalean fungi (i.e., this is the first), our results allow us to make some predictions. In the case of RhopMRE, both our genome-scale and 16S phylogenies suggest that horizontal transmission occurred from an arbuscular mycorrhizal fungus since RhopMRE is nested in a clade of AMF-associated MRE. The source host of the lineage including *Mycoavidus* sp. *SOG* is less clear since our phylogenomic reconstructions suggest this novel EHB diverged at some time after the split between *Mortierella*-associated *Mycoavidus* and AMF-associated Ca. Glomeribacter. That said, the ancestral position of *Mycoavidus* sp. SOG relative to other *Mycoavidus* strains suggests that it was horizontally transferred from a fungus in the Mortierellomycotina, as opposed to AMF. Unlike for MRE, at the time of writing, there are no BRE 16S sequences available outside of those strains with sequenced genomes.

While our work clearly demonstrates that EHB associate with fungi in the Zoopagomycota, it remains unclear how many others there are and if Zoopagomycota-associated EHB form monophyletic clades within their respective lineages of EHB, the extant examples of singular host-switches from mucoromycotan to zoopagomycotan hosts. Alternatively, Zoopagomycota-associated EHB could be paraphyletic, the results of more common, independent host switches. Based on trends in codivergence described elsewhere for obligate EHB of mucoromycotan fungi and the altered selective pressures that an abrupt interphylum host switch must impose, rare transmission events followed by codiversification with novel hosts seems the appropriate null hypothesis to test as more Zoopagomycota-associated EHB are discovered (Mondo et al., 2012; Toomer et al., 2015).

Using an AIS-based approach, we identified seven RhopMRE genes that were likely horizontally transferred from fungi, animals, or protozoans (Figure 3.3) (Alexander et al., 2016). This confirms that, like other MRE, the genome of RhopMRE is mosaic (Naito et al., 2015; Sun et al., 2019; Torres-Cortés et al., 2015). The majority of these HGT candidates (i.e., 5/7) appear to originate from fungi in the Mucoromycota, with the other two originating from a protozoan and an animal, respectively (Figure 3.4). The predicted functions of most of these genes were either unknown or previously identified in classes identified by other authors as HGT candidates originating from fungi or otherwise important in MRE endosymbiotic biology (Naito et al., 2015; Sun et al., 2019; Torres-Cortés et al., 2015).

We identified three HGT candidates that contained AIG1 domains, which have known function in antimicrobial defense in plants (Reuber and Ausubel, 1996). AIG1 domain-containing proteins are clearly present in mucoromycotan genomes (Figure 3.3), but phylogenetic analyses conducted elsewhere suggest that the ancestor of fungal AIG1-containing genes originates in nonfungal eukaryotes, perhaps amoebae (Torres-Cortés et al., 2015). We also identified an HGT candidate that encoded a Jacalin-type lectin domain. Like AIG1, the Jacalin-type lectin is best-known as an agent of plant antimicrobial defense (Esch and Schaffrath, 2017). The importance of lectins in host-parasite interactions is broadly appreciated, where their high-specificity carbohydrate-binding functionality enables recognition of, for example, carbohydrates displayed on the nematode cuticle (Andersson et al., 2014; Lai et al., 2014; Rosenzweig et al., 1985; Tunlid et al., 1992). Lectins have not been previously identified in sets of HGT candidates from MRE genomes (Naito et al., 2015; Sun et al., 2019; Torres-Cortés et al., 2015). This makes our discovery of this lectin-containing gene in the genome of RhopMRE, an intracellular endosymbiont of nematophagous *Rhopalomyces* sp., novel and compelling. Without transcriptional data it is impossible to say for sure whether this protein is expressed *in situ*, however the retention of any gene in the highly reduced and volatile landscapes of MRE genomes suggests its active use. It is tempting to hypothesize that this horizontally transferred lectin-containing gene functions in the nematophagous lifestyle of *Rhopalomyces* sp. (Barron, 1973).

We also identified two HGT candidates that appear to have originated from nonfungal eukaryotes, one of unknown function from an amoeba and one mexicain protease from arthropods (Figure 3.4, orange columns). While the source and functions of the former is ambiguous, the latter is clearly a mexicain cysteine protease from an arthropod. This is made clear by the fact that our gene tree contains neither bacterial nor fungal tips but is instead composed entirely of mexicain proteins from arthropods, nematodes, and chordates, where the RhopMRE gene clusters with a mexicain from *Galendromus occidentalis* (Figure 3.6, black asterisk; Figure 3.7). Since neither *Rhopalomyces* nor its closest relatives predate or parasitize arthropods, the horizontal transfer path of this gene to RhopMRE is unclear. We interpret the presence of this mexicain gene in the genome of RhopMRE in one of three ways: either (i) it was transferred to an ancestor of RhopMRE engaged in endosymbiotic interactions with a entomopathogenic host fungus (i.e., not *Rhopalomyces*) and subsequently inherited vertically by RhopMRE, (ii) it was recently transferred to RhopMRE prior to a host switch from an entomopathogenic fungus to *Rhopalomyces* sp., or (iii) or it was transferred to RhopMRE from a nematode host, indicating that our gene tree does not accurately reflect its relationship to other animal mexicains. Based on the increased propensity for host switches in MRE (i.e., compared to BRE), the relatively low phylogenetic distance between the RhopMRE mexicain and other animal mexicains, and the relatively high support for nodes in our mexicain gene tree, we see the second path as the most likely (Toomer et al., 2015). That is, that RhopMRE underwent a host switch from an entomopathogenic fungus following horizontal transfer of this mexicain from an arthropod. This would suggest the existence of a so far unknown third MRE associate of some arthropod-associated fungus, perhaps in the Entomophthorales, an early-diverging order of entomopathogenic fungi (Gryganskyi et al., 2012).

Our high-level comparative genomic analyses demonstrate that Zoopagomycota-associated EHB are like previously sequenced EHB genomes in terms of size (i.e., reduced) and diversity of function (Figure 3.6). We note some marked differences in PFAM domain profiles between BRE genera with more PFAMs unique to *Mycoavidus* (58) or Ca. Glomeribacter (46) than

Mycetohabitans (9) despite the latter constituting an independent lineage of BRE (Figure 3.5A). This could be explained by the nonoverlapping host ranges of these sister lineages of BRE. On a larger scale, our ordination shows that endosymbiotic members of the family are more diverse in their complement of PFAM domains than their free-living relatives (Figure 3.6B). This variation exemplifies a history of dependence on endosymbiosis in this derived group of EHB and could be due to loss of ancestral genes, gain of novel genes through HGT, rapid rates of evolution, or a combination of all three. First, the genomes of *Mycoavidus* sp. SOG and its BRE relatives are generally reduced relative to their free-living ancestors (Figure 3.6A, non-purple bars/points versus purple points/bars). This reductive trajectory is common in diverse lineages of obligate endosymbionts colonizing diverse hosts, and suggests genetic complementation and metabolic exchange between partners (Araldi-brondolo et al., 2017; Moran et al., 2008; Torres-Cortés et al., 2015). Second, BRE are known to have experienced genomic rearrangements and expansions or retractions in the occupancy of functional gene classes, both of which suggest elevated rates of genomic change (Figure 3.5A), although there are other lineages of obligate endosymbiotic bacteria that exhibit more extreme cases, such as MRE or those that colonize insects (Guo et al., 2020; Moran et al., 2008; Torres-Cortés et al., 2015). Finally, we did not detect a clear signature of HGT in the genome of *Mycoavidus* sp. SOG. This agrees with other studies in BRE and is consistent with patterns observed in separate lineages of obligate bacterial endosymbionts (i.e., insect-associated), but differs dramatically from the genomic mosaicism that characterizes MRE (Moran et al., 2008; Torres-Cortés et al., 2015). The variation we observe in our ordination could be explained by stochastic loss of ancestral genes in different lineages of BRE or gain through evolutionary resolution of lineage-specific genomic rearrangements.

We observed similar reductive evolution, but less dramatic functional separation, when comparing MRE genomes to their non-MRE ancestors (Figure 3.6C,D). In our ordination, MRE forms a separate, but mostly overlapping, cluster relative to non-MRE Mollicutes (Figure 3.6C). Given the background of advanced genome reduction from which MRE emerged (i.e., Mollicutes), it is possible that sustained loss of functional gene classes runs up against an unsustainable bare minimum (Naito et al., 2015). MRE genomes, including RhopMRE, are significantly reduced in size and gene complement compared to non-MRE Mollicutes (Figure 3.6D, yellow points/bars versus red points/bars), but our PFAM-based ordination cannot detect gene loss that is not accompanied by functional loss. Still, it is surprising that rampant uptake of horizontally transferred genes in MRE genomes does not drive more separation, though this could be due to the coarse grain of our PFAM-based analyses.

Our resolution of the MRE clade in the Mollicutes is at odds with past placements based on either rDNA loci or sets of single copy housekeeping genes (Naito et al., 2017, 2015; Sun et al., 2019; Torres-Cortés et al., 2015). In general, trees based on 16S rDNA sequences resolve MRE as sister to the later-diverging *Mycoplasma hominis* group (Hominis Cluster in Figure 3.2A) (Naito et al., 2017; Sun et al., 2019; Torres-Cortés et al., 2015). More comprehensive analyses based on sets of housekeeping genes place MRE as sister to the Pneumoniae cluster within Mollicutes (Naito et al., 2017, 2015; Sun et al., 2019; Torres-Cortés et al., 2015). Our result is particularly interesting given the size of our marker set (i.e., 151 COGs in *mollicutes_odb10*), which is the largest used to place the MRE lineage in Mollicutes to date. The evolutionary history of Mollicutes is currently incomplete and in flux, made clear by the paraphyly of genera in modern phylogenomic reconstructions of the group (Figure 3.2A) (Naito et al., 2017; Torres-Cortés et al., 2015). While the resolution of deep nodes or delineation of novel taxonomic groupings within the Mollicutes is outside the scope of this work, our robustly supported phylogeny could inform future work to resolve inconsistencies in Mollicutes taxonomy. It also underpins the importance of marker set selection and demonstrates that it can have dramatic impacts on topology in concatenated analyses. Like in our phylogeny, MRE in past phylogenies is often connected to the Mollicutes by a long branch. Although this is consistent with rapid evolutionary rates and genome mosaicism, it should also invite scrutiny of our MRE placement as it is well-known that long branch lengths can cause the wrong topology to be strongly supported (Bergsten, 2005). On the other hand, our resolution of a paraphyletic BRE within the Burkholderiaceae agrees with recent genome-scale phylogenetic analyses in BRE (Figure 3.2B) (Guo et al., 2020).

We detect MRE and BRE symbionts in a phylum where they have never been detected before (Araldi-brondolo et al., 2017). Our findings require that the distribution of MRE and BRE in Fungi be expanded to include the Zoopagomycota, an early-diverging phylum of fungi that are not plant-associated (Drechsler, 1959; Keller, 1997; Spatafora et al., 2016; Tanabe et al., 2000).

The concentration of obligate EHB in plant-associated fungi has led to their framing in the context of plant-fungal interactions (Araldi-brondolo et al., 2017; Pawlowska et al., 2018). This framing is especially relevant to BRE where mutualistic bipartite and tripartite interactions are becoming increasingly appreciated, such as the causal role of *Mycetohabitans rhizoxinica* in rice seedling blight (Partida-Martinez and Hertweck, 2005). Expansions in the distribution of EHB to include the Zoopagomycota, a phylum where plant-fungal associations are rare, necessitates a redrafting of the plant-centric framing of EHB diversity and function toward one that includes parasitic and predatory interactions with animals, protozoans, and other fungi.

3.6 Literature Cited

Ahrendt, S.R., Quandt, C.A., Ciobanu, D., Clum, A., Salamov, A., Andreopoulos, B., Cheng, J.-F., Woyke, T., Pelin, A., Henrissat, B., Reynolds, N.K., Benny, G.L., Smith, M.E., James, T.Y., Grigoriev, I.V., 2018. Leveraging single-cell genomics to expand the fungal tree of life. Nature Microbiology 3. https://doi.org/10.1038/s41564-018-0261-0

Alexander, W.G., Wisecaver, J.H., Rokas, A., Hittinger, C.T., 2016. Horizontally acquired genes in early-diverging pathogenic fungi enable the use of host nucleosides and nucleotides. Proc. Natl. Acad. Sci. U. S. A. 113, 4116–4121.

Amses, K.R., Davis, W.J., James, T.Y., 2020. SCGid, a consensus approach to contig filtering and genome prediction from single cell sequencing libraries of uncultured eukaryotes. Bioinformatics 36, 1194–2000.

Andersson, K.-M., Kumar, D., Bentzer, J., Friman, E., Ahrén, D., Tunlid, A., 2014. Interspecific and host-related gene expression patterns in nematode-trapping fungi. BMC Genomics 15, 968.

Araldi-brondolo, S.J., Spraker, J., Shaffer, J.P., Woytenko, E.H., Baltrus, D.A., Gallery, R.E., Arnold, A.E., 2017. Bacterial Endosymbionts: Master Modulators of Fungal Phenotypes. The Fungal Kingdom 981–1004.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477.

Barron, G.L., 1973. Nematophagous fungi: Rhopalomyces elegans. Can. J. Bot. 51, 2505–2507.

Bergsten, J., 2005. A review of long-branch attraction. Cladistics 21, 163–193.

Bianciotto, V., Lumini, E., Bonfante, P., Vandamme, P., 2003. "Candidatus Glomeribacter gigasporarum" gen. nov., sp. nov., an endosymbiont of arbuscular mycorrhizal fungi. Int. J. Syst. Evol. Microbiol. 53, 121–124.

Davis, W.J., Amses, K.R., Benny, G.L., Carter-house, D., Chang, Y., Grigoriev, I., Smith, M.E., Spatafora, J.W., Stajich, J.E., James, T.Y., 2019. Genome-scale phylogenetics reveals a monophyletic Zoopagales (Zoopagomycota , Fungi). Mol. Phylogenet. Evol. 133, 152–163.

Drechsler, C., 1959. Several Zoopagaceae Subsisting on a Nematode and on Some Terricolous Amoebae. Mycologia 51, 787–823.

Drechsler, C., 1936. A new species of Stylopage preying on nematodes. Mycologia 28, 241–246.

Eddy, S.R., HMMER development team, 2015. HMMER.

Esch, L., Schaffrath, U., 2017. An Update on Jacalin-Like Lectins and Their Role in Plant Defense. Int. J. Mol. Sci. 18. https://doi.org/10.3390/ijms18071592

Gawad, C., Koh, W., Quake, S.R., 2016. Single-cell genome sequencing : current state of the science. Nat. Rev. Genet. 17, 175–188.

Gryganskyi, A.P., Humber, R.A., Smith, M.E., Miadlikovska, J., Wu, S., Voigt, K., Walther, G., Anishchenko, I.M., Vilgalys, R., 2012. Molecular phylogeny of the Entomophthoromycota. Mol. Phylogenet. Evol. 65, 682–694.

Guo, Y., Takashima, Y., Sato, Y., Narisawa, K., Ohta, H., Nishizawa, T., 2020. Mycoavidus sp. Strain B2-EB: Comparative Genomics Reveals Minimal Genomic Features Required by a Cultivable Burkholderiaceae-Related Endofungal Bacterium. Appl. Environ. Microbiol. 86. https://doi.org/10.1128/AEM.01018-20

Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119.

James, T.Y., Stajich, J.E., Hittinger, C.T., Rokas, A., 2020. Toward a Fully Resolved Fungal Tree of Life. Annu. Rev. Microbiol. 74, 291–313.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.Y., Lopez, R., Hunter, S., 2014. InterProScan 5: Genome-scale protein function classification. Bioinformatics 30, 1236–1240.

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

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.

Keller, S., 1997. The genus Neozygites (Zygomycetes , Entomophthorales) with special reference to species found in tropical regions. Sydowia 49, 118–146.

Lai, Y., Liu, K., Zhang, Xinyu, Zhang, Xiaoling, Li, K., Wang, N., Shu, C., Wu, Y., Wang, C., Bushley, K.E., Xiang, M., Liu, X., 2014. Comparative genomics and transcriptomics analyses reveal divergent lifestyle features of nematode endoparasitic fungus hirsutella minnesotensis. Genome Biol. Evol. 6, 3077–3093.

Mikhailov, K.V., Simdyanov, T.G., Aleoshin, V.V., Belozersky, A.N., 2016. Genomic survey of a hyperparasitic microsporidian Amphiamblys sp. (Metchnikovellidae). Genome Biol. Evol. 9, 454–467.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., Lanfear, R., 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. 37, 1530–1534.

Mondo, S.J., Toomer, K.H., Morton, J.B., Lekberg, Y., Pawlowska, T.E., 2012. Evolutionary stability in a 400-million-year-old heritable facultative mutualism. Evolution 66, 2564–2576.

Moran, N.A., McCutcheon, J.P., Nakabachi, A., 2008. Genomics and evolution of heritable bacterial symbionts. Annu. Rev. Genet. 42, 165–190.

Naito, M., Desirò, A., González, J.B., Tao, G., Morton, J.B., Bonfante, P., Pawlowska, T.E., 2017. ‘Candidatus Moeniiplasma glomeromycotorum’, an endobacterium of arbuscular mycorrhizal fungi. Int. J. Syst. Evol. Microbiol. 67, 1177–1184.

Naito, M., Morton, J.B., Pawlowska, T.E., 2015. Minimal genomes of mycoplasma-related endobacteria are plastic and contain host-derived genes for sustained life within Glomeromycota. Proc. Natl. Acad. Sci. U. S. A. 112, 7791–7796.

Partida-Martinez, L.P., Hertweck, C., 2005. Pathogenic fungus harbours endosymbiotic bacteria for toxin production. Nature 437, 884–888.

Pawlowska, T.E., Gaspar, M.L., Lastovetsky, O.A., Mondo, S.J., Real-Ramirez, I., Shakya, E., Bonfante, P., 2018. Biology of Fungi and Their Bacterial Endosymbionts. Annu. Rev. Phytopathol. 56, 289–309.

Reuber, T.L., Ausubel, F.M., 1996. Isolation of Arabidopsis genes that differentiate between resistance responses mediated by the RPS2 and RPM1 disease resistance genes. Plant Cell 8, 241–249.

Revell, L.J., 2012. phytools: An R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution.

Rosenzweig, W.D., Premachandran, D., Pramer, D., 1985. Role of trap lectins in the specificity of nematode capture by fungi. Can. J. Microbiol. 31, 693–695.

Seemann, T., 2015. Barrnap. Github.

Seppey, M., Manni, M., Zdobnov, E.M., 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods Mol. Biol. 1962, 227–245.

Spatafora, J.W., Chang, Y., Benny, G.L., Lazarus, K., Smith, M.E., Berbee, M.L., Bonito, G., Corradi, N., Grigoriev, I., Gryganskyi, A., James, T.Y., O'Donnell, K., Roberson, R.W., Taylor, T.N., Uehling, J., Vilgalys, R., White, M.M., Stajich, J.E., 2016. A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data. Mycologia 108, 1028–1046.

Sun, X., Chen, W., Ivanov, S., MacLean, A.M., Wight, H., Ramaraj, T., Mudge, J., Harrison, M.J., Fei, Z., 2019. Genome and evolution of the arbuscular mycorrhizal fungus Diversispora epigaea (formerly Glomus versiforme) and its bacterial endosymbionts. New Phytol. 221, 1556–1573.

Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H., UniProt Consortium, 2015. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31, 926–932.

Tanabe, Y., O'Donnell, K., Saikawa, M., Sugiyama, J., 2000. Molecular phylogeny of parasitic zygomycota (Dimargaritales, zoopagales) based on nuclear small subunit ribosomal DNA sequences. Mol. Phylogenet. Evol. 16, 253–262.

Toomer, K.H., Chen, X., Naito, M., Mondo, S.J., den Bakker, H.C., VanKuren, N.W., Lekberg, Y., Morton, J.B., Pawlowska, T.E., 2015. Molecular evolution patterns reveal life history features of mycoplasma-related endobacteria associated with arbuscular mycorrhizal fungi. Mol. Ecol. 24, 3485–3500.

Torres-Cortés, G., Ghignone, S., Bonfante, P., Schüßler, A., 2015. Mosaic genome of endobacteria in arbuscular mycorrhizal fungi: Transkingdom gene transfer in an ancient mycoplasma-fungus association. Proc. Natl. Acad. Sci. U. S. A. 112, 7785–7790.

Tunlid, A., Jansson, H.-B., Nordbring-Hertz, B., 1992. Fungal attachment to nematodes. Mycol. Res. 96, 401–412.

Wekesa, V.W., Moraes, G.J., Knapp, M., Jr, I.D., 2007. Interactions of two natural enemies of Tetranychus evansi , the fungal pathogen Neozygites floridana (Zygomycetes : Entomophthorales) and the predatory mite , Phytoseiulus longipes (Acari : Phytoseiidae). Biol. Control 41, 408–414.

Whisler, H.C., Travland, L.B., 1974. The rotifer trap of Zoophagus. Arch. Microbiol. 101, 95–107.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the tidyverse. J. Open Source Softw. 4, 1686.

Yu, G., Lam, T.T.-Y., Zhu, H., Guan, Y., 2018. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. Mol. Biol. Evol. 35, 3041–3043.

Yu, G., Smith, D.K., Zhu, H., Guan, Y., Lam, T.T., 2017. Ggtree : An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. 8, 28–36.

3.7 Figures

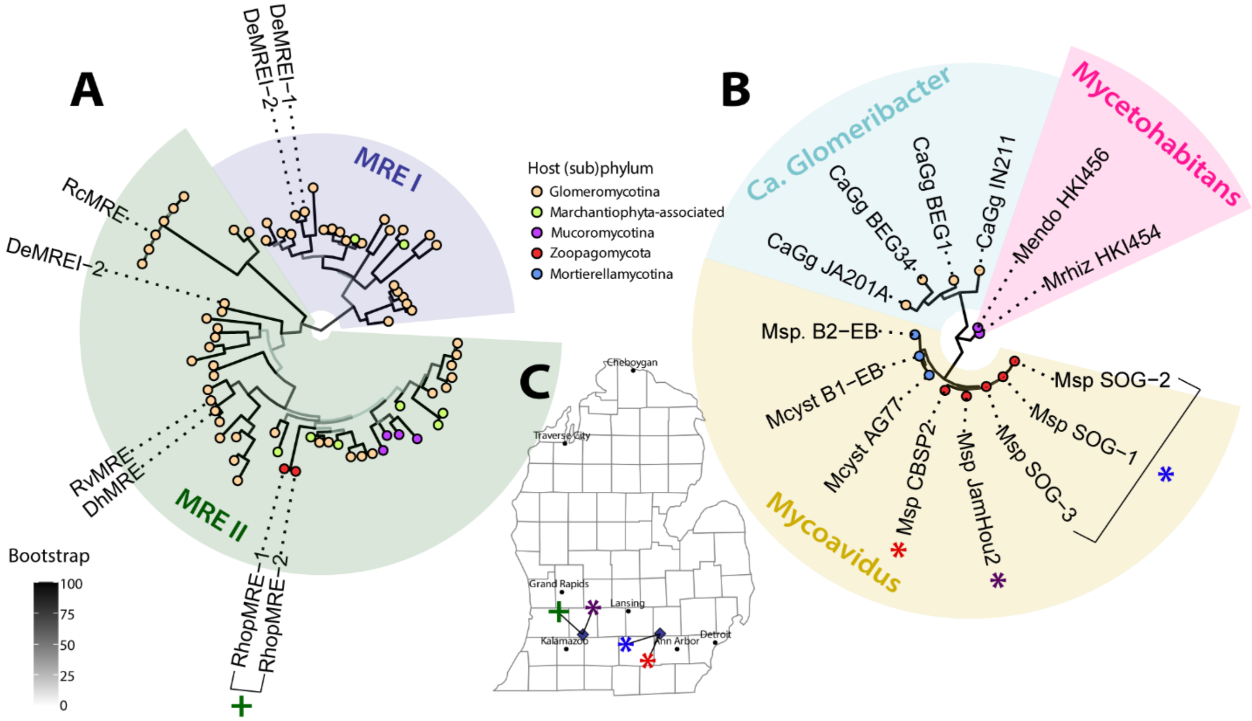


Figure 3.1. ML trees of fungal EHB with novel Zoopagomycota-associated EHB based on 16S rDNA and accompanying map of isolate collection locations. Tree tips are colored by the taxonomic phylum of their known hosts. Bootstraps are shown on tree edges. (A) ML tree of MRE showing the placement of RhopMRE in the MRE II clade (green plus symbol). RhopMRE is represented by two tips, the sequences of which were extracted *in silico* from single-library genome assemblies. MRE with publicly available genome sequence data and included in this study have additional text tip labels. (B) ML tree of BRE showing the placement of *Mycoavidus* sp. SOG in *Mycoavidus* represented by three tips (blue asterisk); 16S rDNA sequences for these tips was extracted from single-library genome assemblies *in silico*. Other novel *Stylopage*-associated BRE are represented by two tips, the sequences of which were PCR-amplified from MDA extracts that were not genome sequenced (blue and red asterisk). Other BRE with publicly available genomes and included in this study have additional text tip labels. (C) Map of lower Michigan counties showing collection sites and isolates collected there. Blue diamonds represent GPS coordinates of collection sites and surrounding symbols indicate which isolates were collected at those sites.

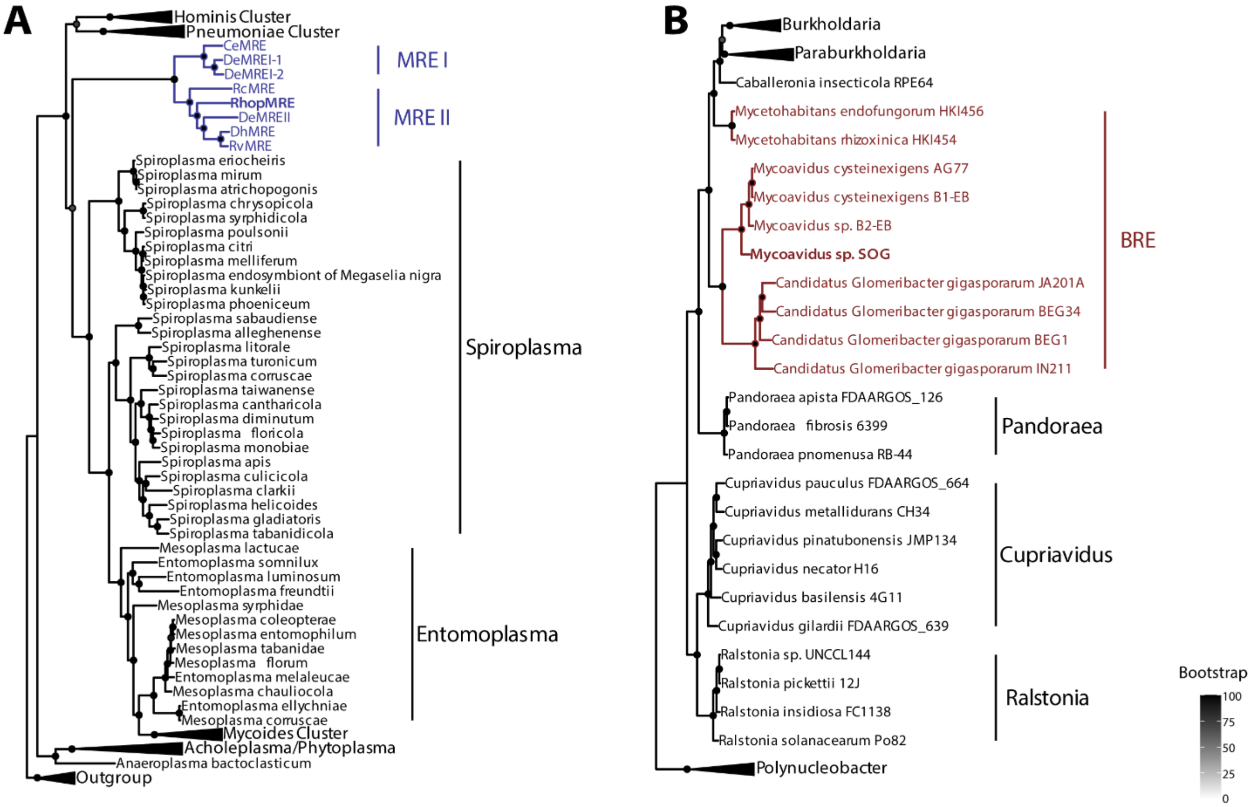


Figure 3.2. Phylogenomic concatenated ML trees of Zoopagomycota-associated EHB and related bacteria. Support values resulting from 10,000 ultrafast bootstraps are shown as shaded circular points on nodes. (A) ML tree of the Mollicutes and MRE based on 42,240 amino acids from 151 high occupancy markers contained in the BUSCO *mollicutes_odb10* database. MRE clade is highlighted in blue and RhopMRE is bolded. (B) ML tree of the Burkholderiaceae and related EHB based on 229,854 amino acids from 688 high occupancy markers contained in the BUSCO *burkholderiaceae_odb10* database. EHB clades are highlighted in red and *Mycoavidus* sp. SOG is bolded.

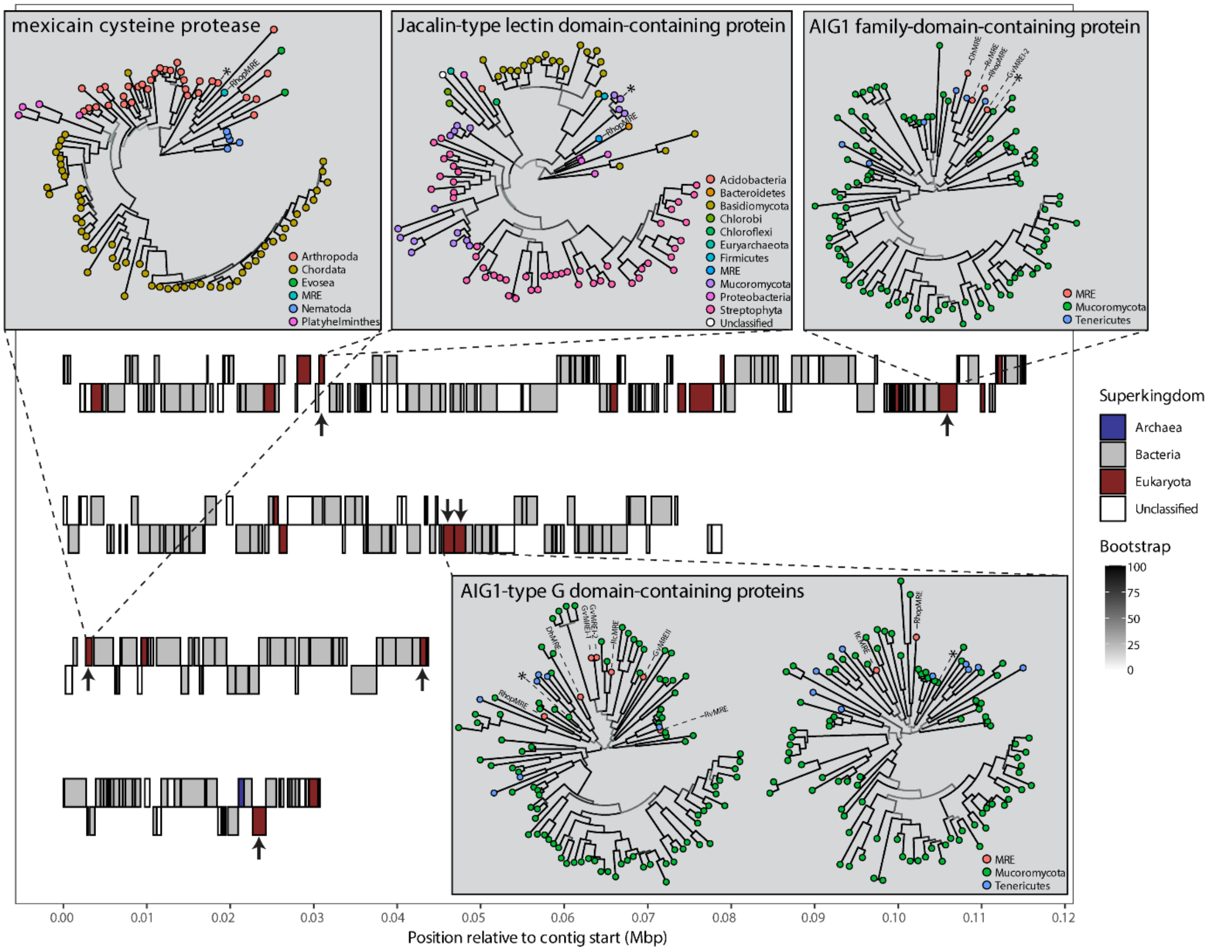
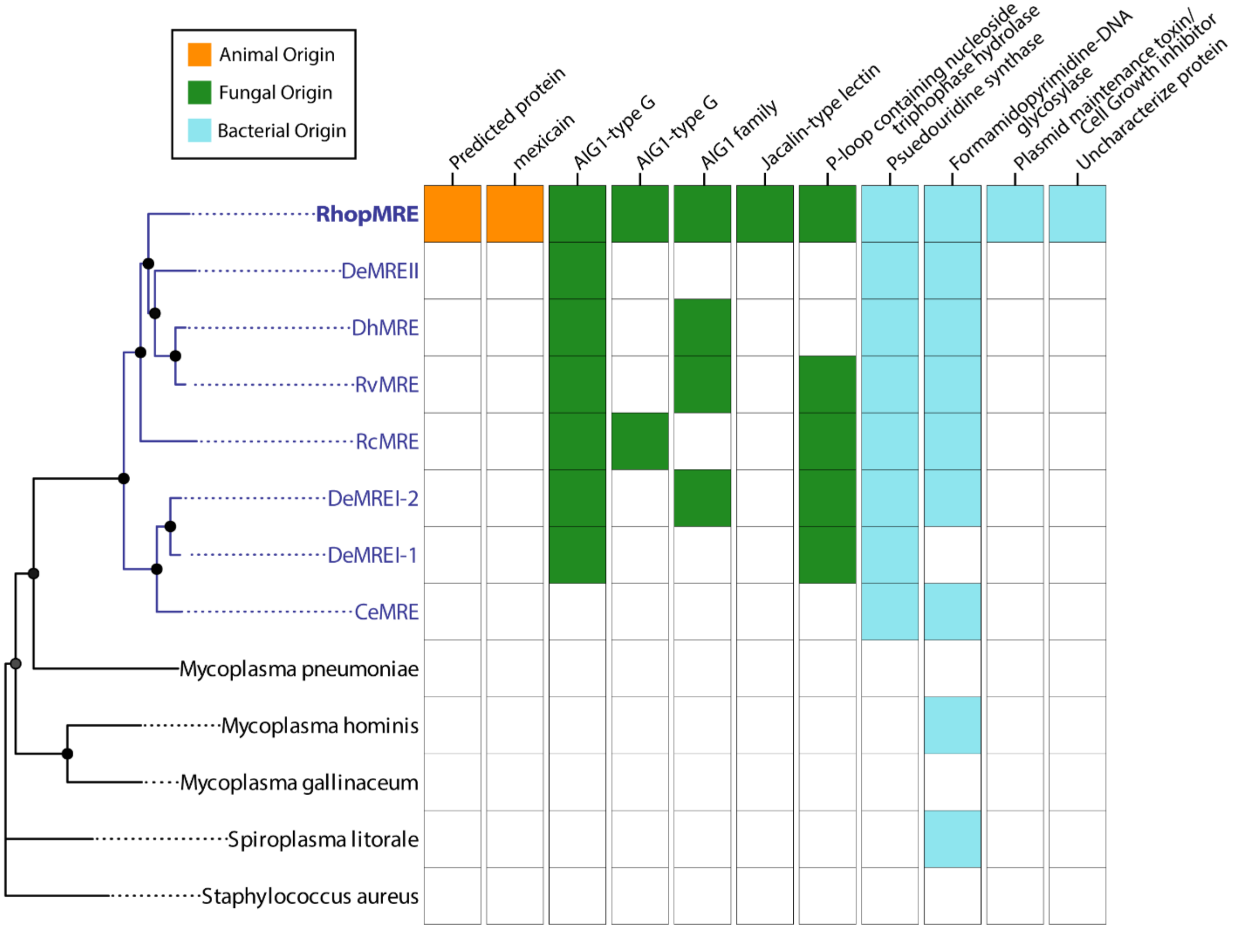


Figure 3.3. Genetic map of annotated genes on four contigs of the RhopMRE draft genome assembly on which HGT candidates were identified. Genes are arranged along the x-axis based on their position on each contig and colored by the superdomain of the top BLAST hit of the protein they encode in the custom UniRef database. Arrows indicate HGT candidate genes. A selection of HGT candidate gene trees is shown in grey insets and dashed lines connect each inset to its corresponding gene. Gene trees in insets are ML reconstructions of an HGT candidate and up to its strongest 100 hits in the custom database in addition to any significant hits (e-value = 1e-50) from other published MRE genomes. Tips are colored by the taxonomic phylum of the protein they represent according to each inset legend. Legends vary by inset. MRE genomes have tip labels and are colored according to each inset legend. Each tree edge is colored by the support value resulting from 10,000 ultrafast bootstraps. Asterisks indicate the top hit to each HGT candidate in the custom database.



	Predicted protein	mexicain	AIG1-type G	AIG1-type G	AIG1 family	Jacalin-type lectin	P-loop containing nucleoside triphophase hydrolase	Psuedouridine synthase	Formamidopyrimidine-DNA glycosylase	Plasmid maintenance toxin/ Cell Growth inhibitor	Uncharacterize protein
RhopMRE	🟧	🟧	🟩	🟩	🟩	🟩	🟩	🟦	🟦	🟦	🟦
DeMREII			🟩					🟦	🟦		
DhMRE			🟩		🟩			🟦	🟦		
RvMRE			🟩		🟩		🟩	🟦	🟦		
RcMRE			🟩	🟩			🟩	🟦	🟦		
DeMREI-2			🟩		🟩		🟩	🟦	🟦		
DeMREI-1			🟩				🟩	🟦			
CeMRE								🟦	🟦		
Mycoplasma pneumoniae											
Mycoplasma hominis									🟦		
Mycoplasma gallinaceum											
Spiroplasma litorale									🟦		
Staphylococcus aureus											

Figure 3.4. Reduced concatenated ML phylogenomic tree of MRE, non-MRE Mollicutes, and *Staphylococcus aureus* outgroup based on 23,566 amino acid positions with HGT candidate presence or absence data mapped on to species tips. Matrix cells show whether a potential homolog was detected in that genome by BLAST (e-value = 1e-50). Colored squares indicate presence while white squares indicate absence. HGT candidates are colored by their putative origin as inferred from gene trees (animal: orange; fungi: green; bacterial: blue). Genes of bacterial origin are not HGT candidates in the context of RhopMRE, but were likely transferred from MRE to fungi.

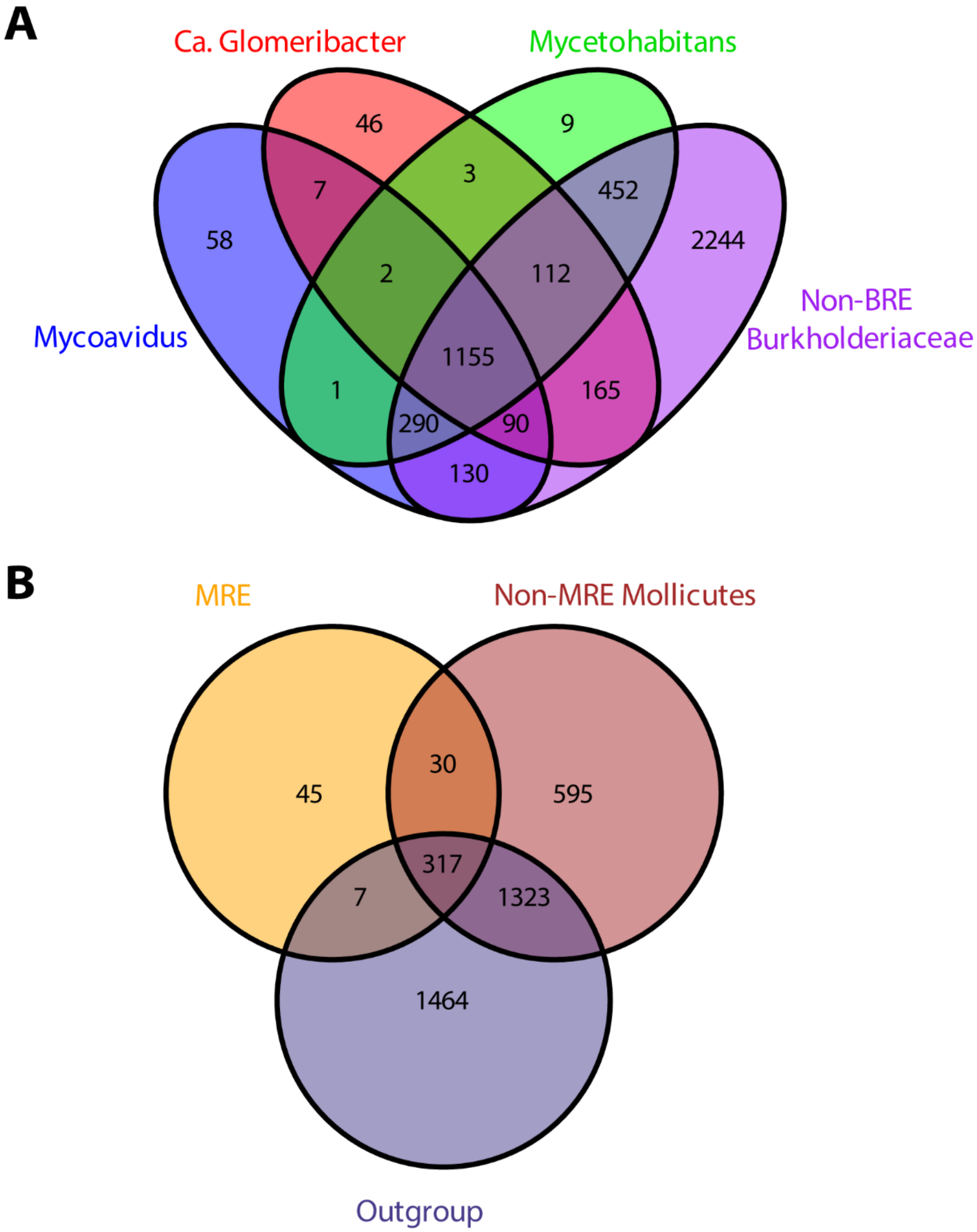


Figure 3.5. Venn diagrams showing counts of unique to or shared PFAM domains between different groupings of EHB. (A) PFAM domains unique to or shared between different genera of BRE and non-BRE Burkholderiaceae. (B) PFAM domains unique to or shared between MRE, non-MRE Mollicutes, or a non-Mollicutes outgroup.

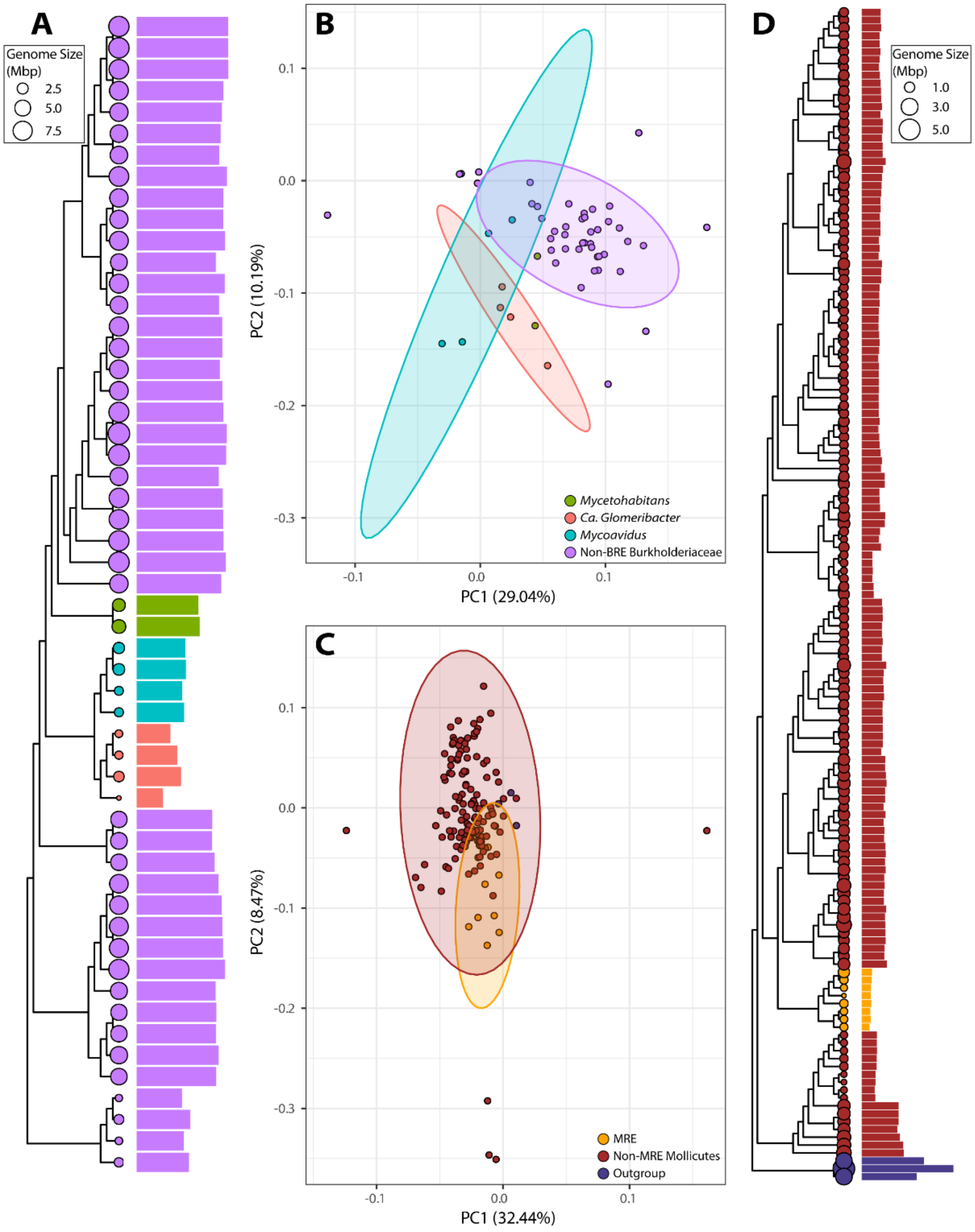


Figure 3.6. Summary of high-level, PFAM-based comparative genomic analyses between EHB and their non-EHB relatives. (A) BRE ML phylogenomic tree (from Figure 2B) annotated with genome size (points on tips) and counts of unique PFAM domains annotated in predicted proteomes (bars). (B) Phylogenetically scaled PCA clustering BRE and non-BRE Burkholderiaceae genomes by their PFAM domain profiles. Uncertainty ellipses shown for groups with n > 3. (C) Phylogenetically scaled PCA clustering MRE, non MRE Mollicutes, and non-Mollicutes outgroup genomes by their PFAM domain profiles. Uncertainty ellipses shown for groups with n > 3. (D) MRE ML phylogenomic tree (from Figure 2B) annotated with genome size (points on tips) and counts of unique PFAM domains annotated in predicted proteomes. Color codes: Non-BRE Burkholderiaceae (purple), *Mycoavidus* (light blue), *Mycetohabitans* (green), Ca. Glomeribacter (light red), MRE (yellow) non-MRE Mollicutes (red), non-Mollicutes outgroup (dark blue).

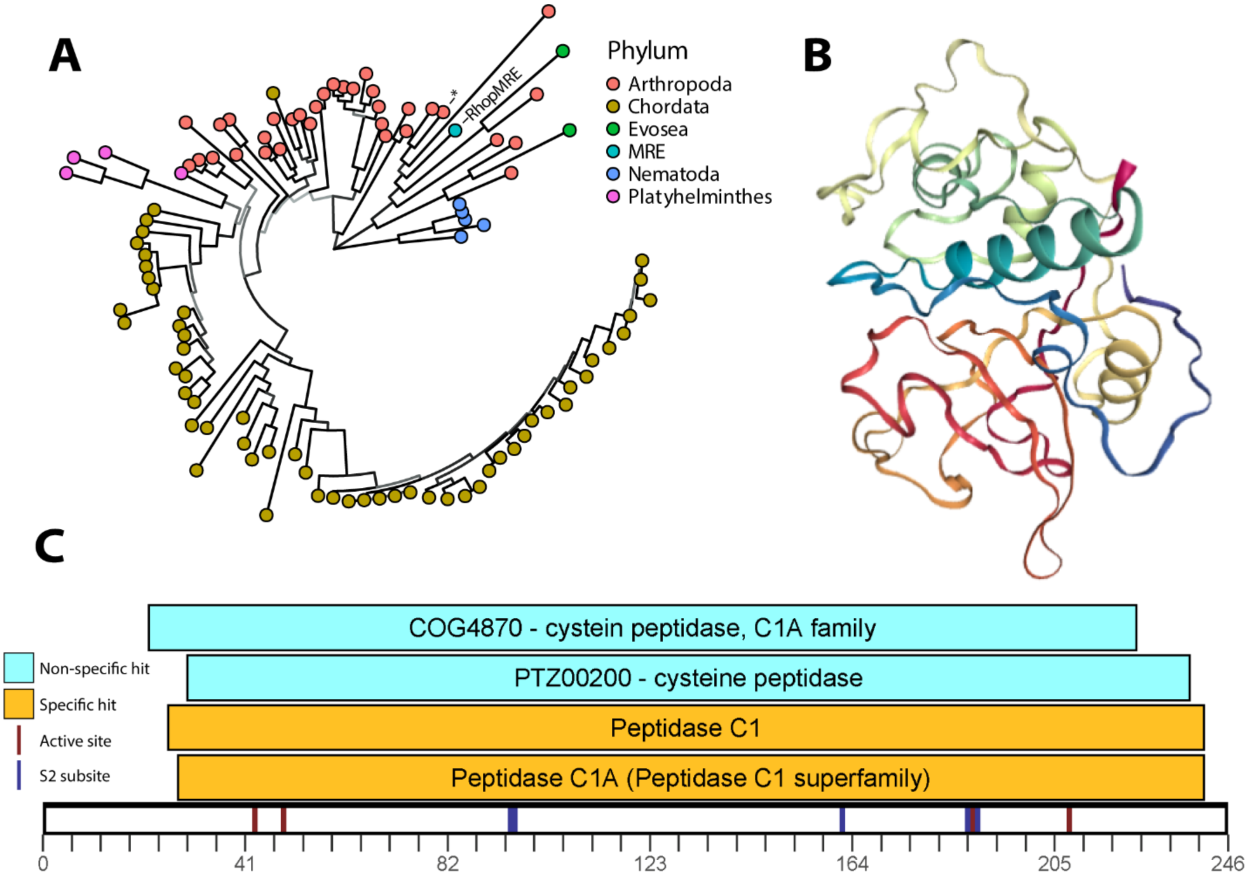


Figure 3.7. Further characterization of the putative animal-derived mexicain protein encoded in the genome of RhopMRE. (A) ML tree of the RhopMRE protein and its top 100 best BLAST hits in our custom UniProt database. Bootstraps are shown by the coloring of nodes on a gradient from grey to black (0-100). (B) The MODELLER-predicted tertiary structure of the RhopMRE mexicain protein. (C) BLASTP domain map of RhopMRE showing its homology to MEROPS class C1A cysteine peptidases.

Chapter 4: Phylogenomic analysis of zoosporic true fungi suggests most early diverging lineages have diploid-dominant life cycles.

4.1 Abstract

The majority of fungal species diversity is contained within two phyla, Ascomycota and Basidiomycota (subkingdom Dikarya), and knowledge about the kingdom is heavily influenced by traits of Dikarya, such as aerial spores and haplontic life cycles dominated by haploid mitosis. Yet, we now appreciate that the evolutionary history of fungi is much deeper and comprises numerous lineages that diversified before Dikarya, but the phylogeny and genetic characteristics of these lineages are poorly understood. Here we greatly increase the genomic sampling of the zoosporic early diverging lineages of true fungi and produce a robust phylogenetic hypothesis based on 487 protein coding sequences. Our phylogeny recovers 5 paraphyletic lineages of zoosporic fungi, placing the Blastocladiomycota with alternation of haploid and diploid generations as likely sister to *Olpidium*+the remaining terrestrial fungi. Using estimates of heterozygosity based on genome sequence data we find that both zoosporic lineages as well as the Zoopagomycota are primarily characterized by diploid mitosis. We mapped key ancestral traits shared with Metazoa and other eukaryotes on the phylogeny, such as use of the amino acid selenocysteine and ancestral cell cycle regulators, and reveal these traits to have been subject to rampant, parallel loss. Together, these results show a gradual transition in the genetics and cell biology of fungi from their protist-like ancestor and caution against assuming that traits only measured in Dikarya are transferable to the earlier diverging lineages.

4.2 Introduction

Kingdom Fungi evolved from a protist-like ancestor (Opisthokont) that was shared with Animalia, yet the two groups have diverged in ways that make their kinship barely recognizable. Fungi grow within their food and mostly feed by osmotrophy, while animals eat things smaller than themselves and feed by phagotrophy or ingestion, and this difference is the basis for massive differences in morphology, including loss of motility during feeding and polarized cell growth in fungi (James and Berbee, 2012; Richards et al., 2017). In addition to these major distinctions, another major difference between the two is the life cycle wherein fungi show zygotic meiosis (haploid dominant life cycle) and animals show gametic meiosis (diploid dominant), and this is believed to be a major distinction between the kingdoms. Early diverging fungal (EDF) lineages, however, here equivalent to all phyla outside of Dikarya (i.e., non-Dikarya) have retained traits from the Opisthokont ancestor shared with Metazoa, such as motility and flagellation, presence of cholesterol in the membrane, actin structures, vitamin dependencies, and cell cycle genes (Medina et al., 2016; Naranjo-Ortiz and Gabaldón, 2019; Orłowska et al., 2021; Prostak et al., 2021; Weete et al., 2010). However, the number and pattern of character transitions between the Opisthokont ancestor and the descendent kingdoms are shrouded because we lack a robustly sampled and consistently supported phylogeny of EDF. Moreover, recent discoveries reveal there is a much greater phylogenetic diversity of EDF than previously appreciated (James et al., 2020; Seto et al., 2020; Tedersoo et al., 2018), and the divergence times are so old that it is difficult to adequately estimate the phylogeny. The goal of this paper is to provide the most comprehensive phylogeny of the earliest fungal lineages with emphasis on the zoosporic fungi to reassess critical transitions in characters during early fungal diversification.

The systematics of the zoosporic fungi is in great flux due to discovery or analysis of new taxa and uncertainty in phylogenetic relationships (James et al., 2020). Zoosporic fungi comprise 8 phyla that form a paraphyletic grade along the earliest branches of the fungal tree: Aphelidiomycota, Blastocladiomycota, Chytridiomycota, Monoblepharidomycota, Neocallimastigomycota, Olpidiomycota, Rozellomycota Sanchytriomycota, whose relationships remain contentious (Chang et al., 2021; Galindo et al., 2020; James et al., 2020; Li et al., 2021; Tedersoo et al., 2018). The earliest diverging phyla, Rozellomycota and Aphelidiomycota, are endoparasites that have retained from the most recent common ancestor (MRCA) of Fungi and Metazoa the ability to utilize phagocytosis, which they employ for devouring host cytoplasm (Karpov et al., 2014; Powell et al., 2017). Zoosporic fungi have simple vegetative thalli that may be unicellular or more mycelium like, while the more complex ultrastructure of zoospores has been used for taxonomic revision (Letcher et al., 2006; Letcher and Powell, 2014; Longcore and Simmons, 2012). The largest and most diverse group, Chytridiomycota, has an estimated 14 orders, and the relationships among these orders is largely unresolved (James et al., 2020; Seto et al., 2020; Simmons et al., 2020).

The Blastocladiomycota, containing the well-known water mold *Allomyces*, is phylogenetically distinct from the core group of zoosporic fungi, the Chytridiomyceta (Chytridiomycota + Monoblepharidomycota + Neocallimastigomycota). Yet, the placement of the Blastocladiomycota has been controversial (Chang et al., 2015; Galindo et al., 2020; James et al., 2020; Li et al., 2021; Liu et al., 2009), with nearly equivocal support for the blastoclads diverging before the divergence of the Chytridiomyceta or after the Chytridiomyceta. Several traits of blastoclads ally them with the terrestrial fungi: closed mitosis, the presence of a cellular vesicular structure termed a Spitzenkörper, beta 1,3 glucans in the cell wall, and a true mycelium-like growth in some members (James et al., 2006b; Ruiz-Herrera and Ortiz-Castellanos, 2019). However, the most distinctive characteristic of the group is the presence of sporic meiosis with alternating haploid and diploid thalli as observed in most blastoclads (James et al., 2014). This may indicate that the taxon is intermediate between the MRCA of animals/fungi that likely are diplontic and the majority of fungi that are haplontic. At least a couple reports indicate that blastoclads may have mating types distinguished as gametophytes with differing colors, while mating types are not known from Chytridiomyceta (Idnurm et al., 2007; Whisler et al., 1975).

Overall, mating and sexuality is poorly described in zoosporic fungi. The textbook life cycles of Chytridiomycota imply a zygotic meiosis for most taxa, but the majority of assumptions of meiotic stages are unconfirmed by cytology. Moreover, there is a complete lack of genetic evidence to back up any of these inferred life cycles, and mating in the lab is not developed for any species of Chytridiomyceta. Importantly, the best studied chytrid fungus, *Batrachochytrium dendrobatidis*, has a life cycle that appears to be dominated by asexually reproducing diploid or higher ploidy thalli (Rosenblum et al., 2013; Schloegel et al., 2012). More recently, evidence based on single cell genomes showing that non-Dikarya phyla show heterozygosity indicative of diploidy (Ahrendt et al., 2018) implies that the assumption of zygotic meiosis for the Chytridiomyceta may be false.

A major limitation in answering questions on the phylogeny, life cycles, and character evolution of zoosporic fungi is the absence of genomic sequence data for many taxonomic orders and families (James et al., 2020). A fully sampled tree will also require integrating uncultured taxa using single cell genomics or metagenomic approaches (Ahrendt et al., 2018; Amses et al., 2020; Chang et al., 2021). In order to adequately trace the evolution of morphological and genetic characters, a robust phylogeny is needed to determine which characteristics are informative and which are homoplasious. Here, we sampled 68 zoosporic fungal genomes using both cultures and single cell approaches to provide a strongly supported phylogeny for understanding taxonomy and the evolution of key characters, with an emphasis on resolving the evolution of life cycles and ploidy. We specifically leveraged our sequencing data to estimate the heterozygosity of our genomes and discovered that the majority of zoosporic fungal phyla demonstrate diplontic life cycles.

4.3 Methods

Strains, vouchers, and genome sequencing methods

We generated 20 high coverage genome sequences, 45 low coverage sequences, and 3 sequences using single cell/low input approaches. The full list of taxa used in our analyses and sequencing methods are found in Table C1. Most material is deposited in a cryopreserved state in the CZEUM collection (Simmons et al., 2020). Details on growth and extraction of DNA from cultured strains for sequencing can be found in Simmons et al. (2020). RNA extraction was performed using TRIzol reagent (Invitrogen).

For genomes sequenced to high coverage, both PacBio SMRT and Illumina sequencing were used. Sequencing libraries were prepared and sequenced on the PacBio SMRT long-read sequencing platform by our collaborators at JGI, following their standard approach. Low coverage genomes were sequenced both at the U. Michigan Advanced Genomics Core and at the

JGI. Library prep was done using a Nextera XT kit (Illumina), and sequencing was done on a HiSeq-4000 using paired end 150 bp mode. Sequencing was performed on two samples using low input methods. DNA extracts of *Rozella multimorpha*, an endoparasite on the water mold *Pythium*, were prepared for sequencing and sequenced on the Illumina sequencing platform using a ThruPLEX DNA-Seq Kit. A single cell of the alga *Micrasterias* cf. *truncata* (PSC023) infected with an endoparasitic chytrid was used for single cell sequencing. The sample was prepared for sequencing according to (Davis et al., 2019).

Assembly and Annotation

For high coverage genomes, assembly and annotation was conducted by the JGI according to their in-house pipelines.

For low coverage genomes, to facilitate downstream ploidy estimation, we generated haploidized draft assemblies using ploidy-aware assembly methods. To make *a priori* estimates of genome ploidy, we assembled reads for each taxon with both a haploid assembly algorithm (*SPAdes* v3.11.1) (Bankevich et al., 2012) and a diploid assembly algorithm (*dipSPAdes* v3.11.1) (Safonova et al., 2015). Based on the cumulative lengths and N50 values of initial assemblies, we sorted genomes into “likely diploid” and “likely haploid” bins. Specifically, we determined a genome to be “likely diploid” if the *dipSPAdes* length was less than or equal to 90% of the *SPAdes* length and the *dipSPAdes* N50 was at least 10% higher than the *SPAdes* N50. Otherwise, we considered the assembly to be “likely haploid”. Following putative ploidy assignment, we restarted the assembly pipeline using either *SPAdes* or *dipSPAdes* accordingly. Resulting genome assemblies were annotated with *funannotate* v1.7.4 (Palmer and Stajich, 2019).

For *R. multimorpha* (ThruPLEX) and PSC023 (SCG), reads libraries were assembled with *SPAdes* v3.11.1, using single-cell mode for PSC023. Following assembly, we used a recently developed single cell binning procedure, *SCGid* (Amses et al., 2020), to identify and remove contaminating sequence data. From the PSC023 SCG metagenome, we recovered two distinct genomes *in silico*, one from the putative chytrid parasite of the alga (*Olpidium*) and another from a presumed hyperparasite of the chytrid (Rozellomycota). These three genome assemblies were annotated with *MAKER* (Cantarel et al., 2008) (*R. multimorpha*) or *funannotate* v1.7.4 (PSC023).

Maximum likelihood phylogenomic analyses

To conduct the genome-scale phylogenomic analyses included in this work we relied on the set of 758 conserved markers comprising the BUSCO *fungi_odb10* database (Simão et al., 2015). Although this marker set is focused on single copy genes, we applied a careful filtering approach to exclude paralogs that were detected and particularly common within EDF.

We searched our predicted proteomes against the *fungi_odb10* database using the *hmmsearch* function included in *HMMER* (Eddy and HMMER development team, 2015). Instead of retaining solely the sequence with the strongest hit to each model, which can in some cases introduce paralogous sequences into phylogenetic analyses, we retained all hits to each protein model detected in each genome as long as the strength of the hit was above the minimum threshold accepted by the BUSCO software pipeline (Seppey et al., 2019). This procedure resulted in zero to many sequences per marker from each genome and subsequently individual locus alignments the size of which far exceeded the 137 taxa in our phylogenetic data set (e.g., 1,349-tip gene tree for *fungi_odb10* marker 6377at4751).

To filter these alignments to homologs, we employed an iterative approach that involved both automated and manual gene tree-curation steps that yielded individual locus alignments that included, at most, one sequence per taxon. Initially we removed low-occupancy markers (<75% marker occupancy). Generation of trees before and after filtering involved alignment with *hmmalign*, trimming alignments with *trimal* v1.2rev59 (Capella-Gutiérrez et al., 2009), manual removal of highly gapped sequences, computing new gene trees with *fasttree* v2.1.10 (Price et al., 2010), and then evaluation of criteria as described below.

We implemented an automated gene tree filtering approach that traversed trees and determined whether all tips corresponding to each taxon were monophyletic or not. If they were all monophyletic, the best hit was taken since the taxon placement was not in conflict. If tips were not monophyletic, we first compared the scores of each protein (i.e., tips) to see if the conflict in taxon placement was the result of one or more particularly low-scoring clusters of tips (i.e., <= 70% the score of the highest-scoring tip for that taxon). We then removed these particularly low-scoring tips and checked again for monophyly. If removal of these low-scoring tips led to monophyly of the taxon, we took the highest-scoring tip among the high-scoring cluster of tips. If monophyly did not result from the removal of these particularly low-scoring tips, we still permanently removed the particularly low-scoring tips but retained all those tips with scores in the higher-scoring bin, despite their polyphyly. Following this first round of monophyly- and taxon-specific score filtering, we regenerated gene trees.

In a second round of taxon-agnostic, score-based filtering, we removed tips with scores that were lower than 1.5 standard deviations from the mean tip score calculated from all tips in each gene tree. In this way we removed low-scoring tips from the tree that received a score higher than 70% of the taxon-specific maximum but were low-scoring relative to the entire tree. Again, we regenerated gene trees following tip removal and computed taxon monophyly across the resultant trees. At this point, the only conflicting placements remaining in gene trees should have been the result of high scoring, but non-monophyletic, clusters of tips. To resolve remaining conflicts, we manually curated all 519 gene trees and tagged tips and nodes for removal based on their clear paralogous position in trees, for example, a cluster of tips corresponding to fungi from multiple phyla that was positioned as an outgroup to all fungi.

Upon removal of these paralogous or otherwise erroneous tips from our continually shrinking sequence set, we regenerated gene trees for one more round of manual curation that aimed to replenish data that may have been automatically removed in error. We noticed that through our automated filtering approach we had reduced the representation of some taxa to very low levels (e.g., ~15% occupancy in gene trees). Many of these taxa are understood as extreme cases both in terms of their long divergence times from the rest of the kingdom and in patterns of genome evolution (e.g., *Mitosporidium daphniae*). We then manually selected poorly represented taxa for which we "spiked" tips back into our alignments from the original, unfiltered sequence set. To assert that these "spiked-in" sequences were not paralogous, we looked at all 368 gene trees to manually determine whether each re-inserted sequence should be kept or removed in line with the decision made by automated filtering. Through this final manual-curation step, we were able to raise all included taxa above ~40% occupancy in gene trees with confidence that paralogous sequences were not included while poorly scoring true orthologs were retained.

Through our four-round part-automated and part-manual iterative filtering approach, we were able to reduce the number of per-taxon sequences in each gene tree to a single sequence. For those cases where we could not determine which sequence to use as the representative sequence for a taxon, we simply removed that taxon from that gene tree. Following this final filter to remove unreconcilable taxa from our gene trees, were generated a set of finalized gene trees for the remaining 487 markers with alignment in *Mafft* (Katoh and Standley, 2013) and tree calculation in *IQ-TREE* v2.0.5 (Minh et al., 2020) with 100 nonparametric bootstraps. Substitution models per gene tree were selected by *ModelFinder* (AIC/BIC highest scoring model) in *IQ-TREE* (Kalyaanamoorthy et al., 2017).

Our filtering approach reduced the 758 conserved marker set from *fungi_odb10* to 487 markers with a mean occupancy of 82.02% and represented in up to 137 proteomes in our taxon set. We concatenated our 487 alignments into a 197,423 amino acid alignment and computed phylogenetic trees in *IQ-TREE* with both unpartitioned and partitioned models. For the unpartitioned analysis, we used the most frequent best (AIC/BIC) substitution model among individual alignments. For the partitioned analysis, we allowed each partition to be modeled by its best estimated model as calculated by *ModelFinder*. The tree topologies from the two models were identical, and therefore we utilized the tree resulting from the unpartitioned analysis. We ran 100 nonparametric bootstraps of our unpartitioned dataset in *IQ-TREE* and annotated the final tree topology with the resultant support values. Support measures were also computed using quartet internode certainty using the program *QuartetScores* (Zhou et al., 2019) and gene Concordance Factors using *IQ-TREE* (Minh et al., 2020).

ASTRAL

We used our 487 gene trees selected above to generate a species tree with local posterior probabilities in *ASTRAL* 5.7.3 (Zhang et al., 2018) using default settings.

Time calibrated phylogeny

We used the concatenated protein ML tree to generate a time-calibrated phylogeny with divergence times between major lineages estimated with the penalized likelihood method implemented in *r8s* v1.81 (Sanderson, 2003). Fossil-based calibration points were used to constrain the minimum ages of the most recent common ancestor (MRCA) of several clades, following (Chang et al., 2019): Blastocladiomycota=407 mya, Chytridiomycota=407 mya, Ascomycota=407 mya, Basidiomycota=330 mya, Mucorales=315 mya. We constrained the tree using a range of allowable dates for the MRCA of Dikarya (500-650 mya), which is based on various reasonable extremes (Lücking et al., 2009). Additional parameters for rate estimation were: smoothing=1000 (chosen using the cross-validation method; num_time_guesses=10; penalty=log).

Ancestral state reconstruction and gene searches

We compiled an ultrastructural data matrix of 32 characters from previously examined taxa Chytridiomyceta species in the AFTOL Structural and Biochemical database (aftol.unm.edu) to which we added ploidy state as estimated below. Where possible, for taxa that were included in our phylogenomic analyses but missing from the AFTOL database, we determined and entered their known or observed character states. Ancestral character states were inferred across our matrix via marginal ancestral state reconstruction with *phytools* (Revell, 2012) in R. For ploidy state, we confirmed the results of the marginal ancestral state reconstruction via stochastic ancestral state reconstruction (Bayesian MCMC), again using *phytools*. Inferred ancestral states were annotated onto the nodes of a reduced-size (Chytridiomycota only) ASTRAL tree (ultrastructural characters) or the full ML phylogeny (ploidy).

We determined the presence of EF1-**α**, EFL (EF1-**α**-like), and cobalamin-dependent genes in the genomes used to construct our phylogeny. Exemplary protein sequences were downloaded from

GenBank from a search of related taxa or selection from a previously published list (Orłowska et al., 2021). We aligned sequences in Muscle 3.8.31, and we used alignments to construct hidden Markov model profiles and search genomes for homologous proteins using *HMMER* v3.1b2. We assessed homologs to eliminate paralogs by aligning sequences in Muscle, creating phylogenetic trees in *Geneious* 9.1.8 ("Geneious," 2019), and deleting long-branch taxa from further analyses. We determined the use of the amino acid selenocysteine using searches for the selenocysteine tRNA gene with *Secmarker* (Santesmasses et al., 2017). Confirmation of selenocysteine usage was complemented by searches for the gene phosphoseryl-tRNA kinase, a consistent marker for fungal utilization of selenocysteine (Mariotti et al., 2019).

Assessing Support and Conflict for Contentious Relationships

We analyzed the individual gene phylogenies for their support for alternative resolution of contentious relationships in the fungal phylogeny. We focused on the resolutions of 5 contentious relationships: the placements of Blastocladiomycota, Monoblepharidomycota, Aphelidiomycota, Neocallimastigomycota, *Olpidium*, and Polychytriales (James et al., 2020; Li et al., 2021). As a control, we analyzed the alternative resolutions of the Ascomycota subphyla, which mostly resolves Taphrinomycotina as the earliest diverging lineage in mitochondrial or multilocus analyses (James et al., 2006a; Liu et al., 2009; Rosling et al., 2011). Our analyses focused on quartets in our ML tree including these focal taxa, and calculated support for alternative resolutions of these quartets in the form of local posterior probabilities (LPP), frequencies of quartets (Q), and differences in log likelihood between alternative constraint trees (deltalnL). LPP and Q values were calculated for the quartets using *ASTRAL* 5.7.3 (Zhang et al., 2018), using modification of scripts available at https://github.com/smirarab/1kp/tree/master/scripts/hypo-test following Li et al. (2021). deltalnL values were calculated for each of the 487 protein alignments by searching for ML trees under topological constraints conforming to each of the three resolutions of the quartets in question using *IQ-TREE* with best models for each protein. Log likelihood values were compared among the three recovered ML trees found under constraint searches as well as an optimal, unconstrained search ML tree. We considered the constraint trees recovered from each protein to be sufficient to identify a most likely quartet only when the quartet constrained search with the highest likelihood was not more than 2 log units less likely than the unconstrained tree. In this way, a widely variable number of proteins were found to provide support for resolution of the quartet, ranging from 20 for the *Olpidium*-focused quartet to 314 for the Ascomycota subphyla quartet.

Ploidy Estimation

To estimate the ploidy of fungal assemblies in our dataset, we employed a two-pronged approach that generated: (i) kmer histograms by counting 23-mers present in raw reads (kmer approach) and (ii) allele frequency histograms by counting SNPs identified by mapping reads to our draft assemblies (AF approach). Our AF approach required that de novo assemblies contain the single haplotype of haploid genomes or one haplotype of the two or more haplotypes present in genomes with 2N+ ploidy. Since our approach was not optimized for long reads, the pipeline was only employed when Illumina short reads were available. For assemblies based on PacBio long reads, we used ploidy estimations provided by our collaborators at JGI. In some cases where only mRNA short reads were available (i.e., RNA-seq), we mapped these to our assemblies instead. Assemblies for which DNA or mRNA short reads were not available were excluded from these ploidy analyses. Where possible, we drew consensus from the literature to assign ploidy or left them scored as ploidy uncertain. Within our 137-taxon dataset, there were 112 assemblies for which DNA or mRNA short reads were available.

To count 23-mers present in raw Illumina reads, we ran the *kmercountexact* algorithm included in *BBMap* (Bushnell, 2014) on all of the Illumina read libraries generated in this study or in published data sets. In cases where multiple Illumina read libraries were available on NCBI SRA, ENA, or JGI Genome Portal, we selected read libraries for use on a case-by-case basis (i.e., based on determined quality in past studies) or otherwise simply based on being the most voluminous library in terms of raw sequence data. We used the output files from individual *kmercountexact* runs to generate kmer frequency histograms using custom scripts and *ggplot2* in R (Wickham, 2016) (https://www.github.com/Michigan-Mycology/Chytrid-Phylogenomics).

We employed a standard SNP-calling pipeline for estimating heterozygosity (SNP rate). Using this approach, we generated Variant Call Format (VCF) files that documented heterozygous positions relative to the reference assembly. Briefly, we mapped raw reads to their corresponding assembly using *bwa mem* (Li, 2013) sorted and removed PCR duplicates with a variety of *samtools* utilities (Li et al., 2009), and finally generated VCFs using *GATK HaplotypeCaller* (Van der Auwera and O'Connor, 2020) specifying that DepthPerAlleleBySample be included in final VCFs. We then filtered these VCF files using custom scripts (https://www.github.com/Michigan-Mycology/Chytrid-Phylogenomics) utilizing functions from *pyvcf* (Casbon, 2016) that removed homozygous positions and heterozygous positions with more than one alternate allele (i.e., likely artifactual). We filtered this SNP set further to exclude low quality SNPs by excluding SNPs with a measured depth (*GATK* DP parameter) outside one standard deviation of the genome-wide mean, that occurred outside of the genome assembly L50 contig set, and had a measured MapQualityRankSum (MQRS) test value that was not equal to 0; that is, we forced MappingQuality of reads bearing the reference allele to be identical to that of reads bearing the alternate allele. With our filtered set of high-quality SNPs, we generated allele frequency histograms, with *ggplot* in R, that plotted the distribution of SNPs by the allele frequency (*GATK* AF parameter) of their reference versus alternate alleles.

In order to make our metagenomic assemblies (i.e., rozellid and *Olpidium*-like members of PSC023) compatible with our allele frequency mapping approach, we first mapped the metagenomic reads to the filtered draft assemblies for each member. We filtered the resulting SAM files using *samtools view* to remove unmapped reads, pairs where one read was orphaned, and all supplementary alignments before extracting forward and reverse read files from the filtered SAMs. Segregate read libraries were used as input into our standard allele frequency mapping approach as described above.

The histograms generated by each prong of our two-pronged approach were visualized as a pair and used to estimate the ploidy of the 112 assembly-reads pairs in our ploidy dataset. We assessed the validity of our method by evaluation of known diploid species such as *Batrachochytrium dendrobatidis* (Rosenblum et al., 2013) and *Allomyces javanicus* (Emerson, 1941), which displayed bimodal kmer frequency histograms and unimodal allele frequency histograms centered at or around 50% allele frequency, both indicative of genome-wide heterozygosity consistent with the presence of two sets of homologous chromosomes.

Based on the variable quality of different genome assemblies impacting both types of histograms (low coverage of some being a main contributor), we conducted subsequent analyses to separate genomes of questionable ploidy based on the mean SNP density across the L50 contig set and the fit of the allele frequency distribution to the expected distribution under the assumption that a diploid assembly should be evidenced by a binomial distribution centered at AF = 50% with a standard deviation relative to coverage of the underlying genome (i.e., lower coverage means higher standard deviation, and vice versa). We measured the fit of each allele frequency histogram to this expected distribution by counting the number of filtered SNPs that fell within one standard deviation of its corresponding hypothetical binomial distribution. This 2D plot was visualized with *ggplot2* in R. Taxa were then grouped into two categories, haploid mitosis and above haploid mitosis using the following evidence: 1) existing literature supporting well documented life cycles, 2) kmer histograms with two or more peaks, 3) high quality SNPs at a density of >0.002 and >50% of SNP allele frequencies within 1 SD of 0.5, 4). In cases where the data were unclear, typically higher density of SNPs with non-diploid-like allele frequency distribution, the taxon was coded as uncertain.

4.4 Results

Phylogenomic analyses reveals a robustly supported paraphyly of zoosporic fungi.

We generated draft genome assemblies for 68 previously-unsequenced zoosporic fungi. Our analysis of one single cell (PSC023) of the alga *Micrasterias* cf. *truncata* parasitized by a chytrid revealed two fungal genomes, one of which grouped with Rhizophydiales and the other with Rozellomycota. Assembly sizes ranged from 11.70 - 81.19 Mb, with gene numbers of 5,512 - 16,599, and genome completeness values of 34.30% - 94.99% *BUSCO* completeness values (*fungi_odb10*, *BUSCO* protein mode) (Table C1). We used the *BUSCO fungi_odb10* ortholog set of 758 markers to search our genomes for a gene set enriched in single copy orthologs. After filtering our data for genes with low occupancy of taxa and high paralogy, we limited our data set to 487 markers. The average of occupancies following filtering was 82.02% with a range of: 69.34% - 96.35%.

Our phylogenomic reconstructions based on concatenation covering 197,423 amino acid positions generated a robustly supported tree by ML analysis with 100% bootstrap support for all nodes (Figure 4.1; see Figure C1 for ML tree with all support values). We recovered a paraphyletic grade of 5 lineages containing zoosporic fungi: Rozellomycota, Aphelidiomycota, Chytridiomyceta, Blastocladiomycota, and Olpidiomycota, in this order **(**Figure 4.1**)**. These relationships are largely consistent with other phylogenomic analyses of zoosporic fungi, although many studies place Blastocladiomycota closer to the fungal root than Chytridiomyceta (Chang et al., 2021; Li et al., 2021; Torruella et al., 2018), whereas others place Chytridiomyceta more basal (Galindo et al., 2020; James et al., 2013). Within Chytridiomyceta, the Monoblepharidomycota is sister to the Neocallimastigomycota. The relationships within classes and orders of Chytridiomycota showed poor support with one exception. Strong support was observed for a group containing Rhizophydiales+Spizellomycetales+Rhizophlyctidales+*Blyttiomyces helicus* (Rhizophydiomycotina nom. prov.).

Olpidium bornovanus was placed as the most recent zoosporic taxa to diverge from the rest of the terrestrial fungi. Among the terrestrial fungi, Zoopagomycota was recovered as diverging first, with Mucoromycota supported as the sister clade to Dikarya.

Gene trees provide support for some but not all controversial nodes

Because of the often biased perspective of phylogenetic support based on nonparametric bootstrapping (Rokas and Carroll, 2006), we also assessed support from individual genes via gene concordance factors and internode certainty, which are highly conservative, less biased metrics based on splits or quartets in underlying gene trees. These results show support across genes is mostly consistent within a taxonomic order, however, interordinal relationships are rarely supported by these measures. We also generated a coalescence-based tree using individual gene trees with *ASTRAL* (Figure C2). The ASTRAL tree was largely congruent with the concatenated tree, with only 9 nodes differing. In the concatenated analysis *Polychytrium* was placed as sister to Chytridiales, whereas in the ASTRAL tree it was sister to Chytridiales+Rhizophydiomycotina. In both analyses, the newly described algal parasite *Quaeritorhiza haematococcus* grouped with Lobulomycetales. However, the relationship between this clade, *Caulochytrium*, Cladochytriales, Synchytriales, is weakly supported and different between both ASTRAL and concatenated analyses. In the concatenated phylogeny the enigmatic *Basidiobolus* groups as sister to the remaining Zoopagomycota, whereas in the ASTRAL tree, it is sister to Mucoromycota. Finally, in the concatenated tree Neocallimastigomycota groups with Monoblepharidomycota, whereas in the ASTRAL tree, Monoblepharidomycota groups with Chytridiomycota.

Because on one hand bootstrap and local posterior probabilities provide strong support for the controversial nodes, whereas gCF and IC values generally indicate minimal support, we queried individual genes in order to test whether they show significant support for one quartet resolution relative to the other two resolutions (Smith et al., 2020). Constrained searches consistent with each of the three quartets were conducted, removing all genes where all constraints were less likely (> 2 logL units) than unconstrained. Using the quartet resolving subphyla of Ascomycota as a control for the method, we recovered strong support among individual genes to support the currently accepted hypothesis of Pezizomycotina and Saccharomycotina as sister (Figure 4.2F). Applying this result to 5 controversial nodes, we found that individual genes generally supported the relationship in the concatenated phylogeny, even when this conflicted with the ASTRAL tree (Figure 4.2A-E). The support is particularly convincing in that the genes favor Neocallimastigomycota with Monoblepharidomycota, rather than Monoblepharidomycota with Chytridiomycota. Of the three resolutions of the quartet containing *Olpidium*, its placement as sister to terrestrial fungi has clear support. Blastocladiomycota was supported by 44% of genes trees as branching with terrestrial fungi and *Olpidium* (Q1), with 32% of trees favoring Blastocladiomycota branching before Chytridiomycota (Q2). Two genes strongly support Blastocladiomycota with terrestrial fungi and *Olpidium*: 26329 (RPA2) and 359482 (SEC22).

Most major branches in the fungi are diplontic

We used our bipartite kmer and allele frequency (AF) approach to infer ploidy for 112/137 of the taxa included in our phylogenomic analyses. The resulting kmer and AF histograms were systematically binned by ploidy based on their similarity to canonical examples of kmer (Figure 4.3A) and AF (Figure 4.3B) histograms, in addition to measured SNP density post-filtering. We identified 62 as diploid, 59 as haploid, 14 as uncertain, and 1 triploid. Although a portion of our histograms evidenced canonical distributions (as is shown in Figure 4.3A,B), there were many others that showed non-canonical distributions. These non-canonical distributions appear to result from low sequencing depth, high assembly fragmentation, whole genome amplification, and suboptimal read mapping, among other factors. In general, kmer histograms were unreliable at low sequencing coverage, and we therefore relied more heavily on AF histograms, and their associated SNP densities, to make ploidy calls in marginal cases.

When the SNP density of each genome was plotted against the proportion of SNPs falling within 1 SD of depth-scaled binomial distributions (i.e., the “expected” range), we observed that genomes assigned to either haploid or diploid/triploid ploidy clustered into two groups (Figure 4.3C). Haploid-annotated genomes form a tight cluster at low SNP-densities (mean = $4.74 \cdot 10^{-5}$; sd = $6.78 \cdot 10^{-5}$) and low numbers of SNPs within the expected range (mean = 17.97%; sd = 20.49%) (Figure 4.3C). On the other hand, diploid genomes form a broad cluster at high SNP densities (mean = $2.02 \cdot 10^{-3}$; sd = $2.51 \cdot 10^{-3}$) and proportions of SNPs occurring within the expected range (mean = 42.38%, sd = 21.09%) (Figure 4.3C). The relatively large diploid ellipse is indicative of the noisy signal of heterozygosity that characterizes our set of diploid genomes. In many cases, this noise is probably biologically relevant, a product of true variability in genomic heterozygosity and allelic richness across species. As such, our method is likely to underestimate diploidy or higher ploidy, which could influence our reconstruction of ploidy as a character state in Fungi.

According to our marginal ancestral state reconstruction, most of the ancestral nodes in the tree are reconstructed with diploidy as the more likely dominant phase of the life cycle (Figure 4.3D). According to this analysis, haplontic life cycles were derived independently multiple times in

Fungi. We infer the probability of a diploid or higher MRCA as more probable than a haploid MRCA for all phyla diverging prior to the Mucoromycota, except for the Neocallimastigomycota, to be >60%, with the MRCA of all Fungi having a 72.59% probability of being diploid (Figure 4.3E). Our maximum likelihood ancestral state reconstructions were strongly corroborated using a Bayesian MCMC stochastic ancestral state reconstruction based on 1,000 simulations (Figure 4.3E).

Diverse EDF lineages show independent loss of ancestral traits

Flagellation of reproductive propagules is a shared characteristic of many Opisthokonts, but has been lost multiple times the Fungi, and the flagellated fungi are paraphyletic (Figure 4.1). Though the core Chytridiomyceta and Blastocladiomycota still widely possess flagellated zoospores, *Hyaloraphidium curvatum* in the Monoblepharidomycota represents one lineage that has lost a flagellated state entirely. The first branch on the fungal tree groups the water mold parasites in *Rozella* with taxa classified as either short-branch Microsporidia or part of a larger Cryptomycota/Rozellomycota that parasitize amoebae or zooplankton and lack flagellation, *Paramicrosporidium* and *Mitosporidium*. Our single cell genome from the Rozellomycota (PSC023) appeared to possess a flagellum on the basis of BLASTp searches against the proteome using 10 flagellar proteins sequences as query.

The utilization of the cofactor cobalamin in metabolic pathways is assumed to be lacking in Dikarya fungi based on the absence of cobalamin-associated enzymes. However, recent searches of genomes have found that most early diverging taxa do possess a subset of these enzymes. Our search for eight enzymes in our collected genomes found cobalamin-associated enzymes to be generally present in most flagellated fungi examined, with certain lineages containing subsets of enzymes or being completely devoid of detectable enzymes (Figure 4.1). Higher taxonomic groups with no cobalamin-associated enzymes were the Neocallimastigomycota, as previously reported, and the majority of the Rhizophydiales, including species of the amphibian pathogenic genus *Batrachochytrium*. However, in nearly all orders there are some taxa with all cobalamin-associated enzyme genes, while others lack any of the known genes. These data speak to a dependence on cobalamin for many hundreds of millions of years as these phyla separated into major lineages. Notably, many parasitic species, e.g., *Caulochytrium protostelioides*, *Coelomomyces lativittatus*, *Mitosporidium daphniae*, appear to lack cobalamin-dependent enzyme genes, perhaps indicative that parasitism can lead to a reduction in cobalamin-dependence due to scarcity within the host.

Similar to the utilization of cobalamin, selenoproteins, or proteins containing the twenty-first amino acid selenocysteine, have been considered widely absent from Dikarya fungi, though they have recently been detected in the other fungal lineages (Mariotti et al., 2019). Our search of the collected genomes has found a scant scattering of selenocysteine across multiple Chytridiomyceta and Blastocladiomycota taxa (Figure 4.1). Within the Blastocladiomycota, the potential for selenocysteine was present in only *Paraphysoderma sedebokerense*. The Monoblepharidiomycota contained two taxa that utilized selenocysteine, *Hyaloraphidium curvatum* and *Gonapodya prolifera*. The Chytridiomycetes had the most diverse taxonomic assemblage of selenocysteine-utilizing taxa, with targets in the Polychytriales (*Polychytrium aggregatum*), Synchytriales (*Synchytrium microbalum*), and an undescribed lineage (*Quaeritorhiza haematococci*). In practically all cases of positive selenocysteine detections, these taxa represent poorly sampled, divergent groups, with potentially more species capable of selenoprotein utilization.

4.5 Discussion

The most important result from this research is the finding that most of the non-Dikarya phyla of the Fungi are characterized by diploidy. Although there were indications that non-dikaryotic fungi can be diploid based on population genomics and single cell genome sequencing (Ahrendt et al., 2018; Rosenblum et al., 2013), our results extend diploidy to additional lineages, such as Aphelidiomycota, Entomophthoromycotina, Monoblepharidomycota, and Olpidiomycota. Our results are based primarily on heterozygosity of cultivated strains, which presumably represent the dominant life cycle stage of these organisms. This finding demands a reconsideration of the canonical life cycle of fungi as being primarily haplontic and lacking mitosis at the diploid stage. Instead, the MRCAs of Fungi and most phyla of fungi were likely diplontic, and/or transitions between haplontic - diplontic life cycles are fluid and frequent in the non-Dikarya lineages.

Indeed, we found that transitions to haplontic life cycles have occurred multiple times independently in major groups of non-Dikarya (e.g., Mucoromycota, Neocallimastigomycota, Spizellomycetales).

One caveat to interpreting our ploidy estimates is that because they rely on genome-wide heterozygosity, homozygous genomes will appear haploid. Therefore, our estimation of ploidy is likely an underestimate. On the other hand, heterozygosity can be overestimated by sequencing or genome amplification errors or from mis-mapping of reads, for example from recent gene duplications. However, our SNP filtering approach drastically reduced the size of the final SNP set, likely favoring the removal of true heterozygosity over including falsely heterozygous positions. Our pipeline showed robust inference of high levels of heterozygosity in some taxa well appreciated as being diploid, such as *Batrachochytrium* spp. (Farrer et al., 2017; Rosenblum et al., 2013) and *Allomyces javanicus* (Emerson, 1941), and confirmed haploidy is some well-known taxa, such as the gametophyte stage of *Coelomomyces* (Whisler et al., 1975), and Mucorales (Lee et al., 2010). On the other hand, the life cycles of some zoosporic fungi, including *Chytriomyces hyalinus* (Moore and Miller, 1973), *Catenaria anguillulae* (Olson and Reichle, 1978), and *Paraphysoderma sedebokerense* (Letcher et al., 2016) were suspected to be haplontic. All three of these species were observed to have highly heterozygous genomes. For the vast majority of zoosporic fungi, there is no information on life cycles, and some large lineages such as Neocallimastigomycota and Spizellomycetes are only known as asexual. What has been lacking in the vast majority of zoosporic taxa is either a confirmation of meiosis via microscopy or genetic analysis. In some instances, such as the Blastocladiomycota taxa *P. sedebokerensis* (Letcher et al., 2016) and *Catenaria anguillulae* (Olson and Reichle, 1978), meiosis is documented by ultrastructural determination of synaptonemal complexes and was used to define life cycles as haplontic.

How can these cytological observations that seem to suggest haplontic life cycles be reconciled with heterozygosity data which suggest diploidy in the dominant vegetative phase? One possible explanation is that the life cycles of these fungi cycle between diploid and tetraploid. Although this is conceivable, it is also unlikely given the errors that are likely to occur in autotetraploid meiosis (Comai, 2005). Despite the finding that at least *Rozella allomycis* is triploid, no cases of tetraploidy were uncovered as might be expected if this was the case. Another possibility is that many taxa have undergone recent whole genome duplication creating a scenario of two similar genomes wherein the heterozygosity is actually divergence between paralogs. If this was the case one might expect to see more divergent alleles and patchy heterozygosity as duplicated regions of the genome either diverge or are lost over time. Both the low level of heterozygosity and evenness of heterozygosity across the genome (data not shown) indicate that this model is unlikely to be accurate. However, there is at least one example where whole genome duplication appears to have occurred. *Cladochytrium replicatum* was identified to have a substantial amount of segmental duplication (~70% of assembly duplicated), but with low amino acid identity between copies (~83%). Unlike with other assemblies, the mapping on this genome perhaps unsurprisingly did not show typical binomial distribution of allele frequencies, likely as a result of poor mapping of reads to the correct paralog. Another possible explanation for the discrepancy between expected and observed rates of diploidy is the possibility that the ultrastructural studies are misinterpreted. There is precedent of synapsis without meiosis in some somatic cells, such as *Drosophila* (McKee, 2004), but the presence of tripartite synaptonemal complexes is considered a meiosis specific hallmark. Both enhanced studies relating DNA replication and pairing to formation of synaptonemal complexes and studies tracing genetic segregation in appropriate zoosporic fungal models are needed to resolve these discrepancies.

Ploidy is not homogenously distributed across the non-dikaryotic lineages. For example, the Neocallimastigomycota and Spizellomycetes were all estimated as haploid. For Blastocladiomycota, we identified both haploid and diploid genomes, consistent with the haplo-diplontic or alternation of generations known for the group. The importance of the placement of Blastocladiomycota as an intermediate step between diplontic or haplontic life cycles is diminished, because our results show that lineages closer to Dikarya, specifically Olpidiomycota and Zoopagomycota, are also often diploid. Within the Mucoromycota, we did not identify any diploid genomes, however, diploidy is known throughout the Dikarya, in particular among the yeasts which are also scattered throughout lineages of Dikarya (Nagy et al., 2014), as well as the

Armillaria mushrooms which are known for their incredibly large genetic individuals (Ullrich and Anderson, 1978).

Beyond a better basic understanding of fungal life cycles, what are the implications of diploidy among the EDF? The advantages and disadvantages of ploidy differences have been much discussed (Otto and Gerstein, 2008). Diploidy is associated with larger organisms as it buffers from somatic mutations, while haploidy is associated with parasitism (Nuismer and Otto, 2004). Proximately, haploidy can increase the rate of adaptive evolution over diploids depending on the dominance of beneficial mutations (Zeyl et al., 2003). None of these mechanisms are likely to fully explain diploidy of EDF. There are multicellular and pathogenic taxa found in many of the EDF phyla, and the particularly haploid lineages are not distinguished by any obvious additional characteristics. In contrast, one model posits that fitness differs depending on the ploidy of the cell due to differences in cell or nucleus size, given that cells with higher ploidy typically are larger (Weiss et al., 1975). EDF contrast with Dikarya generally in that they are coenocytic with individual cells or hyphal filaments with many nuclei, whereas Dikarya have haploid nuclei that are more compartmentalized. As a possible clue that nuclear number per cell may be critical for determining ploidy in fungi, by and large most diploid Dikarya grow as yeasts rather than hyphae. On the other hand, the most coenocytic of the fungi, the Mucoromycota, containing the Glomeromycotina, were all haploid based on our results.

The sequencing of new zoosporic fungal genomes facilitated a robustly supported phylogenetic inference of early divergences in fungi. The results here, and a recent analysis by Galindo et al. (2021), indicate that the Blastocladiomycota are more closely related to the terrestrial fungi than the Chytridiomyceta. By the end of the Cambrian, we estimate that most of the EDF phyla and orders of Chytridiomycota were already diversifying, consistent with the fossil record evidencing a wide diversity chytrid-like fossils by the Devonian (Berbee et al., 2020) and consistent with diversification of fungi alongside algae, which were the dominant photosynthetic life at this time period (Chang et al., 2015). Yet, support for relationships dating to this period are weak, with several of the nodes that conflict between the ASTRAL and ML trees dating from around the Cambrian.

Our analyses show that support for contentious nodes varies greatly between genes. In some cases, such as the clade containing Neocallimastigomycota and Monoblepharidomycota, a majority of decisive genes lend support. On the other hand, for the most difficult to place Blastocladiomycota, there is only marginal support for the ML relationship based on a plurality of genes, a likely factor explaining the oft changing location of the taxon in phylogenetic analyses. Most of the other poorly supported nodes, such as placement of *Polychytrium* and *Caulochytrium* are those taxa represented with single samples. This begs for the discovery and inclusion of additional taxon sampling. For taxa that have never been cultivated, single cell approaches as demonstrated here and elsewhere (Davis et al., 2019; Galindo et al., 2020) can be used to increase taxon sampling. One novel relationship that we uncover here with strong support is the unification of Rhizophydiales/Spizellomycetales/Rhizophlyctidales/*Blyttiomyces helicus*, which is supported by a zoospore ultrastructural character, the absence of an electron dense plug at the base of the flagellum.

Although the phylogenies supporting first divergences in Kingdom Fungi have been clarified recently through whole genome sequencing, the branch at which to designate true Fungi from non-fungal opisthokonts has been subject to considerable debate (James et al., 2020; Karpov et al., 2014; Torruella et al., 2018). The majority of the zoosporic fungi included in this study feed like most fungi, which is across a chitinous cell wall, and some species are even mycelium-like. The discovery that the Aphelidiomycota, Rozellomycota, and Microsporidia that produce chitinous spores are allied with Fungi yet feed primarily across a naked plasma membrane, and in some cases even undergo phagocytosis, has blurred the distinction between Fungi and Metazoa. As shown here and in other recent studies, it is clear that rather than a stark distinction between animals/protists and fungi, the boundary, however you describe it, is far from distinct, and that Fungi gradually diverged over time to become Dikarya-like. Numerous traits of EDF such as amoeboid and flagellated motility of spores (i.e., crawling and swimming), vitamin and mineral dependencies, and cell cycle proteins are just as striking as the oft touted hallmark of Fungi as osmotrophy across a cell wall. Fungal osmotrophy, while striking and characteristic of all Dikarya, is not a synapomorphy for even a more restricted definition of Fungi (Richards and Talbot, 2018).

There has been more renewed focus on the EDL of Fungi in recent years, in part due to phylogenetic studies which revealed a much hidden and underappreciated diversity in gene content (Hérivaux et al., 2017; Medina et al., 2016). Here, we extend this genetic diversity to life cycle differences, and highlight trends in cell structure and biochemistry that are shifting from characteristics of the opisthokont ancestor to those of the Dikarya. These data suggest a pressing need for a reconsideration of life cycle evolution in the fungi and beg for detailed studies on basic Mendelian genetics and cytology in these overlooked organisms. Given that multiple phylogenomic analyses arrive at different resolutions of the most controversial nodes, it is clear that having whole genome sequences of many taxa is not in and of itself sufficient to resolve these relationships. Instead, careful consideration of each step of phylogenetic analysis must be made: homolog retrieval, alignment, model selection, and tree construction. The lack of phylogenetic resolution also suggests areas of the tree that will require further taxon sampling. Through the process of reciprocal illumination, we may identify morphological traits or rare genomic changes, such as gene fusions or insertion/deletions in proteins that resolve some of these difficult relationships.

4.6 Literature Cited

Ahrendt, S.R., Quandt, C.A., Ciobanu, D., Clum, A., Salamov, A., Andreopoulos, B., Cheng, J.-F., Woyke, T., Pelin, A., Henrissat, B., Reynolds, N.K., Benny, G.L., Smith, M.E., James, T.Y., Grigoriev, I.V., 2018. Leveraging single-cell genomics to expand the fungal tree of life. Nat. Microbiol. 3, 1417–1428. https://doi.org/10.1038/s41564-018-0261-0

Amses, K.R., Davis, W.J., James, T.Y., 2020. SCGid: a consensus approach to contig filtering and genome prediction from single-cell sequencing libraries of uncultured eukaryotes. Bioinformatics 36, 1994–2000. https://doi.org/10.1093/bioinformatics/btz866

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477.

Bushnell, B., 2014. sourceforge.net/projects/bbmap: BBMap.

Berbee, M.L., Strullu-Derrien, C., Delaux, P.-M., Strother, P.K., Kenrick, P., Selosse, M.-A., Taylor, J.W., 2020. Genomic and fossil windows into the secret lives of the most ancient fungi. Nat. Rev. Microbiol. 18, 717–730. https://doi.org/10.1038/s41579-020-0426-8

Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., Yandell, M., 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18, 188–196.

Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973.

Casbon, J., 2016. PyVCF. Github.

Chang, Y., Desirò, A., Na, H., Sandor, L., Lipzen, A., Clum, A., Barry, K., Grigoriev, I.V., Martin, F.M., Stajich, J.E., Smith, M.E., Bonito, G., Spatafora, J.W., 2019. Phylogenomics of Endogonaceae and evolution of mycorrhizas within Mucoromycota. New Phytol. 222, 511–525. https://doi.org/10.1111/nph.15613

Chang, Y., Rochon, D., Sekimoto, S., Wang, Y., Chovatia, M., Sandor, L., Salamov, A., Grigoriev, I.V., Stajich, J.E., Spatafora, J.W., 2021. Genome-scale phylogenetic analyses confirm Olpidium as the closest living zoosporic fungus to the non-flagellated, terrestrial fungi. Sci. Rep. 11, 3217. https://doi.org/10.1038/s41598-021-82607-4

Chang, Y., Wang, S.S., Sekimoto, S., Aerts, A.L., Choi, C., Clum, A., LaButti, K.M., Lindquist, E.A., Ngan, C.Y., Ohm, R.A., Salamov, A.A., Grigoriev, I.V., Spatafora, J.W., Berbee, M.L., 2015. Phylogenomic analyses indicate that early fungi evolved digesting cell walls of algal ancestors of land plants. Genome Biol. Evol. 7, 1590–1601. https://doi.org/10.1093/gbe/evv090

Comai, L., 2005. The advantages and disadvantages of being polyploid. Nat. Rev. Genet. 6, 836–846. https://doi.org/10.1038/nrg1711

Davis, W.J., Amses, K.R., Benny, G.L., Carter-House, D., Chang, Y., Grigoriev, I., Smith, M.E., Spatafora, J.W., Stajich, J.E., James, T.Y., 2019. Genome-scale phylogenetics reveals a monophyletic Zoopagales (Zoopagomycota, Fungi). Mol. Phylogenet. Evol. 133, 152–163. https://doi.org/10.1016/j.ympev.2019.01.006

Eddy, S.R., HMMER development team, 2015. HMMER.

Emerson, R., 1941. An experimental study of the life cycles and taxonomy of Allomyces. Lloydia 4, 77–144.

Farrer, R.A., Martel, A., Verbrugghe, E., Abouelleil, A., Ducatelle, R., Longcore, J.E., James, T.Y., Pasmans, F., Fisher, M.C., Cuomo, C.A., 2017. Genomic innovations linked to infection strategies across emerging pathogenic chytrid fungi. Nat. Commun. 8, 11. https://doi.org/10.1038/ncomms14742

Galindo, L.J., López-García, P., Torruella, G., Karpov, S., Moreira, D., 2020. Phylogenomics of a new fungal phylum reveals multiple waves of reductive evolution across Holomycota. bioRxiv 2020.11.19.389700. https://doi.org/10.1101/2020.11.19.389700

Geneious. 2019. https://www.geneious.com.

Hérivaux, A., Bernonville, T.D. de, Roux, C., Clastre, M., Courdavault, V., Gastebois, A., Bouchara, J.-P., James, T.Y., Latgé, J.-P., Martin, F., Papon, N., 2017. The identification of phytohormone receptor homologs in early diverging fungi suggests a role for plant sensing in land colonization by fungi. mBio 8. https://doi.org/10.1128/mBio.01739-16

Idnurm, A., James, T.Y., Vilgalys, R., 2007. Sex in the rest: mysterious mating in the Chytridiomycota and Zygomycota, in: J. Heitman, J.K., J.W. Taylor, and L.A. Casselton (Ed.), Sex in Fungi: Molecular Determination and Evolutionary Implications. ASM Press, Washington, D. C., pp. 407–418.

James, T.Y., Berbee, M.L., 2012. No jacket required- new fungal lineage defies dress code. BioEssays 34, 94–102.

James, T.Y., Kauff, F., Schoch, C., Matheny, P.B., Hofstetter, V., Cox, C., Celio, G., Gueidan, C., Fraker, E., Miadlikowska, J., Lumbsch, H.T., Rauhut, A., Reeb, V., Arnold, A.E., Amtoft, A., Stajich, J.E., Hosaka, K., Sung, G.-H., Johnson, D., O'Rourke, B., Crockett, M., Binder, M., Curtis, J.M., Slot, J.C., Wang, Z., Wilson, A.W., Schüßler, A., Longcore, J.E., O'Donnell, K., Mozley-Standridge, S., Porter, D., Letcher, P.M., Powell, M.J., Taylor, J.W., White, M.M., Griffith, G.W., Davies, D.R., Humber, R.A., Morton, J.B., Sugiyama, J., Rossman, A.Y., Rogers, J.D., Pfister, D.H., Hewitt, D., Hansen, K., Hambleton, S., Shoemaker, R.A., Kohlmeyer, J., Volkmann-Kohlmeyer, B., Spotts, R.A., Serdani, M., Crous, P.W., Hughes, K.W., Matsuura, K., Langer, E., Langer, G., Untereiner, W.A., Lücking, R., Büdel, B., Geiser, D.M., Aptroot, A., Diederich, P., Schmitt, I., Schultz, M., Yahr, R., Hibbett, D., Lutzoni, F., McLaughlin, D., Spatafora, J., Vilgalys, R., 2006a. Reconstructing the early evolution of the fungi using a six gene phylogeny. Nature 443, 818–822.

James, T.Y., Letcher, P.M., Longcore, J.E., Mozley-Standridge, S.E., Porter, D., Powell, M.J., Griffith, G.W., Vilgalys, R., 2006b. A molecular phylogeny of the flagellated fungi (Chytridiomycota) and description of a new phylum (Blastocladiomycota). Mycologia 98, 860–871.

James, T.Y., Pelin, A., Bonen, L., Ahrendt, S., Sain, D., Corradi, N., Stajich, J.E., 2013. Shared signatures of parasitism and phylogenomics unite Cryptomycota and Microsporidia. Curr. Biol. 23, 1548–1553. https://doi.org/10.1016/j.cub.2013.06.057

James, T.Y., Porter, T.M., Martin, W.W., 2014. 7 Blastocladiomycota, in: McLaughlin, D.J., Spatafora, J.W. (Eds.), Systematics and Evolution: Part A, The Mycota. Springer, Berlin, Heidelberg, pp. 177–207. https://doi.org/10.1007/978-3-642-55318-9_7

James, T.Y., Stajich, J.E., Hittinger, C.T., Rokas, A., 2020. Toward a fully resolved Fungal Tree of Life. Annu. Rev. Microbiol. 74, 291–313. https://doi.org/10.1146/annurev-micro-022020-051835

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

Karpov, S.A., Mamkaeva, M.A., Aleoshin, V.V., Nassonova, E., Lilje, O., Gleason, F.H., 2014. Morphology, phylogeny, and ecology of the aphelids (Aphelidea, Opisthokonta) and proposal for the new superphylum Opisthosporidia. Front. Microbiol. 5, 112. https://doi.org/10.3389/fmicb.2014.00112

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.

Lee, S.C., Ni, M., Li, W., Shertz, C., Heitman, J., 2010. The evolution of sex: a perspective from the fungal kingdom. Microbiol. Mol. Biol. Rev. 74, 298–340.

Letcher, P.M., Lee, P.A., Lopez, S., Burnett, M., McBride, R.C., Powell, M.J., 2016. An ultrastructural study of Paraphysoderma sedebokerense (Blastocladiomycota), an epibiotic parasite of microalgae. Fungal Biol. 120, 324–337. https://doi.org/10.1016/j.funbio.2015.11.003

Letcher, P.M., Powell, M.J., 2014. Hypothesized evolutionary trends in zoospore ultrastructural characters in Chytridiales (Chytridiomycota). Mycologia 106, 379–396. https://doi.org/10.3852/13-219

Letcher, P.M., Powell, M.J., Churchill, P.F., Chambers, J.G., 2006. Ultrastructural and molecular phylogenetic delineation of a new order, the Rhizophydiales (Chytridiomycota). Mycol. Res. 110, 898–915. https://doi.org/10.1016/j.mycres.2006.06.011

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079.

Li, Y., Steenwyk, J.L., Chang, Y., Wang, Y., James, T.Y., Stajich, J.E., Spatafora, J.W., Groenewald, M., Dunn, C.W., Hittinger, C.T., Shen, X.-X., Rokas, A., 2021. A genome-scale phylogeny of the kingdom Fungi. Curr. Biol. S0960-9822, 10.1016/j.cub.2021.01.074. https://doi.org/10.1016/j.cub.2021.01.074

Liu, Y., Steenkamp, E.T., Brinkmann, H., Forget, L., Philippe, H., Lang, B.F., 2009. Phylogenomic analyses predict sistergroup relationship of nucleariids and Fungi and paraphyly of zygomycetes with significant support. BMC Evol. Biol. 9, 11. https://doi.org/10.1186/1471-2148-9-272

Longcore, J.E., Simmons, D.R., 2012. The Polychytriales ord. nov contains chitinophilic members of the rhizophlyctoid alliance. Mycologia 104, 276–294. https://doi.org/10.3852/11-193

Lücking, R., Huhndorf, S., Pfister, D.H., Plata, E.R., Lumbsch, H.T., 2009. Fungi evolved right on track. Mycologia 101, 810–822. https://doi.org/10.3852/09-016

Mariotti, M., Salinas, G., Gabaldón, T., Gladyshev, V.N., 2019. Utilization of selenocysteine in early-branching fungal phyla. Nat. Microbiol. 4, 759–765. https://doi.org/10.1038/s41564-018-0354-9

McKee, B.D., 2004. Homologous pairing and chromosome dynamics in meiosis and mitosis. Biochim. Biophys. Acta 1677, 165–180. https://doi.org/10.1016/j.bbaexp.2003.11.017

Medina, E.M., Turner, J.J., Gordân, R., Skotheim, J.M., Buchler, N.E., 2016. Punctuated evolution and transitional hybrid network in an ancestral cell cycle of fungi. eLife 5, e09492. https://doi.org/10.7554/eLife.09492

Minh, B.Q., Hahn, M.W., Lanfear, R., 2020. New Methods to Calculate Concordance Factors for Phylogenomic Datasets. Mol. Biol. Evol. 37, 2727–2733. https://doi.org/10.1093/molbev/msaa106

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., Lanfear, R., 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. 37, 1530–1534.

Moore, E.D., Miller, C.E., 1973. Resting body formation by rhizoidal fusion in Chytriomyces hyalinus. Mycologia 65, 145–154. https://doi.org/10.1080/00275514.1973.12019413

Nagy, L.G., Ohm, R.A., Kovács, G.M., Floudas, D., Riley, R., Gácser, A., Sipiczki, M., Davis, J.M., Doty, S.L., de Hoog, G.S., Lang, B.F., Spatafora, J.W., Martin, F.M., Grigoriev, I.V., Hibbett, D.S., 2014. Latent homology and convergent regulatory evolution underlies the repeated emergence of yeasts. Nat. Commun. 5, 4471. https://doi.org/10.1038/ncomms5471

Naranjo-Ortiz, M.A., Gabaldón, T., 2019. Fungal evolution: diversity, taxonomy and phylogeny of the Fungi. Biol. Rev. 94, 2101–2137. https://doi.org/10.1111/brv.12550

Nuismer, S.L., Otto, S.P., 2004. Host-parasite interactions and the evolution of ploidy. Proc. Natl. Acad. Sci. U. S. A. 101, 11036–11039.

Olson, L.W., Reichle, R., 1978. Meiosis and diploidization in the aquatic Phycomycete Catenaria anguillulae. Trans. Br. Mycol. Soc. 70, 423–437.

Orłowska, M., Steczkiewicz, K., Muszewska, A., 2021. Utilization of cobalamin is ubiquitous in early-branching fungal phyla. Genome Biol. Evol. https://doi.org/10.1093/gbe/evab043

Otto, S.P., Gerstein, A.C., 2008. The evolution of haploidy and diploidy. Curr. Biol. 18, R1121–R1124.

Palmer, J., Stajich, J., 2019. nextgenusfs/funannotate: funannotate v1.5.3. https://doi.org/10.5281/zenodo.2604804

Powell, M.J., Letcher, P.M., James, T.Y., 2017. Ultrastructural characterization of the host parasite interface between *Allomyces anomalus* (Blastocladiomycota) and *Rozella allomycis* (Cryptomycota). Fungal Biol. 121, 561–572. https://doi.org/10.1016/j.funbio.2017.03.002

Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 5, e9490.

Prostak, S.M., Robinson, K.A., Titus, M.A., Fritz-Laylin, L.K., 2021. The actin networks of chytrid fungi reveal evolutionary loss of cytoskeletal complexity in the fungal kingdom. Curr. Biol. 31, 1192-1205.e6. https://doi.org/10.1016/j.cub.2021.01.001

Revell, L.J., 2012. phytools: An R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution.

Richards, T.A., Leonard, G., Wideman, J.G., 2017. What defines the “Kingdom” Fungi? Microbiol. Spectr. 5, FUNK-0044-2017. https://doi.org/10.1128/microbiolspec.FUNK-0044-2017

Richards, T.A., Talbot, N.J., 2018. Osmotrophy. Curr. Biol. CB 28, R1179–R1180. https://doi.org/10.1016/j.cub.2018.07.069

Rokas, A., Carroll, S.B., 2006. Bushes in the Tree of Life. PLOS Biol. 4, e352.

https://doi.org/10.1371/journal.pbio.0040352

Rosenblum, E.B., James, T.Y., Zamudio, K.R., Poorten, T.J., Ilut, D., Rodriguez, D., Eastman, J.M., Richards-Hrdlicka, K., Joneson, S., Jenkinson, T.S., Longcore, J.E., Parra Olea, G., Toledo, L.F., Arellano, M.L., Medina, E.M., Restrepo, S., Flechas, S.V., Berger, L., Briggs, C.J., Stajich, J.E., 2013. Complex history of the amphibian-killing chytrid fungus revealed with genome resequencing data. Proc. Natl. Acad. Sci. U. S. A. 110, 9385–9390. https://doi.org/10.1073/pnas.1300130110

Rosling, A., Cox, F., Cruz-Martinez, K., Ihrmark, K., Grelet, G.-A., Lindahl, B.L., Menkis, A., James, T.Y., 2011. Archaeorhizomycetes: unearthing an ancient class of ubiquitous soil fungi. Science 333, 876–879.

Ruiz-Herrera, J., Ortiz-Castellanos, L., 2019. Cell wall glucans of fungi. A review. Cell Surf. 5, 100022. https://doi.org/10.1016/j.tcsw.2019.100022

Safonova, Y., Bankevich, A., Pevzner, P.A., 2015. DipSPAdes: Assembler for highly polymorphic diploid genomes. J. Comput. Biol. 22, 528–545.

Sanderson, M.J., 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19, 301–302. https://doi.org/10.1093/bioinformatics/19.2.301

Santesmasses, D., Mariotti, M., Guigó, R., 2017. Computational identification of the selenocysteine tRNA (tRNASec) in genomes. PLOS Comput. Biol. 13, e1005383. https://doi.org/10.1371/journal.pcbi.1005383

Schloegel, L.M., Toledo, L.F., Longcore, J.E., Greenspan, S.E., Vieira, C.A., Lee, M., Zhao, S., Wangen, C., Ferreira, C.M., Hipolito, M., Davies, A.J., Cuomo, C.A., Daszak, P., James, T.Y., 2012. Novel, panzootic and hybrid genotypes of amphibian chytridiomycosis associated with the bullfrog trade. Mol. Ecol. 21, 5162–5177. https://doi.org/10.1111/j.1365-294X.2012.05710.x

Seppey, M., Manni, M., Zdobnov, E.M., 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods Mol. Biol. 1962, 227–245.

Seto, K., Van Den Wyngaert, S., Degawa, Y., Kagami, M., 2020. Taxonomic revision of the genus Zygorhizidium : Zygorhizidiales and Zygophlyctidales ord. nov. (Chytridiomycetes , Chytridiomycota). Fungal Syst. Evol. 5, 17–38. https://doi.org/10.3114/fuse.2020.05.02

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Simmons, D.R., Bonds, A.E., Castillo, B.T., Clemons, R.A., Glasco, A.D., Myers, J.M., Thapa, N., Letcher, P.M., Powell, M.J., Longcore, J.E., James, T.Y., 2020. The Collection of Zoosporic Eufungi at the University of Michigan (CZEUM): introducing a new repository of barcoded Chytridiomyceta and Blastocladiomycota cultures. IMA Fungus 11, 20. https://doi.org/10.1186/s43008-020-00041-z

Smith, S.A., Walker-Hale, N., Walker, J.F., Brown, J.W., 2020. Phylogenetic conflicts, combinability, and deep phylogenomics in plants. Syst. Biol. 69, 579–592. https://doi.org/10.1093/sysbio/syz078

Tedersoo, L., Sánchez-Ramírez, S., Kõljalg, U., Bahram, M., Döring, M., Schigel, D., May, T., Ryberg, M., Abarenkov, K., 2018. High-level classification of the Fungi and a tool for evolutionary ecological analyses. Fungal Divers. 90, 135–159.

https://doi.org/10.1007/s13225-018-0401-0

Torruella, G., Grau-Bové, X., Moreira, D., Karpov, S.A., Burns, J.A., Sebé-Pedrós, A., Völcker, E., López-García, P., 2018. Global transcriptome analysis of the aphelid Paraphelidium tribonemae supports the phagotrophic origin of fungi. Commun. Biol. 1, 231. https://doi.org/10.1038/s42003-018-0235-z

Ullrich, R.C., Anderson, J.B., 1978. Sex and diploidy in Armillaria mellea. Exp. Mycol. 2, 119–129.

Van der Auwera, G.A., O'Connor, B.D., 2020. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. O'Reilly Media.

Weete, J.D., Abril, M., Blackwell, M., 2010. Phylogenetic Distribution of Fungal Sterols. PLOS ONE 5, e10899. https://doi.org/10.1371/journal.pone.0010899

Weiss, R.L., Kukora, J.R., Adams, J., 1975. The relationship between enzyme activity, cell geometry, and fitness in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. 72, 794–798. https://doi.org/10.1073/pnas.72.3.794

Whisler, H.C., Zebold, S.L., Shemanchuk, J.A., 1975. Life history of Coelomomyces psorophorae. Proc. Natl. Acad. Sci. U. S. A. 72, 693–696.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis.

Zeyl, C., Vanderford, T., Carter, M., 2003. An evolutionary advantage of haploidy in large yeast populations. Science 299, 555–558. https://doi.org/10.1126/science.1078417

Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S., 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19, 153. https://doi.org/10.1186/s12859-018-2129-y

Zhou, X., Lutteropp, S., Czech, L., Stamatakis, A., von Looz, M., Rokas, A., 2019. Quartet-based computations of internode certainty provide robust measures of phylogenetic incongruence. Syst. Biol. 69, 308–324. https://doi.org/10.1093/sysbio/syz058

4.7 Figures

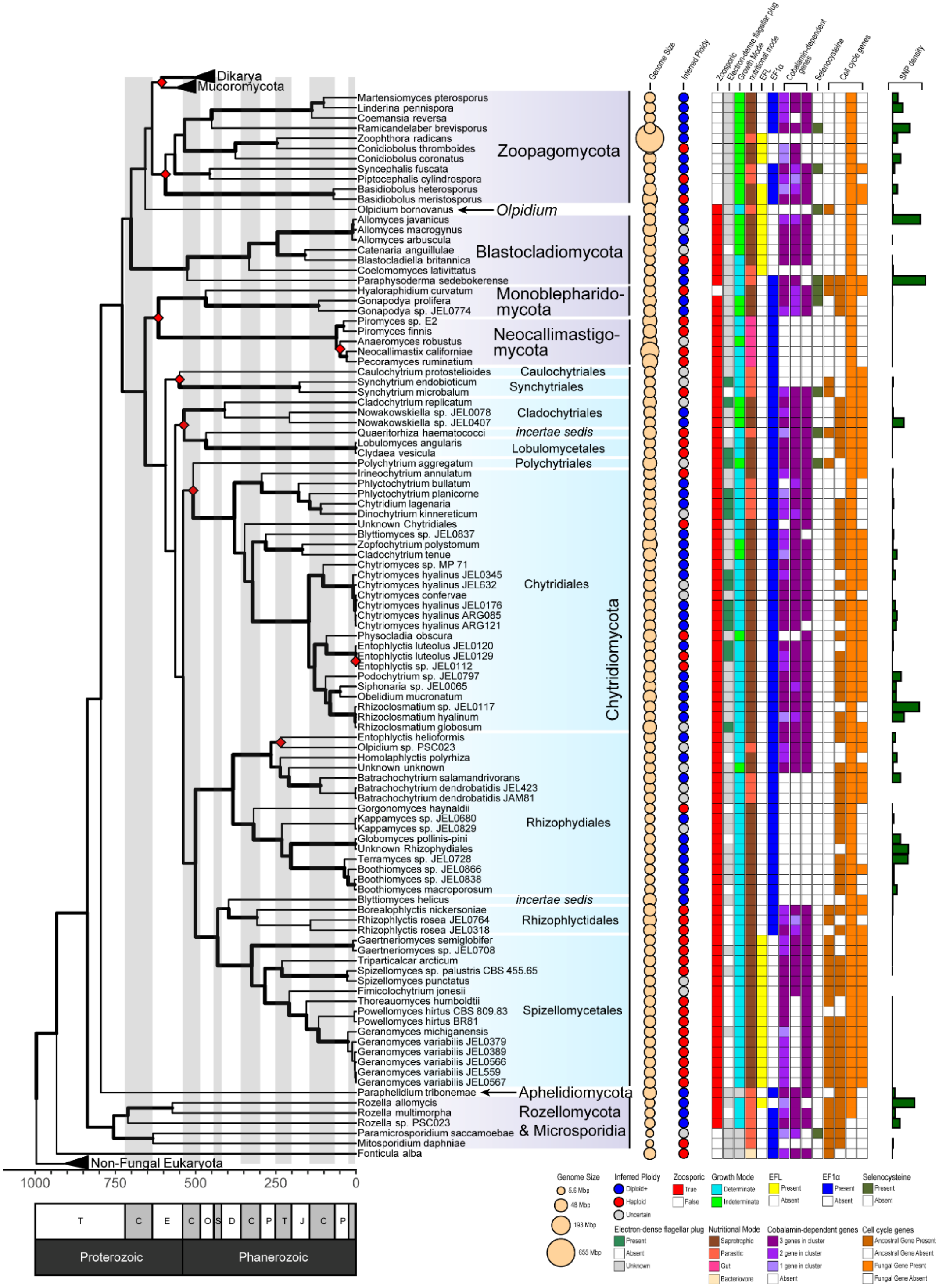


Figure 4.1. Annotated, time-calibrated concatenated ML tree of Kingdom Fungi, including 68 newly sequenced genomes of zoosporic fungi, based on 197,423 amino acid positions. All bootstraps at 100, edge thickness indicates gCF support, red diamonds indicate clades that were not present in ASTRAL tree. Various traits appear in tracks to the right as labeled and colored or sized according to the legend (bottom right). SNP density appears as horizontal bar chart on far right; complete absence of bars indicates taxa for which ploidy was not inferred at all.

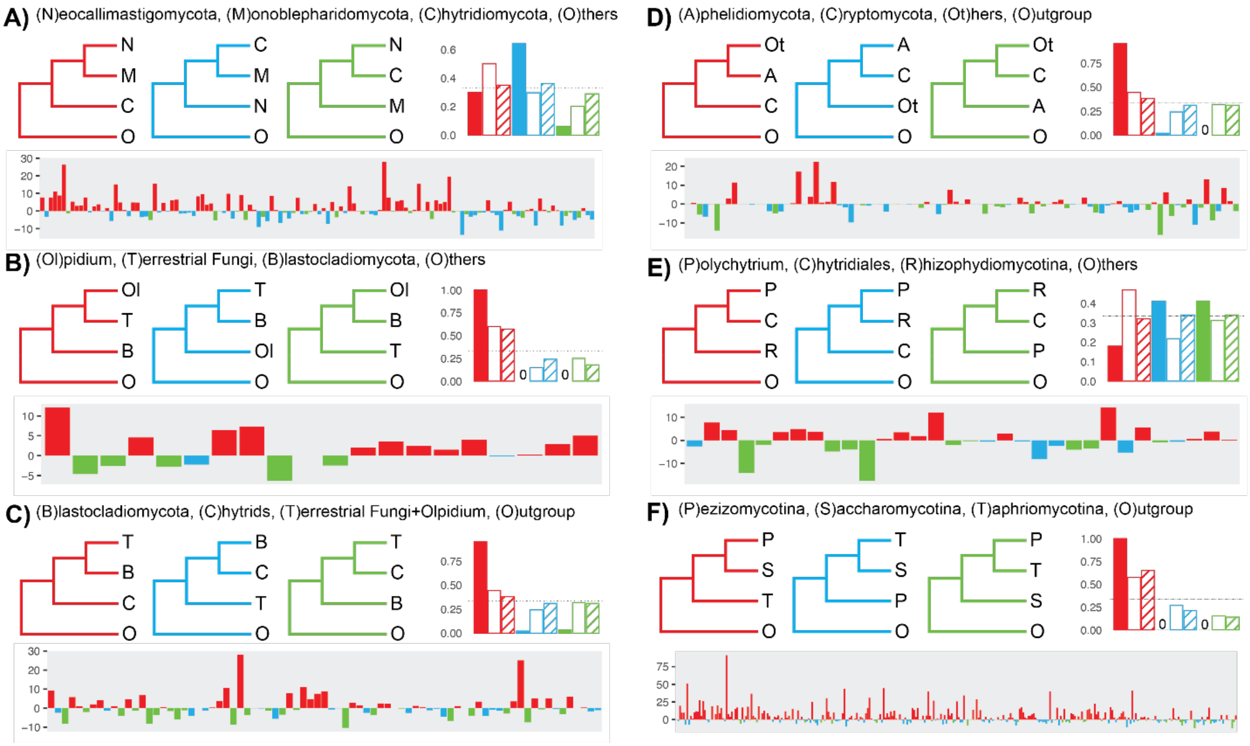


Figure 4.2. Assessment of support for controversial nodes via quartet analyses reveals conflict among genes but also instances where a majority of decisive genes support one quartet arrangement. For each controversial relationship, the three quartet resolutions are shown, with Q1 (red) the relationship in the ML tree, and Q2 (blue) and Q3 (green) the alternative relationships. The bar graph indicates support for each resolution. Solid bars represent local posterior probability values from ASTRAL. Open bars indicate the proportion of genes trees in which the likelihood of the indicated quartet was the highest after ML searches imposing constraints for each quartet. Only decisive genes in which the best scoring quartet under constrained searches was less than 2 logL units different from an unconstrained search. Striped bars indicate quartet frequencies across gene trees. In the bottom half of each panel is plotted the lnL difference between the best scoring quartet and the second best scoring quartet for decisive genes following constrained searches. Best scoring Q1 likelihoods are shown as positive values, and Q2 and Q3 likelihood differences are plotted as negative values.

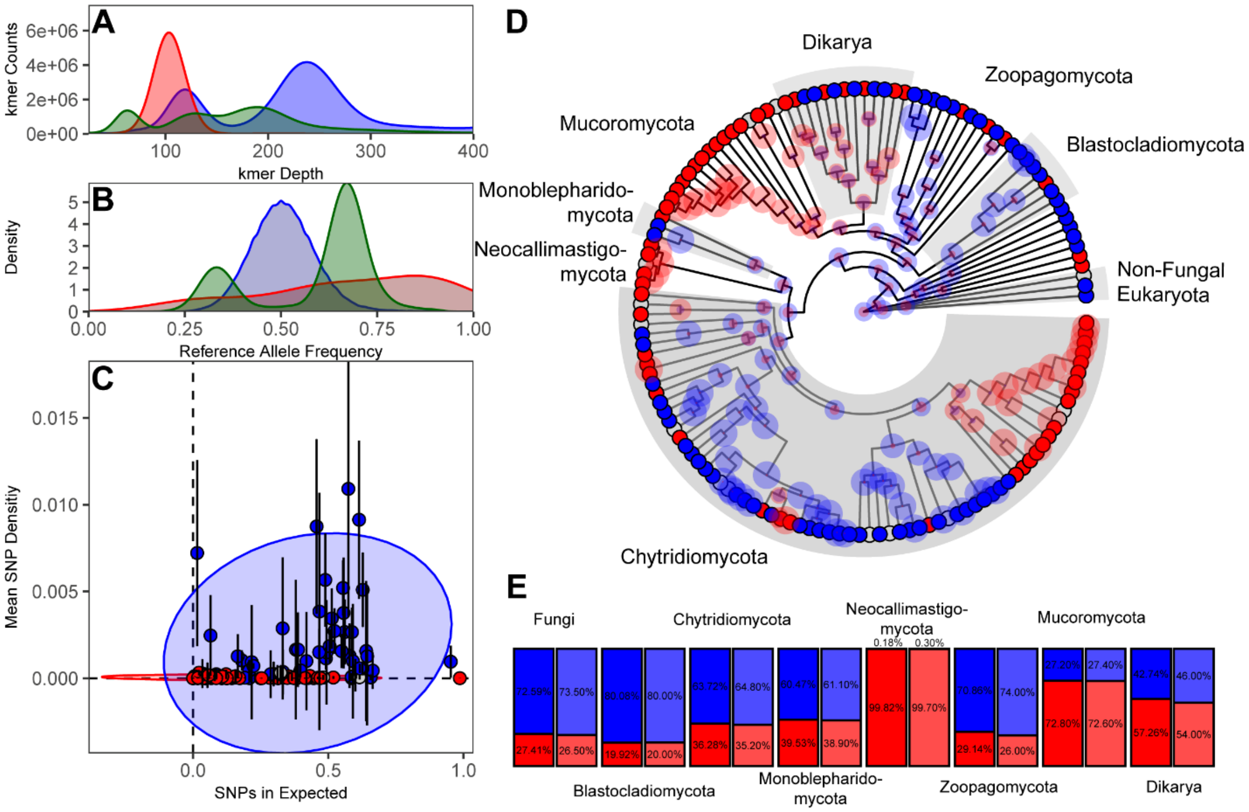


Figure 4.3. Summary of ploidy inference analyses ran for 112 assemblies and their underlying reads. Curves (A,B), points (C,D), or bars (E) are colored by ploidy state (triploid: green; diploid+: blue; haploid: red, uncertain: grey). (A) Overlaid select kmer histograms chosen from 112-plot set of kmer histograms based on their observed fit to canonical kmer distributions for triploid, diploid, and haploid genomes. (B) Overlaid select allele frequency density plots chosen from full 112-plot set of allele frequency histograms based on their observed fit to canonical allele frequency distributions for triploid, diploid, and haploid genomes. (C) Scatter plot of genomes by weighted mean of filtered SNP density across L50 contig set (y-axis) and proportion of filtered SNPs from L50 contig set falling within 1 standard deviation of the mean of each genome's theoretical binomial distribution (x-axis). Ellipses are normal ellipses around diploid+- or haploid-annotated points. Error bars represent standard deviation. Dashed lines indicate the origin. (D) Results of marginal ancestral state reconstruction of ploidy state across Fungi. Tips are colored according to the inferred ploidy of each genome they represent. Concentric circles on internal nodes represent inferred ancestral state probabilities of ploidy state at each node. Major clades are labeled with text and alternating grey-white insets. (E) Grouped stacked bar charts of ancestral state probabilities for major lineages inferred via marginal ancestral state reconstruction (solid) or stochastic ancestral state reconstruction (transparent).

Chapter 5: Single cell sequencing and phylogenomics places the enigmatic arthropod-mummifying fungus *Neozygites* as a distinct lineage in Entomophthorales.

5.1. Abstract

The fungal tree of life has and continues to undergo dramatic reorganizations as genome sequencing and genome-scale phylogenetics further resolves our view of the history of fungal evolution. *Neozygites* is a genus of obligate entomopathogenic fungi that parasitize mites, aphids, and other arthropods, catalyzing devastating epizootic outbreaks that are tightly tied to environmental conditions. *Neozygites* survives suboptimal environmental conditions inside the mummified cadavers of its hosts, lying in wait for conditions to improve and the cycle to begin anew. Based on its life history and characteristic morphology, the genus *Neozygites* was placed into the Entomophthorales, an order of other obligate entomopathogenic fungi. Early molecular phylogenetic analyses based on rDNA markers cast doubt on this placement by resolving *Neozygites* as a member of a separate lineage of fungi that do not share similar morphology nor parasitize insects. Based on long branches, this placement has always been considered artifactual but has never been resolved. In this study, we generate genome-scale data for three species of *Neozygites* and use genome-scale phylogenetics to resolve this conflict between morphological and molecular characters. We use multiple sets of phylogenetic markers, reduced taxon sampling, and separation of sites based on their relative evolutionary rates to assess the stability of the genus in genome-scale phylogenies. Our robustly supported phylogenies place *Neozygites* as an ancestral member of the Entomophthorales and we find this placement to be markedly stable despite long branches and alternative placements in past phylogenetic reconstructions.

5.2. Introduction

Entomopathogenic fungi are those that infect and kill insects and other arthropods, a type of parasitism that has evolved many times within the kingdom. Like most biotrophic strategies in fungi, entomopathogenic fungi can be completely dependent on parasitic nutrition (i.e., obligate) or not (i.e., facultative). As a group, they also show remarkable variability in host range, from compatibility with diverse hosts to compatibility with only a single host species (Fargues and Remaudiere, 1977; Wang et al., 2016). At the same time, some entomopathogenic lineages show the capacity for rampant host switching over evolutionary time scales (Tian et al., 2010; Wang et al., 2016). Obligate entomopathogenic fungi tend to have higher host-specificity, the combination of which makes their cultivation difficult or impossible, requiring highly specialized media formulations where it has been achieved (Delalibera et al., 2011; Grundschober et al., 1998). Unlike obligate fungal parasites and endosymbionts of plants, the genomes of entomopathogenic fungi do not show clear histories of reduction. Quite the opposite, genomes of obligate entomopathogenic fungi belonging to the early-diverging phylum Entomophthoromycotina can be >1 Gbp in length, representing some of the largest fungal genomes sequenced to date (Boyce et al., 2019; Elya et al., 2018).

The genus *Neozygites* is composed of 23 species of entomopathogenic fungi that infect and kill aphids or mites with a high degree of specificity (Delalibera and Hajek, 2004; Yaninek et al., 2002). Under the correct environmental conditions, species of *Neozygites* facilitate epizootic infections of host species, widespread but temporary mass killings of hosts in an area (Delalibera and Hajek, 2004; Steinkraus et al., 2002; Wekesa et al., 2007). Due to its host-specificity and pathogenicity, members of *Neozygites* have long been investigated as biocontrol agents across the world where its hosts facilitate crop declines (Delalibera et al., 2006; Delalibera and Hajek, 2004; Wekesa et al., 2007). The *Neozygites* infection cycle involves the infection of novel hosts through spore dispersal followed by vegetative growth within the host (Wekesa et al., 2007). Following the death of the host, cadavers go through a mummification process where they then remain, awaiting favorable environmental conditions (Delalibera and Hajek, 2004; Wekesa et al., 2007). When temperature and relative humidity rise to sufficient levels, the entomopathogen produces infective asexual propagules that are dispersed to nearby surfaces, catalyzing or amplifying epizootic infections (Delalibera and Hajek, 2004; Oduor et al., 1995; Wekesa et al., 2007).

Based on characteristic morphological traits, *Neozygites* was placed in the Entomophthorales, an order of fungal parasites of arthropods. This set of exceptionally characteristic traits, which rarely occur together in other fungal lineages, includes the capilloconidium, tiered conidial generations, and forcible primary conidial discharge (Keller, 1997). Despite all it shares with other members of the Entomophthorales, *Neozygites* is distinct from other Entomophthorales by virtue of its kinked capilloconidial stalk and unique nuclear behavior during mitosis. In fact, the novelty of its nuclear behavior, and associated distribution of actin microtubules, during mitosis was substantial enough that a new family in the Entomophthorales was erected specifically for *Neozygites* (Butt and Heath, 1988; Butt and Humber, 1989; Keller, 1997). This placement was called in to question by molecular phylogenies based exclusively on rDNA loci placed it in a separate phylum and sister to the mycoparasitic genus *Dimargaris* (Kickxellomycotina) (White et al., 2006). *Dimargaris* lacks all these morphological characteristics and, like other fungi in the Kickxellomycotina, does not parasitize arthropods. Based on its characteristic morphology and abnormal, poorly aligning rDNA sequences that result in long branches in phylogenies, the placement of *Neozygites* in the Kickxellomycotina has always been considered artifactual.

The Kingdom Fungi underwent significant reorganizations with the advent of next generation sequencing and the subsequent use of genome-scale sets of markers in phylogenetic reconstructions (Capella-Gutiérrez et al., 2012; James et al., 2013; Liu et al., 2009; Spatafora et al., 2016). The old early-diverging phylum Zygomycota was split into two phyla, each with three subphyla. Arthropod parasites in the old Entomophthorales were placed into the new subphylum Entomophthoromycotina (Spatafora et al., 2016). Clarification of the position of *Neozygites* in the context of this new classification has not been possible since it has only ever been represented by rDNA sequences (White et al., 2006). Its placement and the resolution of this disagreement between morphological and sequence-based traits remain unsolved.

In the initial phase of this study, we used single-cell genomics (SCG) to generate genome-scale data for *Neozygites* isolated from mummified aphid cadavers in Florida, USA (*Neozygites* sp. UF) toward resolving its placement. In our initial phylogenies we again recovered the *Neozygites–Dimargaris* relationship suggested by rRNA, however the genome assembly quality was poor. In order to produce better data to address this hypothesis, we then re-sequenced our initial isolate as well as two additional isolates of *Neozygites* via culture-based and single-cell whole genome sequencing. We use genome-scale sets of phylogenetic markers extracted from these draft genomes to resolve the placement of *Neozygites* in the fungal tree of life. Further, we investigate the stability of our placement and endeavor to explain the source of its phylogenetic affinity to *Dimargaris* and the Kickxellomycotina.

5.3. Methods

Strain information.

Genome sequencing was conducted on two cultured and one uncultured isolate of *Neozygites*. Cultured isolates were acquired from the United States Department of Agriculture's Agricultural Research Service Collection of Entomopathogenic Fungal Cultures (ARSEF). Two species of *Neozygites*, *Neozygites parvispora* (ARSEF 5620) and *Neozygites floridana* (ARSEF 5376), were acquired from ARSEF. *Neozygites* sp. UF was sequenced from mummified aphid cadavers collected during an epizootic outbreak in 2018. Although *Neozygites* sp. UF was only collected at this time, we isolated cells for sequencing in two rounds. In order to verify or rebuke our past results, we kept data from these two rounds separate in downstream analyses (Round 1: *Neozygites* sp. UF-NeoCoSC; Round 2: *Neozygites* sp. UF-Neo30) despite their shared source.

Culturing, DNA extraction, library preparation, and sequencing.

Upon reception of cultures from ARSEF, we set aside 2 μL aliquots of cell suspensions as a contingency for failure in establishing persistent cultures of these isolates. We attempted to culture both ARSEF isolates in Gibco Unsupplemented Grace's Insect Medium (Thermofisher CAT 11595030) supplemented with 1.65 μl/mL L-methionine and 5% Fetal Bovine Serum (Delalibera et al., 2011). Cultures only reached sufficient densities for conventional sequencing in the case of *N. floridana*. As such, DNA extraction from cultures of *N. floridana* was conducted following a standard CTAB extraction protocol. For *N. parvispora*, we conducted SCG using the QIAGEN REPLI-g Single Cell Kit (Cat. 150343) to generate enough template DNA for sequencing from 2 µL cell suspensions. Briefly, cells were lysed under alkaline conditions, the lysate was neutralized, and exposed genomic DNA was non-specifically amplified by multiple displacement amplification.

We did not attempt to culture *Neozygites* sp. UF collected from nature. Given the complex nature of this sample, we opted to follow the same SCG approach described above to generate whole genome data for this isolate (versus bulk metagenomic sequencing of whole aphids). Single aphids were placed in droplets of water under a stereoscope and agitated with a sterile probe. To collect 5–30 *Neozygites* cells in sterile PBS, we used either small capillary tubes to aspirate microliter amounts of water containing conidia or dental files to physically extract 10–30 fungal cells from droplets. In most cases, we simply deposited cells into 2 µL of PBS (*Neozygites sp.* UF-NeoCoSC). In one case, cells were washed in series of sterile water droplets, allowing us to isolate ~30 washed conidia (*Neozygites sp.* UF-Neo30). To mitigate against contamination in unwashed collections of cells, we prepared four for downstream sequencing (i.e., *Neozygites sp.* UF-NeoCoSC). Prior to sequencing, we verified the presence of *Neozygites* DNA in extracts via PCR amplification of the 18S rDNA locus using the primers SR6.1 (5'-TGTTACGACTTTTASTTCCTCT-3') and NS1.5 (5′-AAGGCAGCAGGCGCGCAAATTAC-3′) (James et al., 2000; Parrent and Vilgalys, 2009).

In all cases, sequencing libraries were generated from DNA extractions (either CTAB DNA extracts or SCG amplification products) using the Illumina Nextera Flex kit. We generated 1, 1, 4, and 1, sequencing libraries for *N. floridana, N. parvispora*, *Neozygites sp.* UF-NeoCoSC, and *Neozygites sp.* UF-Neo30, respectively. Sequencing libraries were sequenced on either the Illumina NextSeq (*Neozygites sp.* UF-NeoCoSC) or Illumina NovaSeq (*Neozygites sp.* UF-Neo30 and ARSEF isolates) sequencing platforms to generate paired-end 150 bp reads. Sequenced read libraries contained from 16,334,121–291,538,337 paired end reads. To account for the broad coverage distributions characteristic of SCG, single cell read libraries were normalized to their mean kmer coverage, as determined by *kmercountexact* in the *bbmap* software package (Bushnell, 2014).

Genome assembly, filtering, and annotation.

Read libraries were assembled with *SPAdes* v3.11.1 (Bankevich et al., 2012), using single-cell mode where applicable. The nonspecific amplification that characterizes SCG requires that SCG assemblies be assessed for contamination and, in cases where contamination is present, filtered prior to downstream analyses. We used SCGid v0.9b (Amses et al., 2020). to assess the degree of contamination and to filter initial assemblies. We detected sometimes significant human, fungal, or bacterial contamination in the assemblies associated with *Neozygites* sp. UF-NeoCoSC. Since contamination in these read libraries led to reduced sequencing depth of the *Neozygites* genome, we opted to filter the read libraries before reassembling all four filtered read libraries together (i.e., coassembly) to yield a single draft assembly for *Neozygites* sp. UF-NeoCoSC for use in downstream analyses. We did not detect contamination in the other three drafts. Genes were annotated in these four draft assemblies with *funannotate* v1.8.4 (Palmer and Stajich, 2019).

Phylogenomic tree inference.

To infer the placement of *Neozygites* in the fungal tree of life, we compiled a set of 68 representative fungi and 2 non-fungal eukaryotes for which protein annotations were publicly available (Table D1). Combined with our 4 *Neozygites* predicted proteomes, our phylogenomic dataset included 74 proteomes.

We inferred phylogenomic trees based on two different sets of genome-scale core orthologous genes (COGs): (i) BUSCO *fungi_odb10* (758 COGs) (Seppey et al., 2019) and (ii) *JGI_1086* (434 COGs) (Stajich, 2020). Protein sequences homologous to each marker were extracted from proteomes, aligned, and trimmed using a standard approach. Briefly, marker HMMs were downloaded and combined with *hmmpress* (Eddy and HMMER development team, 2015). Predicted proteins were searched against these multi-HMM marker files with *hmmsearch*. The resulting domain tables were filtered such that the predicted protein with the most significant hit (as determined by *hmmsearch*) to each marker HMM was selected for inclusion in marker gene alignments. For *fungi_odb10*, in cases where the most significant hit to the HMM from a proteome was below the BUSCO internal score cutoff for that marker, no protein sequence was selected, and gaps were inserted instead (Seppey et al., 2019). For *JGI_1086*, the best hit was selected so long as it was more significant than the e-value cutoff used for *hmmsearch* (i.e., $1e^{-5}$). The resulting FASTA files, which contained up to 74 protein sequences, were aligned with *Mafft* v7.310 (Katoh and Standley, 2013). To determine the most appropriate substitution model with which to infer individual gene trees, alignments were run through *ModelFinder* in *IQ-TREE* v2.0.5 (Kalyaanamoorthy et al., 2017; Minh et al., 2020). ML gene trees were inferred for each marker in *IQ-TREE* using the most appropriate model and 100 nonparametric bootstraps.

To infer concatenated ML trees, individual marker gene alignments were concatenated to yield a single multiple sequence alignment of 74 taxa, with gap sequences inserted for genes that were missing in individual alignments. Concatenated ML trees were computed in *IQ-TREE* using the most popular best substitution model among individual marker alignments (LG+I+G4 for both COG databases) with 10,000 ultrafast bootstrap replicates. Gene Concordance Factor support values (gCF) were calculated in *IQ-TREE*. ASTRAL consensus trees were inferred in *ASTRAL* v5.7.7 (Zhang et al., 2018) from gene trees. All phylogenetic trees presented in this study were visualized using *ggtree* (Yu et al., 2018, 2017) and *tidyverse* (Wickham et al., 2019) in R.

Segregation of fast and slow sites.

Relative evolutionary rates of sites in concatenated alignments were calculated using the Likelihood Estimation of Individual Site Rates (LEISR) functionality of *HyPhy* (Spielman and Kosakovsky Pond, 2018). Fast and slow sites were identified and binned from the outputs of these analyses according to different thresholds (e.g., 75% slowest sites and 25% fastest sites). Custom scripts were used to remove sites in fast and slow bins from alignments, yielding alignments of just fast or slow sites. Phylogenetic trees based on their segregate alignments were inferred in *IQ-TREE*, as described above, and visualized with *ggtree*. Tree-related plots were visualized with *ggplot2* in R (Wickham, 2016).

5.4. Results

Phylogenomic analyses robustly resolve the placement of Neozygites as an ancestral member of the Entomophthoromycotina.

We generated four draft genome assemblies for three isolates of *Neozygites*. Our draft genome assemblies ranged in size from 55.64–197.93 Mbp, encoding from 2,564–11,026 predicted proteins, and with completeness estimates ranging from 2.90%–33.38% (*BUSCO*, protein mode). To compile a gene set for phylogenomic analyses, we searched our annotated genomes against the 758 COG markers contained in *fungi_odb10*. Marker recovery rates across our entire 74-taxon dataset ranged from 9.20%–99.87% (mean = 82.44%). For *Neozygites* annotated genomes, marker recovery rates ranged from 9.2%–47.49% (mean = 26.85%). Across the 758 markers included in our phylogenomic reconstructions, marker occupancy rate varied from 32.43%–100.00% (mean = 60.51%).

Our ML phylogenomic reconstructions based on concatenation of 294,589 amino acid positions robustly resolves the placement the genus *Neozygites* as an ancestral member of the Entomophthoromycotina (Figure 5.1). Our phylogeny resolves this node with maximum bootstrap and high gene Concordance Factor (gCF = 42.9) support. gCF support values, which indicate the percentage of gene trees that support a clade, represent a more informative measure of support for large phylogenomic reconstructions, where maximal bootstrap values are easily attained in analyses based on large concatenated alignments (Salichos et al., 2014). Our phylogenomic analyses produced this well supported placement for the genus despite relatively low genome completeness estimates and marker recovery rates for *Neozygites* draft assemblies.

ASTRAL coalescent based on 758 individual marker gene trees supports placement of Neozygites as an ancestral member of Entomophthoromycotina.

To verify the ancestral position of *Neozygites* relative to the Entomophthoromycotina using a gene tree coalescent approach, we generated an ASTRAL consensus tree based on all 758 individual gene trees from the *fungi_odb10* marker set (Figure 5.2). The ASTRAL tree places *Neozygites* in the exact same position as the concatenated ML tree and does so with maximal local posterior probability support.

Placement of Neozygites as an ancestral member of the Entomophthoromycotina based on 758 COGs is significantly better than any alternative placement.

To assess the stability of our placement of the *Neozygites* clade in our concatenated ML tree, we conducted approximately unbiased tests (AU test) in *IQ-TREE* on a fully factorial set of trees that placed the clade at all internal nodes and tips across the tree (Shimodaira, 2002). All alternative placements were significantly worse ($\alpha = 0.05$) than the placement resolved in our concatenated ML tree (i.e., Figure 5.1). The top five alternative placements with pAU values closest to the significance threshold place *Neozygites* in other early-diverging fungal lineages (Figure 5.3). In order of decreasing pAU values, these top five trees place *Neozygites* ancestral to the Mortierellomycotina–Glomeromycotina clade (pAU = 0.00265), two equally significant placements in the Kickxellomycotina (pAU = 0.00222), ancestral to the Mucoromycota (pAU = 0.000963), or sister to *Martensiomyces pterosporus* in the Kickxellomycotina (pAU = 0.000739) (Figure 5.3). Interestingly, none of these top scoring, but nonsignificant, alternative placements resolve *Neozygites* as sister to *Dimargaris*, although 3/5 place *Neozygites* in the Kickxellomycotina.

Neozygites placement remains robustly supported in phylogenomic analyses despite the use of an alternative marker set.

To assess the stability of *Neozygites* placement dependent on marker set, we conducted a parallel phylogenomic construction of our 74-taxon dataset where we compiled a separate gene set based on 434 COGs contained in the *JGI_1086* marker set (used in our initial phylogenies of *Neozygites sp.* UF-NeoCoSC; data not shown). Marker recovery rates across our entire 74-taxon dataset ranged from 32.48%–100.0% (mean = 94.26%). For *Neozygites* annotated genomes, marker recovery rates ranged from 32.48%–82.26% (mean = 56.45%). This represents a marked increase in *Neozygites* representation compared to *fungi_odb10*, although it is important to note that there are no thresholds for score cutoffs like those provided for *fungi_odb10,* leaving only our *hmmsearch* e-value cutoff (i.e., 1e^{-5}) to exclude low-scoring sequences. Across the 434 markers, marker occupancy rates varied from 66.22%–100.00% (mean = 68.23%). We inferred the phylogeny of the resulting 121,262 amino acid concatenated alignment identically to our main phylogenomic reconstruction; that is, we used the most frequent best model among individual alignments (LG+I+G4) and 10,000 ultrafast bootstraps. Again, our independent phylogenomic reconstruction resolves a well-supported placement for the genus *Neozygites* as an ancestral member of the Entomophthoromycotina, at least in terms of bootstrap support (bootstrap = 100, gCF = 20.0) (Figure 5.4).

Neozygites placement is resilient to reductions in taxon sampling.

Since *Neozygites* branches off from the base of the Entomophthoromycotina on a long branch and is internally characterized by long branches between species, we wanted to test the stability of its placement in cases of reduced taxon sampling. To do so, we inferred four ML phylogenies for each marker set based on reduced size concatenated alignments that contained only one of each *Neozygites* isolate. For trees based on *fungi_odb10*, reduced taxon sampling had no impact on the placement of *Neozygites and* individual Neozygites samples were always resolved as ancestral to the Entomophthoromycotina (data not shown). On the other hand, trees based on *JGI_1086* were impacted by reduced taxon sampling for 2/4 species tips. When occurring alone in trees, *Neozygites* sp. UF-Neo30 and *Neozygites* sp. UF-*NeoCoSC* tips were resolved as sister to *Dimargaris* in the Kickxellomycotina. This is the same placement that characterized our previous genome-scale ML trees (data not shown) and past rDNA phylogenies (White et al., 2006). Placements of *N. floridana* and *N. parvispora* were completely resilient to reduced taxon sampling in *JGI_1086*-based trees. To further investigate the *Neozygites–Dimargaris* placement, we generated *JGI_1086* concatenated alignments that included fully factorial combinations of *Neozygites* tips and inferred ML trees. Paired inclusion of *Neozygites* sp. UF-NeoCoSC and *Neozygites* sp. UF-Neo30 resulted in the *Neozygites–Dimargaris* placement. Inclusion of a single resilient species tip (i.e., *N. floridana* or *N. parvispora*) with 1–2 unstable tips (i.e., *Neozygites sp.* UF-NeoCoSC or *Neozygites sp.* UF-Neo30) resulted in the expected placement ancestral to the Entomophthoromycotina.

Neozygites–Dimargaris placement is supported by sites with slow relative evolutionary rates.

To understand the underlying phylogenetic signal supporting the *Neozygites–Dimargaris* relationship in the *JGI_1086* marker set, we calculated the relative evolutionary rates of sites across an alignment including *Neozygites* sp. UF-NeoCoSC and *Neozygites sp.* UF-Neo30 as the sole representatives of *Neozygites.* Using these relative rates, we separated sites into slow and fast bins according to four different thresholds and generated alignments that contained only slow or fast sites (configurations: 0.25/0.75, 0.50/0.50, 0.75/0.25, 0.95/0.25 for slow and fast sites, respectively). We inferred ML trees based on these alignments and inspected them for the *Neozygites–Dimargaris* relationship. Surprisingly, all trees based on slow sites as well as the fast site tree including the highest proportion of slower sites (i.e., the 0.75 fast site tree) resolved *Neozygites* as sister to *Dimargaris* (Figure 5.5A). The 0.50 fast site tree resolved the expected Entomophthoromycotina placement (Figure 5.5B). The 0.25 and 0.05 fast site tree resolved different placements (ancestral to Cryptomycota and sister to *Saccharomyces*) (phylogenies not shown). These results suggests that slow-evolving sites, and not fast evolving sites, are providing support for the *Neozygites–Dimargaris* relationship in the *JGI_1086* alignments (Figure 5.5C,D). We repeated this slow-fast site separation process for full alignments (i.e., all four species tips) for both marker sets, where trees based on all slow or fast site bins resolved the expected placement of *Neozygites* as an ancestral member of the Entomophthoromycotina (data not shown).

5.5. Discussion

We resolve the placement of *Neozygites*, a genus of arthropod-mummifying entomopathogenic fungi, as an ancestral member of the Entomophthoromycotina using a genome-scale set of 758 markers (i.e., *fungi_odb10)*. Our placement is robustly supported by maximum likelihood, ASTRAL coalescent, bootstrapping, and gene Concordance Factor. Further, our placement is markedly stable, resilient to reductions in taxon sampling and inference based on subsets of slow- and fast-evolving sites. We independently verify our placement with a separate set of 434 genome-scale phylogenetic markers, although support is reduced according to some metrics. Our phylogenomic reconstructions should leave little doubt that *Neozygites* is an ancestral member of the Entomophthoromycotina in agreement with past placement of the genus as a distinct family based on morphological characteristics and its entomopathogenic lifestyle (Butt and Heath, 1988; Butt and Humber, 1989; Keller, 1997).

Although our placement of *Neozygites* is robustly supported, we demonstrate a few cases in which a different relationship is better supported; that is, the *Neozygites–Dimargaris* relationship that was suggested by past phylogenies (White et al., 2006). We show that this relationship is supported by sites with relatively slow evolutionary rates instead of those that are fast evolving. This is surprising given the presumption that support for this relationship in past trees was a case of long branch attraction driven by rapid rates of evolutionary change in the *Neozygites* genome (White et al., 2006).

Evolutionary rates aside, the fact that we only resolve the *Neozygites–Dimargaris* placement in trees based on one set of markers (i.e., *JGI_1086*) suggests that it is derived from signal unique to that marker set. We believe this signal to be derived from paralogous genes shared between the *Neozygites* and *Dimargaris* genomes, although we do not explicitly identify the sources of this signal. Unlike gene sets incorporated in our main ML tree, those incorporated into our *JGI_1086*-based ML trees were only filtered by a relatively liberal e-value cutoff (i.e., $1e^{-5}$). This makes it quite possible that paralogous sequences were included in those alignments, a possibility that is further suggested by reduced gCF support for the placement of *Neozygites* (i.e., 42.9 versus 20.0). Other potentially contributing factors are artifacts of mis-annotation in our *Neozygites* genome drafts due to the unsuitability of protein models used by protein annotation software applied to early diverging fungal lineages. The impact of the latter is hinted to by the low number of genes we annotated in our assemblies (i.e., 2,564–11,026 genes) relative to their cumulative size (i.e., 55.64–197.93 Mbp) as well as the low rates of marker recovery (i.e., 9.2%–47.49%) and estimates of genome completeness (i.e., 2.90%–33.38%). To contextualize this with other Entomophthoromycotina, the published genome of *Entomophthora muscae* is composed of 1.23 Gbp with 21,712 annotated genes (Elya et al., 2018). Unsurprisingly, the tips that garnered support for the *Neozygites–Dimargaris* relationship represent the least complete drafts (i.e., 2.90% and 3.56% estimated completeness) with the lowest number of annotated genes (i.e., 6,245 and 2,564 genes, respectively). Future work to address these inconsistencies for *Neozygites* and other early-diverging fungi in the Entomophthoromycotina, and beyond, will require critical evaluation of the methods and models used to annotate fungal genomes in this sector of the tree of life.

Our resolution of *Neozygites* in the fungal tree of life represents a somewhat uncommon situation where past classification based on morphology was called into question by molecular sequence data but later confirmed by more voluminous molecular data on the genome scale. Such examples are relatively rare in fungi, where evolutionary convergence on morphological traits in diverse lineages is rampant (e.g., Basidiomycete fruiting body form) (Binder et al., 2005; Hibbett, 2007; Hibbett et al., 1997). Along with past work, our placement of *Neozygites* asserts that there are morphological traits that contain important, unconflicted signal about the evolutionary histories of fungi. While this concept is not novel and there are other analogous morphological traits in fungi that have held up to the scrutiny of molecular phylogenetics (e.g., the basidium), modern phylogenetics is often quick to disregard morphological classifications when molecular classifications disagree. Conflicting or artifactual signal can exist in all accepted and potentially useful phylogenetic markers, including rDNA, derived from the process of evolution and exacerbated by long divergence times, rapid rates of sequence evolution, and deep paralogy (White et al., 2006). We present a situation where any reclassification of *Neozygites* based on rDNA phylogenies would have been proven wrong in the light of genome-scale data (White et al., 2006). Cases like this are important to remember as we continue to produce better and better reconstructions of the tree of life using more and more voluminous datasets.

5.6. Literature Cited

Amses, K.R., Davis, W.J., James, T.Y., 2020. SCGid, a consensus approach to contig filtering and genome prediction from single cell sequencing libraries of uncultured eukaryotes. Bioinformatics 36, 1194–2000.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477.

Binder, M., Hibbett, D.S., Larsson, K., Larsson, E., Langer, E., Langer, G., 2005. The phylogenetic distribution of resupinate forms across the major clades of mushroom-forming fungi (Homobasidiomycetes). System. Biodivers. 3, 113–157.

Boyce, G.R., Gluck-Thaler, E., Slot, J.C., Stajich, J.E., Davis, W.J., James, T.Y., Cooley, J.R., Panaccione, D.G., Eilenberg, J., De Fine Licht, H.H., Macias, A.M., Berger, M.C., Wickert, K.L., Stauder, C.M., Spahr, E.J., Maust, M.D., Metheny, A.M., Simon, C., Kritsky, G., Hodge, K.T., Humber, R.A., Gullion, T., Short, D.P.G., Kijimoto, T., Mozgai, D., Arguedas, N., Kasson, M.T., 2019. Psychoactive plant- and mushroom-associated alkaloids from two behavior modifying cicada pathogens. Fungal Ecol. 41, 147–164.

Bushnell, B., 2014. sourceforge.net/projects/bbmap: BBMap.

Butt, T.M., Heath, I.B., 1988. The changing distribution of actin and nuclear behavior during the cell cycle of the mite-pathogenic fungus Neozygites sp. Eur. J. Cell Biol. 46, 499–505.

Butt, T.M., Humber, R.A., 1989. An immunofluorescence study of mitosis in a mite-pathogen,Neozygites sp. (Zygomycetes: Entomophthorales). Protoplasma 151, 115–123.

Capella-Gutiérrez, S., Marcet-Houben, M., Gabaldón, T., 2012. Phylogenomics supports microsporidia as the earliest diverging clade of sequenced fungi. BMC Biol. 10, 47.

Delalibera, I., Jr., Demétrio, C.G.B., Manly, B.F.J., Hajek, A.E., 2006. Effect of relative humidity and origin of isolates of Neozygites tanajoae (Zygomycetes : Entomophthorales) on production of conidia from cassava green mite , Mononychellus tanajoa (Acari : Tetranychidae), cadavers. Biol. Control 39, 489–496.

Delalibera, I., Jr, Hajek, A.E., 2004. Pathogenicity and specificity of Neozygites tanajoae and Neozygites floridana (Zygomycetes: Entomophthorales) isolates pathogenic to the cassava green mite. Biol. Control 30, 608–616.

Delalibera, I., Jr, Hajek, A.E., Humber, R.A., 2011. Use of cell culture media for cultivation of the mite pathogenic fungi Neozygites tanajoae and Neozygites floridana 84, 119–127.

Eddy, S.R., HMMER development team, 2015. HMMER.

Elya, C., Lok, T.C., Spencer, Q.E., McCausland, H., Martinez, C.C., Eisen, M.B., 2018. Robust manipulation of the behavior of Drosophila melanogaster by a fungal pathogen in the laboratory. bioRxiv. https://doi.org/10.1101/232140

Fargues, J., Remaudiere, G., 1977. CONSIDERATIONS ON THE SPECIFICITY OF ENTOMOPATHOGENIC FUNGI. MYCOPATHOLOGIA 62, 31–37.

Grundschober, A., Tuor, U.R.S., Aebi, M., 1998. APPLIED MICROBIOLOGY In vitro Cultivation and Sporulation of Neozygites parvispora (Zygomycetes : Entomophthorales). Syst. Appl. Microbiol. 21, 461–469.

Hibbett, D.S., 2007. After the gold rush, or before the flood? Evolutionary morphology of mushroom-forming fungi (Agaricomycetes) in the early 21st century. Mycol. Res. 111, 1001–1018.

Hibbett, D.S., Pine, E.M., Langer, E., Langer, G., Donoghue, M.J., 1997. Evolution of gilled mushrooms and puffballs inferred from ribosomal DNA sequences. Proc. Natl. Acad. Sci. U. S. A. 94, 12002–12006.

James, T.Y., Pelin, A., Bonen, L., Ahrendt, S., Sain, D., Corradi, N., Stajich, J.E., 2013. Shared signatures of parasitism and phylogenomics unite cryptomycota and microsporidia. Curr. Biol. 23, 1548–1553.

James, T.Y., Porter, D., Leander, C.A., Vilgalys, R., Longcore, J.E., 2000. Molecular phylogenetics of the Chytridiomycota supports the utility of ultrastructural data in chytrid systematics. Can. J. Bot. 78, 336–350.

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

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.

Keller, S., 1997. The genus Neozygites (Zygomycetes , Entomophthorales) with special reference to species found in tropical regions. Sydowia 49, 118–146.

Liu, Y., Steenkamp, E.T., Brinkmann, H., Forget, L., Philippe, H., Lang, B.F., 2009. Phylogenomic analyses predict sistergroup relationship of nucleariids and fungi and paraphyly of zygomycetes with significant support. BMC Evol. Biol. 9, 272.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., Lanfear, R., 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. 37, 1530–1534.

Oduor, G.I., Moraes, G.J.D., Yaninek, J.S., Geest, L.P.S.V.D., 1995. Effect of temperature , humidity and photoperiod on mortality of Mononycheiius tanajoa (Acari : Tetranychidae) infected by Neozygites cf . floridana (Zygomycetes : Exp. Appl. Acarol. 19, 571–579.

Palmer, J., Stajich, J., 2019. nextgenusfs/funannotate: funannotate v1.5.3. https://doi.org/10.5281/zenodo.2604804

Parrent, J.L., Vilgalys, R., 2009. Expression of genes involved in symbiotic carbon and nitrogen transport in Pinus taeda mycorrhizal roots exposed to CO2 enrichment and nitrogen fertilization. Mycorrhiza 19, 469–479.

Salichos, L., Stamatakis, A., Rokas, A., 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol. Biol. Evol. 31, 1261–1271.

Seppey, M., Manni, M., Zdobnov, E.M., 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods Mol. Biol. 1962, 227–245.

Shimodaira, H., 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51, 492–508.

Spatafora, J.W., Chang, Y., Benny, G.L., Lazarus, K., Smith, M.E., Berbee, M.L., Bonito, G., Corradi, N., Grigoriev, I., Gryganskyi, A., James, T.Y., O'Donnell, K., Roberson, R.W., Taylor, T.N., Uehling, J., Vilgalys, R., White, M.M., Stajich, J.E., 2016. A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data. Mycologia 108, 1028–1046.

Spielman, S.J., Kosakovsky Pond, S.L., 2018. Relative evolutionary rate inference in HyPhy with LEISR. PeerJ 6, e4339.

Stajich, J., 2020. 1KFG/PHYling_HMMs_fungi: PHYling markers 1.3. https://doi.org/10.5281/zenodo.3630031

Steinkraus, D.C., Boys, G.O., Rosenheim, J.A., 2002. Classical biological control of Aphis gossypii (Homoptera : Aphididae) with Neozygites fresenii (Entomophthorales : Neozygitaceae) in California cotton. Biol. Control 25, 297–304.

Tian, L.-H., Hu, B., Zhou, H., Zhang, W.-M., Qu, L.-H., Chen, Y.-Q., 2010. Molecular phylogeny of the entomopathogenic fungi of the genus Cordyceps (Ascomycota: Clavicipitaceae) and its evolutionary implications. J. Syst. Evol. 48, 435–444.

Wang, J.B., St. Leger, R.J., Wang, C., 2016. Chapter Three - Advances in Genomics of Entomopathogenic Fungi, in: Lovett, B., St. Leger, Raymond J. (Eds.), Advances in Genetics. Academic Press, pp. 67–105.

Wekesa, V.W., Moraes, G.J., Knapp, M., Jr, I.D., 2007. Interactions of two natural enemies of Tetranychus evansi , the fungal pathogen Neozygites floridana (Zygomycetes : Entomophthorales) and the predatory mite , Phytoseiulus longipes (Acari : Phytoseiidae). Biol. Control 41, 408–414.

White, M.M., James, T.Y., Donnell, K.O., Cafaro, M.J., Tanabe, Y., Sugiyama, J., James, T.Y., Carolina, N., Donnell, K.O., 2006. Phylogeny of the Zygomycota based on nuclear ribosomal sequence data. Mycologia 98, 872–884.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the tidyverse. J. Open Source Softw. 4, 1686.

Yaninek, J.S., Moraes, G.J.D.E., Oduor, G.I., 2002. Host specificity of the cassava green mite pathogen Neozygites floridana 61–66.

Yu, G., Lam, T.T.-Y., Zhu, H., Guan, Y., 2018. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. Mol. Biol. Evol. 35, 3041–3043.

Yu, G., Smith, D.K., Zhu, H., Guan, Y., Lam, T.T., 2017. Ggtree : An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. 8, 28–36.

Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S., 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19, 153.

5.7 Figures

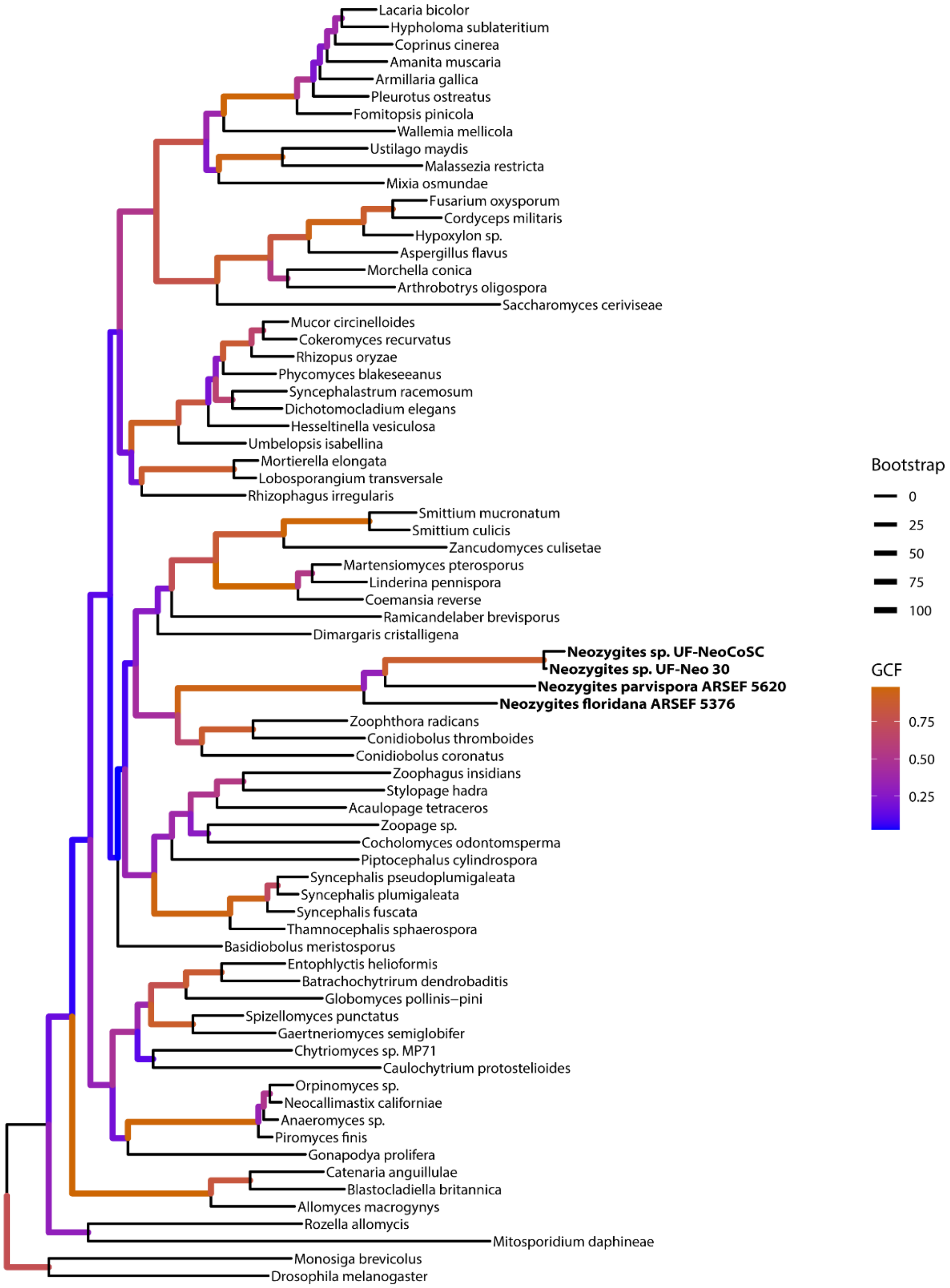


Figure 5.1. Maximum likelihood phylogenomic tree based on a concatenated alignment of 294,589 amino acid positions extracted from 74 fungal genomes based on the BUSCO *fungi_odb10* COG database. Phylogeny resolves *Neozygites* in an ancestral position within the Entomophthoromycotina, which is otherwise represented by *Zoophthora radicans*, *Conidiobolus thromboides*, and *Conidiobolus coronatus*. Bootstrap support values are indicated by the thickness of node edges. Node edges are colored by gCF support values. *Neozygites* tips are bolded.

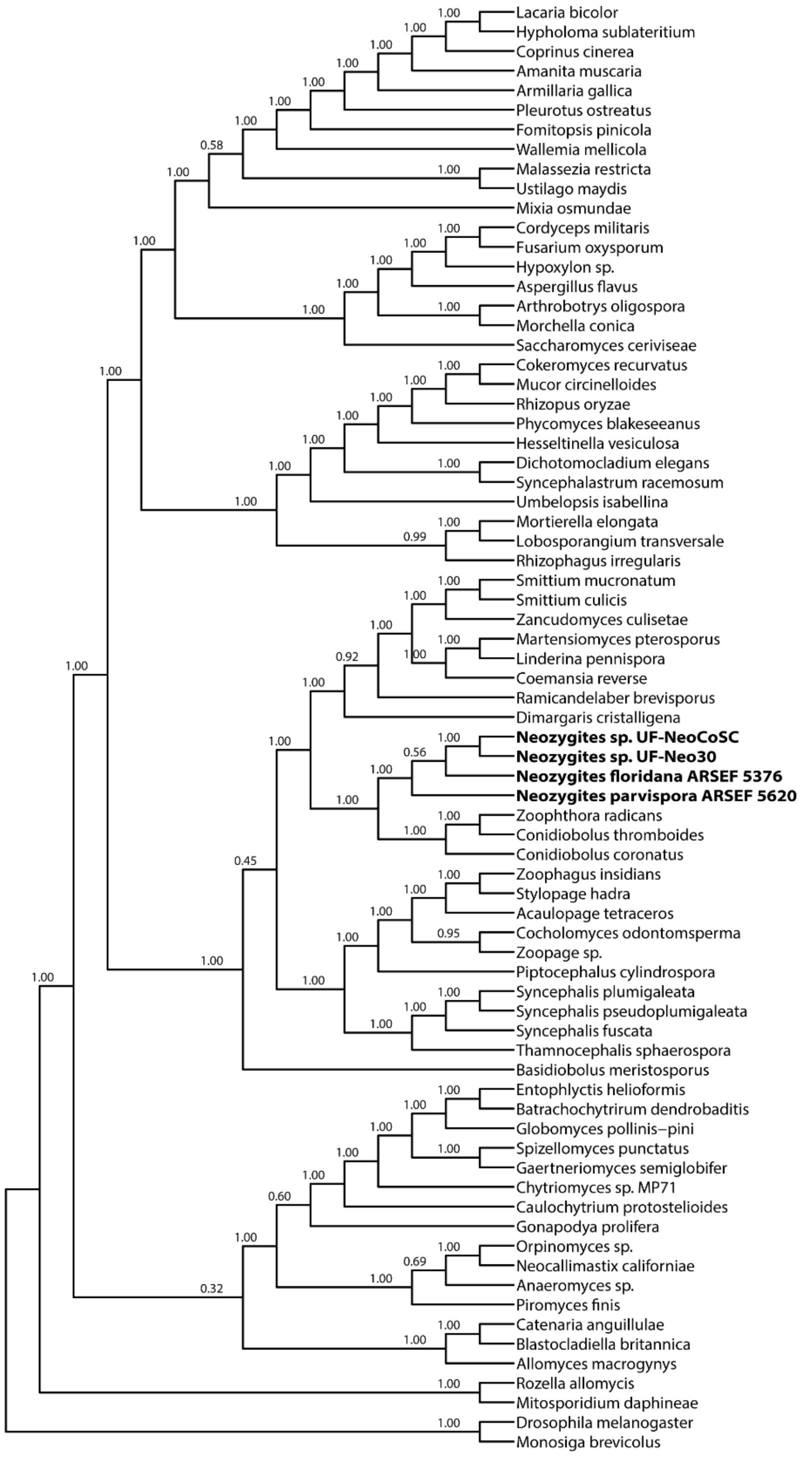


Figure 5.2. ASTRAL coalescent tree based on 758 gene trees computed for phylogenetic markers in the BUSCO *fungi_odb10* COG database. Phylogeny resolves *Neozygites* in an ancestral position within the Entomophthoromycotina, which is otherwise represented by *Zoophthora radicans*, *Conidiobolus thromboides*, and *Conidiobolus coronatus*. ASTRAL local posterior probabilities are indicated by text on nodes. *Neozygites* tips are bolded.

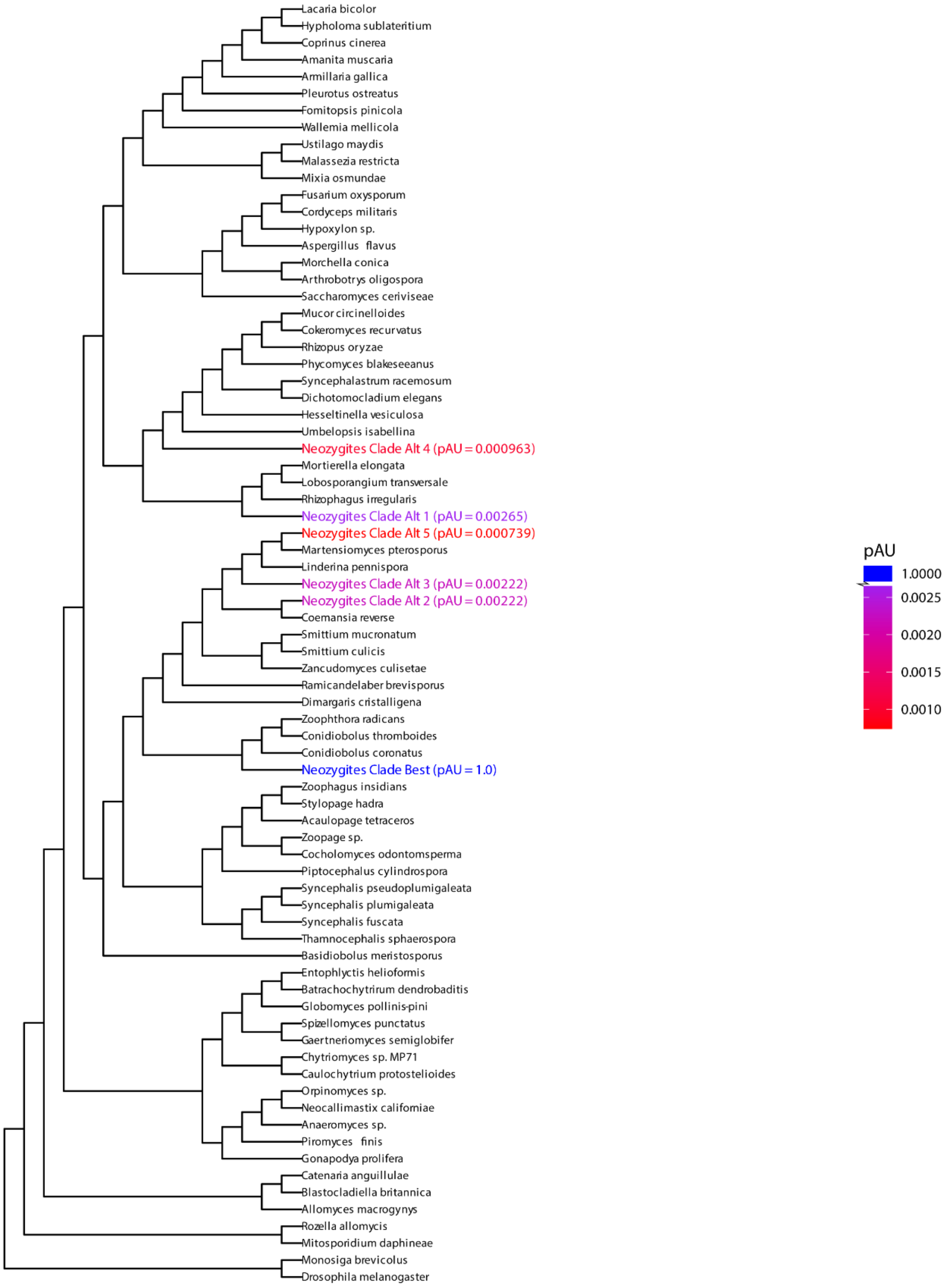


Figure 5.3. Cladogram representation of ML tree from Figure 5.1 annotated with the results of approximately unbiased test (AU test). Best-supported *Neozygites* clade has been collapsed to a single tip (blue tip). Top five highest-score, but nonsignificant, placements of *Neozygites*, according to the AU test, appear as tips colored by the magnitude of their pAU value (red-to-purple tips). Actual pAU values parenthesized in tip labels.

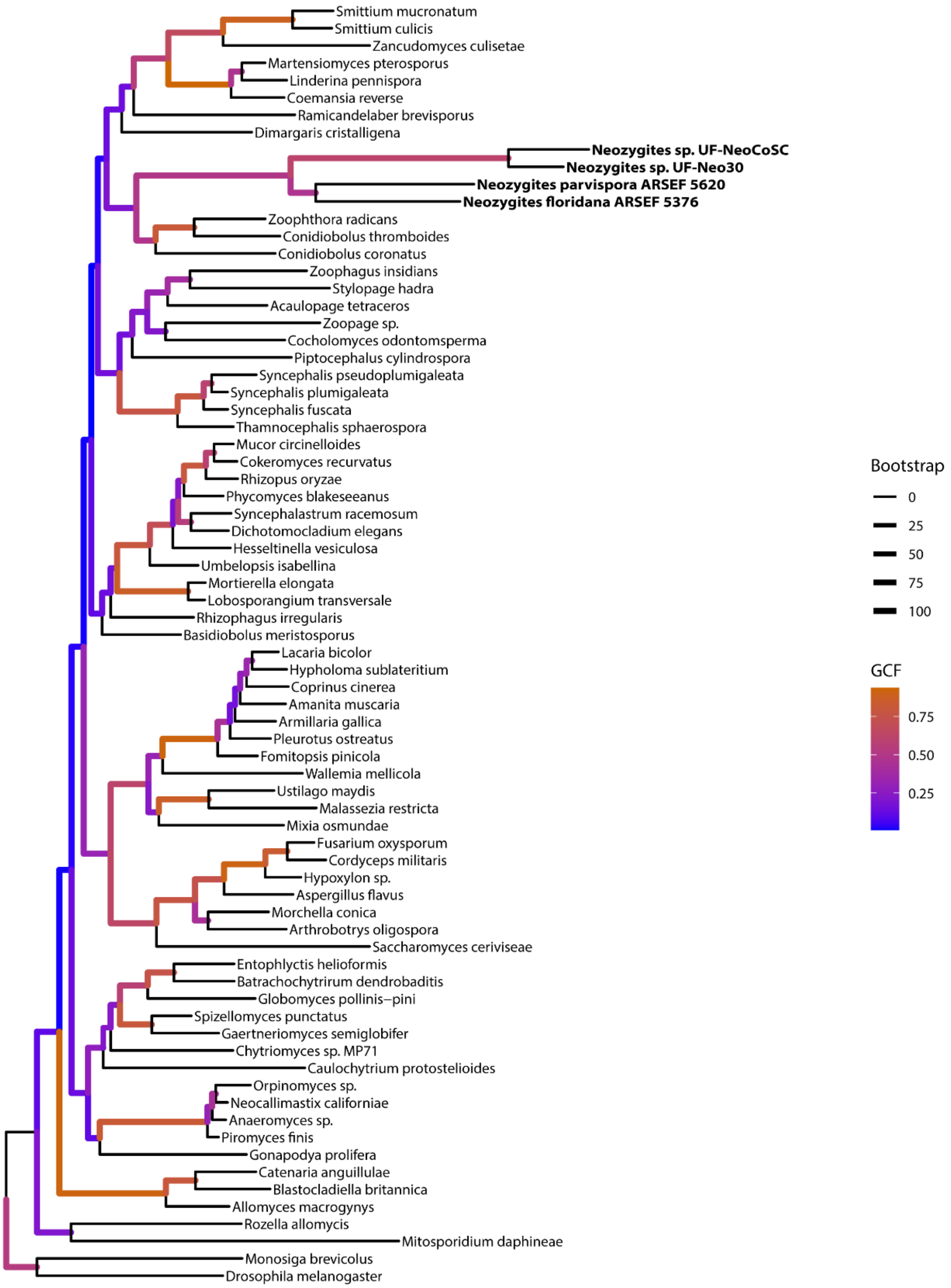


Figure 5.4. Maximum likelihood phylogenomic tree based on a concatenated alignment of 121,262 amino acid positions extracted from 74 fungal genomes based on the *JGI_1086* COG database. Phylogeny resolves *Neozygites* in an ancestral position within the Entomophthoromycotina, which is otherwise represented by *Zoophthora radicans*, *Conidiobolus thromboides*, and *Conidiobolus coronatus*. Bootstrap support values are indicated by the thickness of node edges. Node edges are colored by gCF support values. *Neozygites* tips are bolded.

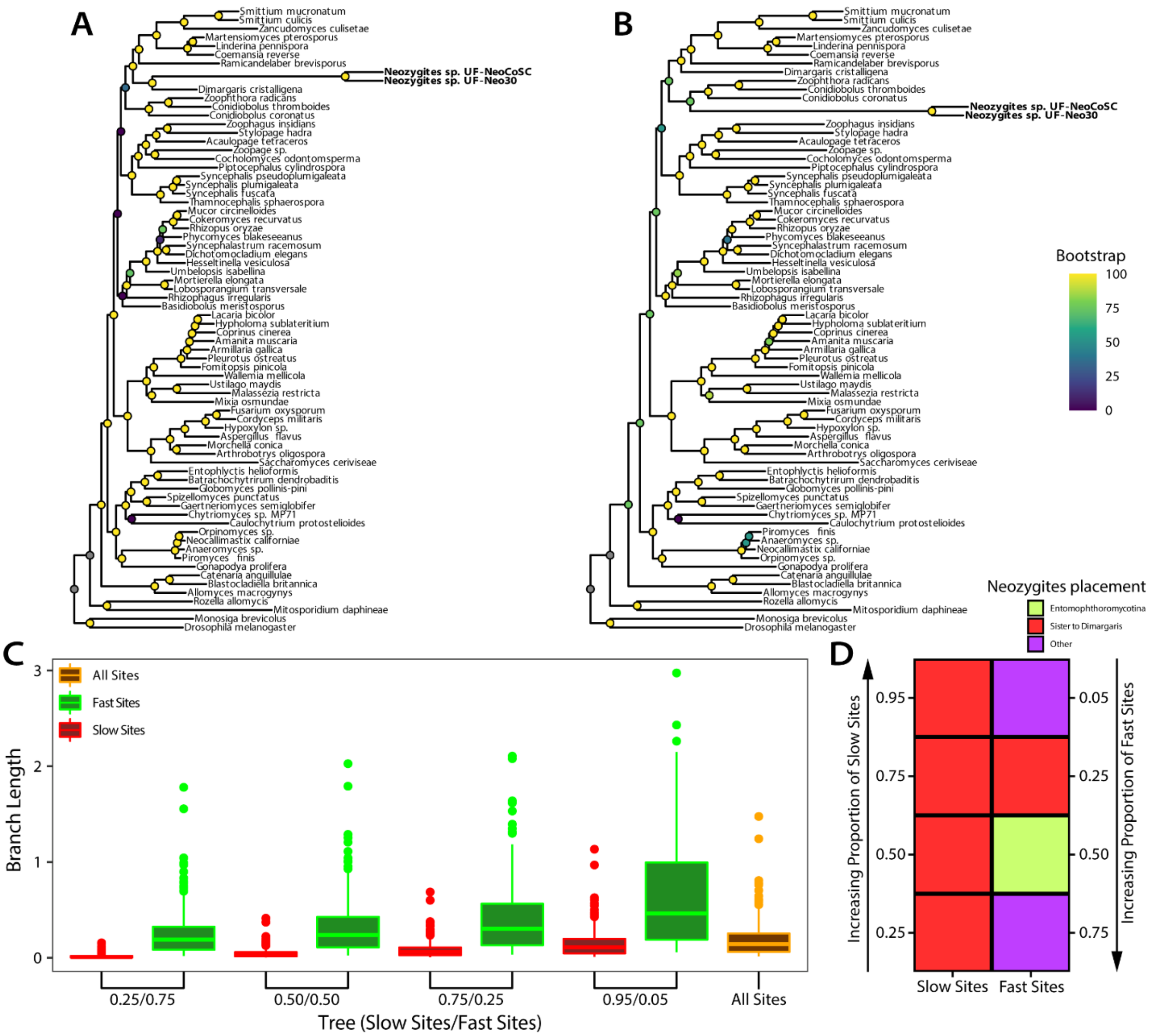


Figure 5.5. Summary of phylogenetic analyses based on bins of fast- and slow-evolving sites segregated from alignments containing only *Neozygites* sp. UF-NeoCoSC and *Neozygites* sp. UF-Neo30 as representatives of *Neozygites*. (A) 0.95 slow site tree based on the bin with the highest proportion of slow sites. Bootstrap support values indicated by colored points on nodes. Phylogeny is 1/5 that resolved the *Neozygites–Dimargaris* relationship. (B) 0.50 fast site tree, which is the only tree that resolved the expected placement of *Neozygites* as an ancestral member of the Entomophthoromycotina. Bootstrap support values indicated by colored points on nodes. (C) Boxplots of branch lengths from pairs of slow (red) and fast (green) site trees from each site-segregation configuration (e.g., 0.25/0.75 slow/fast bins). Branch lengths of the full ML tree (i.e., Figure 5.4) represented by yellow boxplot. (D) Tiled plot showing the placement of the 2-tip *Neozygites* clade resolved by each fast or slow site tree in each of the four site-segregation configurations (ancestral member of Entomophthoromycotina, green; sister to *Dimargaris*, red; some other placement, purple). Y-axes are opposite relative to each other as indicated by arrows.

Chapter 6: Conclusion

6.1 Summary of Dissertation Research

Due to their small stature, microbes exist on a plane outside of human perception in nature. As such, technological and methodological advances to bring microbes onto human scales have been needed to drive microbial research forward (Bonnet et al., 2020; Pace, 1997; Porter, 1976; Tucker et al., 2009; van Niel, 1944). Despite a rich history of innovation-driven microbial research, the vast majority of microbes remain poorly understood. A major factor contributing to and perpetuating this gap in our understanding of microbial diversity on Earth is that most microbes cannot be grown under axenic conditions in the lab, which precludes identification, experimentation, and genome sequencing. Like technological and methodological advances before it, SCG offers a promising way forward to understanding the Earth's vast pool of uncultured microbes (Kalisky and Quake, 2011). SCG is not without bias, but circumvents some of the problems with using other culture-independent sequencing approaches to investigate this untapped pool of biodiversity (Schoch et al., 2012; Yue et al., 2020).

In this dissertation, I use targeted SCG and conventional genome sequencing to conduct genome-enabled biological research in uncultured and under-sampled fungi. I first developed a novel computational approach for filtering SCG metagenomes that expedites the path to draft genomes of uncultured microbes, a process for which there were no automated tools and that was previously dependent on time consuming manual curation (Gawryluk et al., 2016; Mikhailov et al., 2016; Sedlar et al., 2017). In subsequent chapters, I use SCG enabled by this approach to discover and characterize novel endohyphal bacteria (EHB) colonizing uncultured predatory fungi, include uncultured fungi in a large phylogenomic analysis of under-sampled zoosporic fungi that, most importantly, upends the prevailing conception of fungal life cycles as haploid-centric, and resolve the placement of an enigmatic arthropod-mummifying fungal parasite whose placement was conflicted (Keller, 1997; White et al., 2006).

Chapter 2: SCGid, a consensus approach to contig filtering and genome prediction from single-cell sequencing libraries of uncultured eukaryotes.

In this chapter, I develop and test *SCGid*, a metagenome filtering tool designed specifically to address sequence biases introduced by the SCG-associated MDA reaction (Pinard et al., 2006). *SCGid* reimplements existing filtering approaches with SCG-optimizations and uses consensus-based reasoning to generate high fidelity draft genomes for uncultured microbes. I test *SCGid* on mock and genuine SCG datasets where I demonstrate that it can recapitulate the results of time-consuming manual curation in a fast and automated way. My hope is that *SCGid* increases access to metagenome filtering and genome-enabled research for experts in uncultured microbes from across the tree of life. By virtue of its consensus-based approach, *SCGid* has the potential to grow as new methods for metagenome filtering come to light.

Chapter 3: Novel obligate endohyphal bacterial symbionts of uncultured predatory fungi revealed by single cell sequencing implicate recent interphylum host switches.

In this chapter, I use *SCGid* to isolate nearly complete genomes of two novel EHB in association with uncultured predatory fungi in the Zoopagomycota *in silico*. My genome-scale phylogenetic analyses resolve these novel EHB to be nested within the MRE and BRE groups of obligate EHB known predominantly from plant-associated fungi in the Mucoromycota, implicating interphylum host switches in the histories of obligate EHB. In line with past work on MRE, I detected the signature of past horizontal gene transfer from fungi in the genome of the novel MRE-related EHB I discovered (Naito et al., 2015; Sun et al., 2019; Torres-Cortés et al., 2015). Interestingly, I also detected the signature of horizontal gene transfer from non-fungal eukaryotes, which has interesting implications for the flow of genes during fungal predation of animals and protozoans. The detection of these lineages of EHB in fungi that are not associated with plants has major implications for the burgeoning field of EHB biology, which leans heavily on a framing of these EHB as endosymbionts of plant-associated fungi.

Chapter 4: Phylogenomic analysis of zoosporic true fungi suggests most early diverging lineages have diploid-dominant life cycles.

In this chapter, in collaboration with an interinstitutional team of researchers, I use genome sequencing of 65 new fungal genomes and phylogenomics to infer a robustly supported phylogeny of early diverging zoosporic fungi based on 197,423 amino acid positions. Aware of the bias in available sets of genome-scale phylogenetic markers against the early-diverging fungal lineages at the center of these analyses, I develop and employ an automated gene tree filtering approach that, when combined with subsequent manual filtering, excludes the conflicting signal of paralogous and spurious sequences. I also implement a computational pipeline that infers the ploidy of draft genome assemblies from an entirely sequence-based perspective. I deploy this approach to demonstrate that diploid-centric life cycles, versus the haploid-centric ones that characterize Dikarya, have been much more important in fungal evolution than previously appreciated. The results of mapping ploidy and other genetic characters onto the phylogeny paint a history of fungal evolution that has seen the gradual loss of traits that connect them to their MRCA with animals and a transition toward those that characterize Dikarya. Taken together, these two results have major implications for character evolution in fungi, which is too often seen through a Dikarya-centric lens.

Chapter 5: Single cell sequencing and phylogenomics places the enigmatic arthropod-mummifying fungus Neozygites as a distinct lineage in Entomophthorales.

In this final chapter, I generate four draft genomes for three species of *Neozygites*, a genus of obligate parasites of arthropods, and place it in the fungal tree of life using genome-scale phylogenetic analyses. The placement of *Neozygites* based on morphology and life style was called into question when rDNA-based phylogenetic analyses placed it in a phylum of non-entomopathogenic fungi with which it shared little diagnostic morphological similarities (White et al., 2006). I resolved this conflict by resolving *Neozygites* as an ancestral member of the Entomophthorales based on 294,589 amino acid positions with robust support. This is in agreement with past classification of the genus based on morphology and in disagreement with rDNA marker loci (Keller, 1997; White et al., 2006). I assess the stability of the *Neozygites* placement by inferring a phylogenies with different markers set, artificially reducing *Neozygites* taxon sampling, and splitting-up sites with fast and slow relative evolutionary rates. In general, I find the placement of *Neozygites* to be markedly stable. A small subset of phylogenies based on the alternative marker set, which I filter less stringently, and with reduced taxon sampling recapitulate the artifactual *Neozygites–Dimargaris*. Inference of trees based on slow or fast sites suggests that slow-evolving paralogs, erroneously included in alignments, garner support for this artifactual placement. This is at odds with expectations, which implicated fast evolving orthologous sites as the cause of artifactual placement (White et al., 2006).. Future work to identify the source of this signal will solve the phylogenetic mystery of the arthropod-mummifying fungus for good.

6.2 Synthesis and Future Directions

The work of reconstructing a fungal tree of life that accurately represents actual fungal diversity is just beginning. At present, our best drafts of evolutionary diversification in the kingdom is substantially biased toward later-diverging lineages that produce macroscopic forms or can be cultivated under axenic conditions (James et al., 2020). Early diverging lineages where microscopic and uncultured fungi dominate are poorly represented both in numbers of described species and representative biological sequences. Based on the current state of fungal sampling and estimates that put total fungal diversity between 2–4 million species, it is obvious that of all the gaps in our understanding of Kingdom Fungi, gaps in early diverging lineages are the most substantial (Hawksworth and Lücking, 2017; James et al., 2020). These lineages are prime targets for discoveries that further resolve this picture.

My dissertation work demonstrates the utility of SCG in facilitating genome-enabled biological research in uncultured fungi. So long as cells can be collected, SCG massively alleviates the activation energy required to generate genome-scale data. While the more classical approach of quasi-cultivation and manual collection of cells employed herein preserves the organismal identity of uncultured fungi, modern automated cell sorting approaches are significantly higher throughput (Davis et al., 2019; Drechsler, 1959; Rinke et al., 2014). Whether cells are collected by either method is unimportant to the central goal of resolving a more complete fungal tree of life. Similar to other culture-independent sequencing approaches, SCG is not without bias (Pinard et al., 2006). The MDA reaction it involves introduces unique sequence composition biases and exacerbates contamination (Davis et al., 2019; Mikhailov et al., 2016; Pinard et al., 2006). My dissertation endeavors to circumvent the barriers introduced by using SCG in uncultured microbes by introducing an automated SCG metagenome filtering pipeline that is designed with them in mind. The overarching goal of *SCGid* is to make SCG an even more viable option for sequencing the genomes of uncultured fungi, and other microbes, than it is now, thereby encouraging its use to fill gaps in the fungal tree of life.

Sequencing the genomes of members of under-sampled fungal lineages is only the first step in further resolving the fungal tree of life. While sampling certainly increases representation in phylogenetic reconstructions, it does not guarantee their accuracy. Genome-scale data has given molecular phylogenetics a vast wealth of characters with which to infer phylogenetic hypotheses, but it is important to remember that not all data is good data from a phylogenetic standpoint (Choi and Kim, 2017; James et al., 2020; Prasanna et al., 2019). My dissertation underpins the importance of assessing marker compatibility and filtering paralogous and spurious sequence data out of phylogenetic analyses early, before it leads to misrepresentations of evolutionary history. While these considerations are important in phylogenetic reconstructions across the tree of life, they are most important in poorly sampled lineages (Prasanna et al., 2019).

Beyond demonstrating the utility of SCG in resolving the fungal tree of life and inviting caution about genome-scale phylogenetics, my dissertation shows how SCG can be used to discover entirely novel symbioses between separate domains of life. Prior to my dissertation research, obligate EHB of fungi were known predominantly from plant-associated fungi in the Mucoromycota (Araldi-brondolo et al., 2017; Pawlowska et al., 2018). This relatively narrow host range has led to the framing of these EHB as involved members in well-understood plant-fungal interactions (Araldi-brondolo et al., 2017; Partida-Martinez and Hertweck, 2005; Pawlowska et al., 2018). While my dissertation does not contest the involvement of EHB in plant-fungal interactions, it does contest their framing as endosymbionts solely of plant-associated fungi that are only important to plant-fungal interactions. My discovery of EHB in zoopagomycotan fungal predators of nematodes requires the concept of EHB to be broadened significantly. Further, the nested placement of these novel EHB within clades of Mucoromycota-associated EHB implicates host switching on scales that are not currently appreciated in the study of EHB (Araldi-brondolo et al., 2017; Mondo et al., 2012; Toomer et al., 2015). If obligate EHB from these bacterial lineages colonize these zoopagomycotan fungi, they probably colonize others in Zoopagomycota and other early-diverging lineages. Based on the paucity of sampling in early-diverging fungal lineages in general, future work to detect and characterize novel EHB therein will almost undoubtedly succeed.

To understand the breadth of microbial diversity on Earth, we need to understand all of its parts, and it is abundantly clear that we understand some parts much better than others, in Kingdom Fungi and elsewhere. So, toward a better understanding of the awe-inspiring diversity of microbial life, we need to do a better job of sampling its under-represented lineages. Many of these lineages are uncultured and, as such, SCG represents a powerful methodological advance with which to, in line with its innovation-driven history, continue to drive microbial research forward.

6.3 Literature Cited

Araldi-brondolo, S.J., Spraker, J., Shaffer, J.P., Woytenko, E.H., Baltrus, D.A., Gallery, R.E., Arnold, A.E., 2017. Bacterial Endosymbionts: Master Modulators of Fungal Phenotypes. The Fungal Kingdom 981–1004.

Bonnet, M., Lagier, J.C., Raoult, D., Khelaifia, S., 2020. Bacterial culture through selective and non-selective conditions: the evolution of culture media in clinical microbiology. New Microbes New Infect 34, 100622.

Choi, J., Kim, S.-H., 2017. A genome Tree of Life for the Fungi kingdom. Proc. Natl. Acad. Sci. U. S. A. 114, 9391–9396.

Davis, W.J., Amses, K.R., Benny, G.L., Carter-house, D., Chang, Y., Grigoriev, I., Smith, M.E., Spatafora, J.W., Stajich, J.E., James, T.Y., 2019. Genome-scale phylogenetics reveals a monophyletic Zoopagales (Zoopagomycota , Fungi). Mol. Phylogenet. Evol. 133, 152–163.

Drechsler, C., 1959. Several Zoopagaceae Subsisting on a Nematode and on Some Terricolous Amoebae. Mycologia 51, 787–823.

Gawryluk, R.M.R., del Campo, J., Okamoto, N., Strassert, J.F.H., Lukeš, J., Richards, T.A., Worden, A.Z., Santoro, A.E., Keeling, P.J., 2016. Morphological Identification and Single-Cell Genomics of Marine Diplonemids. Curr. Biol. 26, 3053–3059.

Hawksworth, D.L., Lücking, R., 2017. Fungal diversity revisited: 2.2 to 3.8 million species, in: The Fungal Kingdom. ASM Press, Washington, DC, USA, pp. 79–95.

James, T.Y., Stajich, J.E., Hittinger, C.T., Rokas, A., 2020. Toward a Fully Resolved Fungal Tree of Life. Annu. Rev. Microbiol. 74, 291–313.

Kalisky, T., Quake, S.R., 2011. Single-cell genomics. Nat. Methods 8, 311–314.

Keller, S., 1997. The genus Neozygites (Zygomycetes , Entomophthorales) with special reference to species found in tropical regions. Sydowia 49, 118–146.

Mikhailov, K.V., Simdyanov, T.G., Aleoshin, V.V., Belozersky, A.N., 2016. Genomic survey of a hyperparasitic microsporidian Amphiamblys sp. (Metchnikovellidae). Genome Biol. Evol. 9, 454–467.

Mondo, S.J., Toomer, K.H., Morton, J.B., Lekberg, Y., Pawlowska, T.E., 2012. Evolutionary stability in a 400-million-year-old heritable facultative mutualism. Evolution 66, 2564–2576.

Naito, M., Morton, J.B., Pawlowska, T.E., 2015. Minimal genomes of mycoplasma-related endobacteria are plastic and contain host-derived genes for sustained life within Glomeromycota. Proc. Natl. Acad. Sci. U. S. A. 112, 7791–7796.

Pace, N.R., 1997. A molecular view of microbial diversity and the biosphere. Science 276, 734–740.

Partida-Martinez, L.P., Hertweck, C., 2005. Pathogenic fungus harbours endosymbiotic bacteria for toxin production. Nature 437, 884–888.

Pawlowska, T.E., Gaspar, M.L., Lastovetsky, O.A., Mondo, S.J., Real-Ramirez, I., Shakya, E., Bonfante, P., 2018. Biology of Fungi and Their Bacterial Endosymbionts. Annu. Rev. Phytopathol. 56, 289–309.

Pinard, R., de Winter, A., Sarkis, G.J., Gerstein, M.B., Tartaro, K.R., Plant, R.N., Egholm, M., Rothberg, J.M., Leamon, J.H., 2006. Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. BMC Genomics 7, 216.

Porter, J.R., 1976. Antony van Leeuwenhoek: tercentenary of his discovery of bacteria. Bacteriol. Rev. 40, 260–269.

Prasanna, A.N., Gerber, D., Kijpornyongpan, T., Aime, M.C., Doyle, V.P., Nagy, L.G., 2019. Model Choice, Missing Data, and Taxon Sampling Impact Phylogenomic Inference of Deep Basidiomycota Relationships. Syst. Biol. https://doi.org/10.1093/sysbio/syz029

Rinke, C., Lee, J., Nath, N., Goudeau, D., Thompson, B., Poulton, N., Dmitrieff, E., Malmstrom, R., Stepanauskas, R., Woyke, T., 2014. Obtaining genomes from uncultivated environmental microorganisms using FACS – based single-cell genomics. Nat. Protoc. 9, 1038–1048.

Schoch, C.L., Seifert, K. a., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C. a., Chen, W., Consortium, F.B., Bolchacova, E., Voigt, K., Crous, P.W., Miller, a. N., Wingfield, M.J., Aime, M.C., An, K.-D., Bai, F.-Y., Barreto, R.W., Begerow, D., Bergeron, M.-J., Blackwell, M., Boekhout, T., Bogale, M., Boonyuen, N., Burgaz, a. R., Buyck, B., Cai, L., Cai, Q., Cardinali, G., Chaverri, P., Coppins, B.J., Crespo, A., Cubas, P., Cummings, C., Damm, U., De Beer, Z.W., de Hoog, G.S., Del-Prado, R., Dentinger, B., Dieguez-Uribeondo, J., Divakar, P.K., Douglas, B., Duenas, M., Duong, T. a., Eberhardt, U., Edwards, J.E., Elshahed, M.S., Fliegerova, K., Furtado, M., Garcia, M. a., Ge, Z.-W., Griffith, G.W., Griffiths, K., Groenewald, J.Z., Groenewald, M., Grube, M., Gryzenhout, M., Guo, L.-D., Hagen, F., Hambleton, S., Hamelin, R.C., Hansen, K., Harrold, P., Heller, G., Herrera, C., Hirayama, K., Hirooka, Y., Ho, H.-M., Hoffmann, K., Hofstetter, V., Hognabba, F., Hollingsworth, P.M., Hong, S.-B.S.-B.S.-B.S.-B., Hosaka, K., Houbraken, J., Hughes, K., Huhtinen, S., Hyde, K.D., James, T., Johnson, E.M., Johnson, J.E., Johnston, P.R., Jones, E.B.G., Kelly, L.J., Kirk, P.M., Knapp, D.G., Koljalg, U., Kovacs, G.M., Kurtzman, C.P., Landvik, S., Leavitt, S.D., Liggenstoffer, a. S., Liimatainen, K., Lombard, L., Luangsa-ard, J.J., Lumbsch, H.T., Maganti, H., Maharachchikumbura, S.S.N., Martin, M.P., May, T.W., McTaggart, a. R., Methven, a. S., Meyer, W., Moncalvo, J.-M., Mongkolsamrit, S., Nagy, L.G., Nilsson, R.H., Niskanen, T., Nyilasi, I., Okada, G., Okane, I., Olariaga, I., Otte, J., Papp, T., Park, D., Petkovits, T., Pino-Bodas, R., Quaedvlieg, W., Raja, H. a., Redecker, D., Rintoul, T.L., Ruibal, C., Sarmiento-Ramirez, J.M., Schmitt, I., Schussler, A., Shearer, C., Sotome, K., Stefani, F.O.P., Stenroos, S., Stielow, B., Stockinger, H., Suetrong, S., Suh, S.-O., Sung, G.-H., Suzuki, M., Tanaka, K., Tedersoo, L., Telleria, M.T., Tretter, E., Untereiner, W. a., Urbina, H., Vagvolgyi, C., Vialle, A., Vu, T.D., Walther, G., Wang, Q.-M., Wang, Y., Weir, B.S., Weiss, M., White, M.M., Xu, J., Yahr, R., Yang, Z.L., Yurkov, A., Zamora, J.-C., Zhang, N., Zhuang, W.-Y.W.-Y., Schindel, D., 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc. Natl. Acad. Sci. U. S. A. 109, 1–6.

Sedlar, K., Kupkova, K., Provaznik, I., 2017. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. Comput. Struct. Biotechnol. J. 15, 48–55.

Sun, X., Chen, W., Ivanov, S., MacLean, A.M., Wight, H., Ramaraj, T., Mudge, J., Harrison, M.J., Fei, Z., 2019. Genome and evolution of the arbuscular mycorrhizal fungus Diversispora epigaea (formerly Glomus versiforme) and its bacterial endosymbionts. New Phytol. 221, 1556–1573.

Toomer, K.H., Chen, X., Naito, M., Mondo, S.J., den Bakker, H.C., VanKuren, N.W., Lekberg, Y., Morton, J.B., Pawlowska, T.E., 2015. Molecular evolution patterns reveal life history features of mycoplasma-related endobacteria associated with arbuscular mycorrhizal fungi. Mol. Ecol. 24, 3485–3500.

Torres-Cortés, G., Ghignone, S., Bonfante, P., Schüßler, A., 2015. Mosaic genome of endobacteria in arbuscular mycorrhizal fungi: Transkingdom gene transfer in an ancient mycoplasma-fungus association. Proc. Natl. Acad. Sci. U. S. A. 112, 7785–7790.

Tucker, T., Marra, M., Friedman, J.M., 2009. Massively parallel sequencing: the next big thing in genetic medicine. Am. J. Hum. Genet. 85, 142–154.

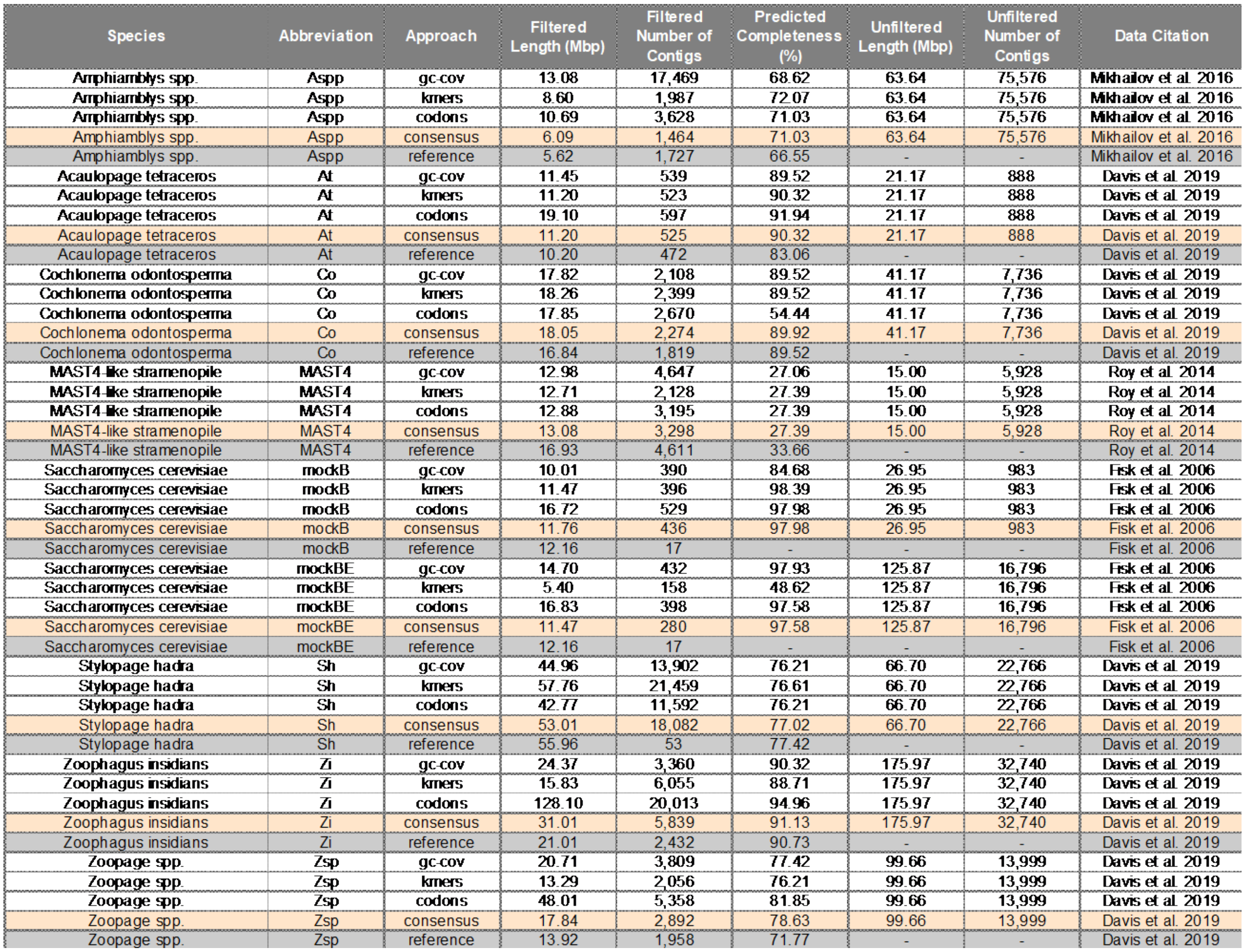
van Niel, C.B., 1944. THE CULTURE, GENERAL PHYSIOLOGY, MORPHOLOGY, AND CLASSIFICATION OF THE NON-SULFUR PURPLE AND BROWN BACTERIA. Bacteriol. Rev. 8, 1–118.

White, M.M., James, T.Y., Donnell, K.O., Cafaro, M.J., Tanabe, Y., Sugiyama, J., James, T.Y., Carolina, N., Donnell, K.O., 2006. Phylogeny of the Zygomycota based on nuclear ribosomal sequence data. Mycologia 98, 872–884.

Yue, Y., Huang, H., Qi, Z., Dou, H.-M., Liu, X.-Y., Han, T.-F., Chen, Y., Song, X.-J., Zhang, Y.-H., Tu, J., 2020. Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. BMC Bioinformatics 21, 334.

Appendix A Supplementary Table and Figures for Chapter 2

Table A1. Comparative assembly statistics for all the SCG mock and genuine datasets used to test SCGid in this study. Statistics are shown before and after automated filtering with SCGid. SCGid consensus and published reference assemblies are highlighted.



Species	Abbreviation	Approach	Filtered Length (Mbp)	Filtered Number of Contigs	Predicted Completeness (%)	Unfiltered Length (Mbp)	Unfiltered Number of Contigs	Data Citation
Amphiamblys spp.	Aspp	gc-cov	13.08	17,469	68.62	63.64	75,576	Mikhailov et al. 2016
Amphiamblys spp.	Aspp	kmers	8.60	1,987	72.07	63.64	75,576	Mikhailov et al. 2016
Amphiamblys spp.	Aspp	codons	10.69	3,628	71.03	63.64	75,576	Mikhailov et al. 2016
Amphiamblys spp.	Aspp	consensus	6.09	1,464	71.03	63.64	75,576	Mikhailov et al. 2016
Amphiamblys spp.	Aspp	reference	5.62	1,727	66.55	-	-	Mikhailov et al. 2016
Acaulopage tetraceros	At	gc-cov	11.45	539	89.52	21.17	888	Davis et al. 2019
Acaulopage tetraceros	At	kmers	11.20	523	90.32	21.17	888	Davis et al. 2019
Acaulopage tetraceros	At	codons	19.10	597	91.94	21.17	888	Davis et al. 2019
Acaulopage tetraceros	At	consensus	11.20	525	90.32	21.17	888	Davis et al. 2019
Acaulopage tetraceros	At	reference	10.20	472	83.06	-	-	Davis et al. 2019
Cochlonema odontosperma	Co	gc-cov	17.82	2,108	89.52	41.17	7,736	Davis et al. 2019
Cochlonema odontosperma	Co	kmers	18.26	2,399	89.52	41.17	7,736	Davis et al. 2019
Cochlonema odontosperma	Co	codons	17.85	2,670	54.44	41.17	7,736	Davis et al. 2019
Cochlonema odontosperma	Co	consensus	18.05	2,274	89.92	41.17	7,736	Davis et al. 2019
Cochlonema odontosperma	Co	reference	16.84	1,819	89.52	-	-	Davis et al. 2019
MAST4-like stramenopile	MAST4	gc-cov	12.98	4,647	27.06	15.00	5,928	Roy et al. 2014
MAST4-like stramenopile	MAST4	kmers	12.71	2,128	27.39	15.00	5,928	Roy et al. 2014
MAST4-like stramenopile	MAST4	codons	12.88	3,195	27.39	15.00	5,928	Roy et al. 2014
MAST4-like stramenopile	MAST4	consensus	13.08	3,298	27.39	15.00	5,928	Roy et al. 2014
MAST4-like stramenopile	MAST4	reference	16.93	4,611	33.66	-	-	Roy et al. 2014
Saccharomyces cerevisiae	mockB	gc-cov	10.01	390	84.68	26.95	983	Fisk et al. 2006
Saccharomyces cerevisiae	mockB	kmers	11.47	396	98.39	26.95	983	Fisk et al. 2006
Saccharomyces cerevisiae	mockB	codons	16.72	529	97.98	26.95	983	Fisk et al. 2006
Saccharomyces cerevisiae	mockB	consensus	11.76	436	97.98	26.95	983	Fisk et al. 2006
Saccharomyces cerevisiae	mockB	reference	12.16	17	-	-	-	Fisk et al. 2006
Saccharomyces cerevisiae	mockBE	gc-cov	14.70	432	97.93	125.87	16,796	Fisk et al. 2006
Saccharomyces cerevisiae	mockBE	kmers	5.40	158	48.62	125.87	16,796	Fisk et al. 2006
Saccharomyces cerevisiae	mockBE	codons	16.83	398	97.58	125.87	16,796	Fisk et al. 2006
Saccharomyces cerevisiae	mockBE	consensus	11.47	280	97.58	125.87	16,796	Fisk et al. 2006
Saccharomyces cerevisiae	mockBE	reference	12.16	17	-	-	-	Fisk et al. 2006
Stylopage hadra	Sh	gc-cov	44.96	13,902	76.21	66.70	22,766	Davis et al. 2019
Stylopage hadra	Sh	kmers	57.76	21,459	76.61	66.70	22,766	Davis et al. 2019
Stylopage hadra	Sh	codons	42.77	11,592	76.21	66.70	22,766	Davis et al. 2019
Stylopage hadra	Sh	consensus	53.01	18,082	77.02	66.70	22,766	Davis et al. 2019
Stylopage hadra	Sh	reference	55.96	53	77.42	-	-	Davis et al. 2019
Zoophagus insidians	Zi	gc-cov	24.37	3,360	90.32	175.97	32,740	Davis et al. 2019
Zoophagus insidians	Zi	kmers	15.83	6,055	88.71	175.97	32,740	Davis et al. 2019
Zoophagus insidians	Zi	codons	128.10	20,013	94.96	175.97	32,740	Davis et al. 2019
Zoophagus insidians	Zi	consensus	31.01	5,839	91.13	175.97	32,740	Davis et al. 2019
Zoophagus insidians	Zi	reference	21.01	2,432	90.73	-	-	Davis et al. 2019
Zoopage spp.	Zsp	gc-cov	20.71	3,809	77.42	99.66	13,999	Davis et al. 2019
Zoopage spp.	Zsp	kmers	13.29	2,056	76.21	99.66	13,999	Davis et al. 2019
Zoopage spp.	Zsp	codons	48.01	5,358	81.85	99.66	13,999	Davis et al. 2019
Zoopage spp.	Zsp	consensus	17.84	2,892	78.63	99.66	13,999	Davis et al. 2019
Zoopage spp.	Zsp	reference	13.92	1,958	71.77	-	-	Davis et al. 2019

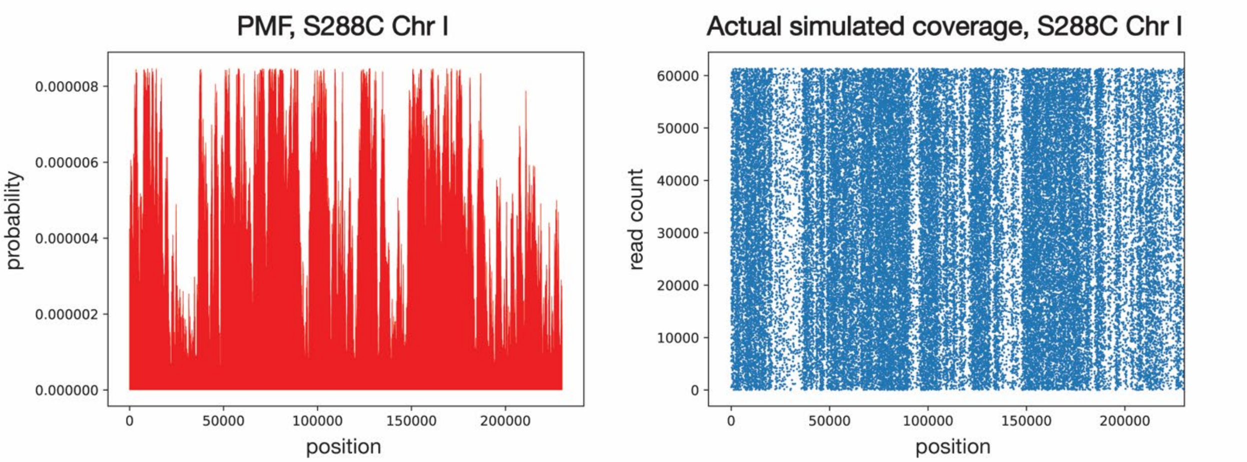


Figure A1 Plots showing exemplar probability mass functions (left) and scatter plot showing draws from that PMF (right). Draws from the PMF represent start locations for 500 bp fragments, the 150bp paired-end reads from which were used to assemble the mock MDA library. PMFs like this were generated for all twenty initial chromosomes constituting the mock metagenome.

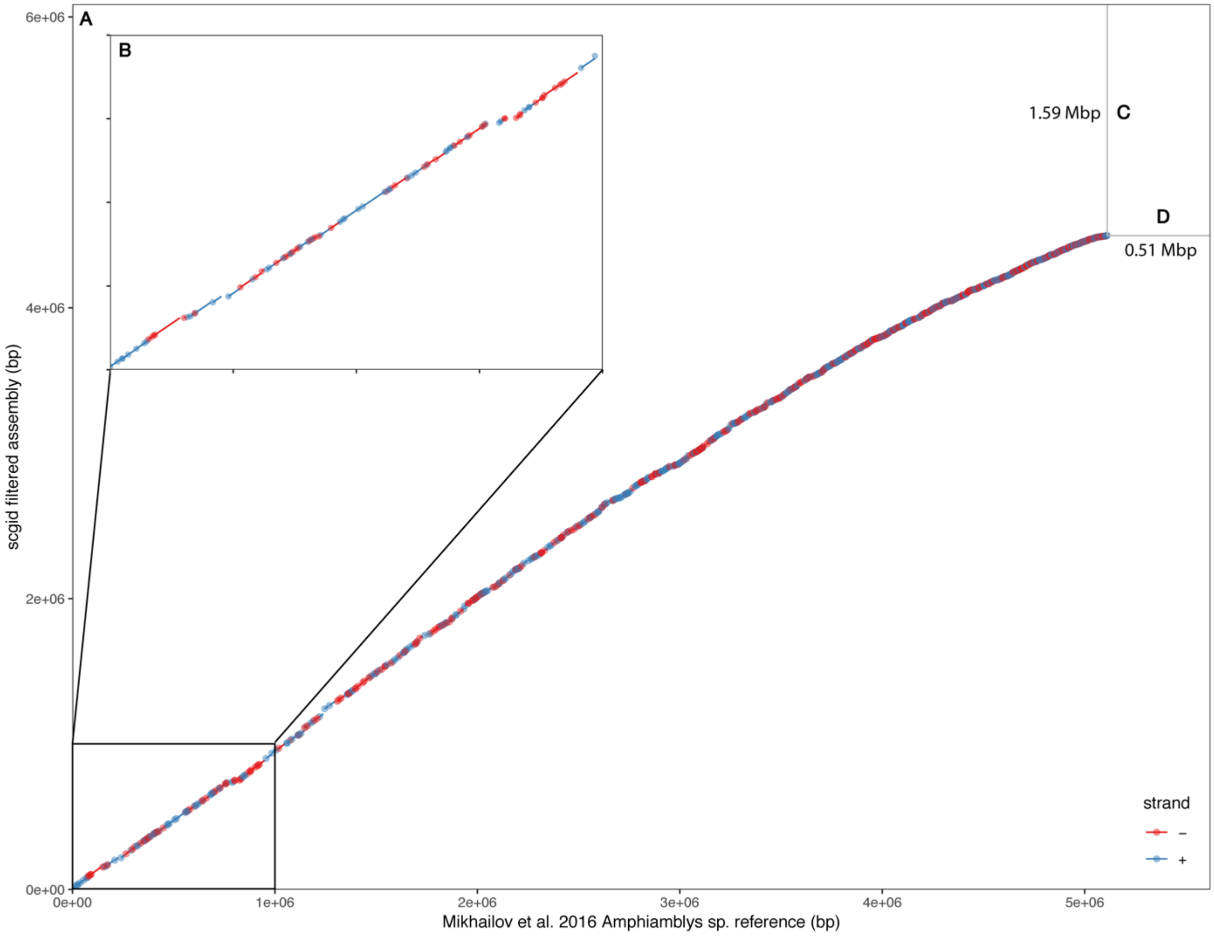


Figure A2. (A)Whole genome alignment showing correspondence between the SCG*id*-predicted genome draft for *Amphiamblys* sp. (y-axis) with that reported by Mikhailov et al. 2016. (x-axis). (B) Expanded view of the whole genome alignment shown in (A) showing the first megabase of the alignment. (C) Cumulative size of SCG*id*-predicted draft genome that does not align with the published draft. (D) Cumulative size of the published draft that does not align with that predicted by SCG*id*.

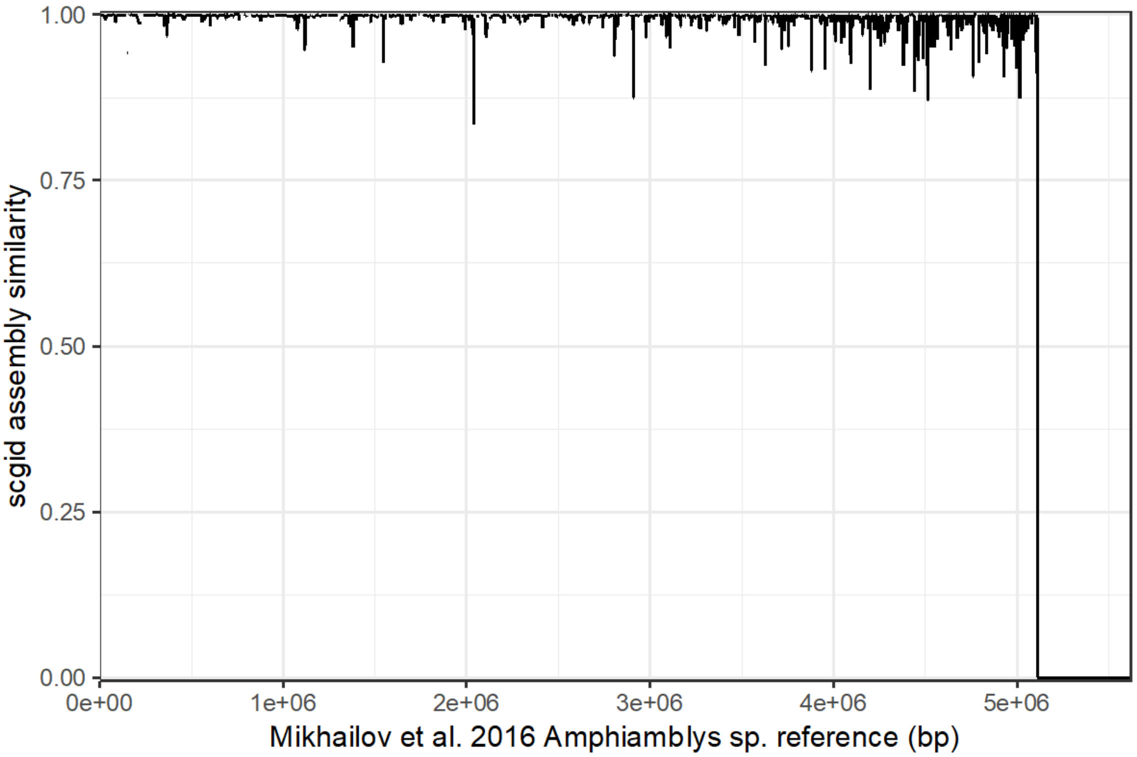


Figure A3. Percent sequence identity along the Mikhailov et al. versus SCG*id*-derived *Amphiamblys* sp. whole genome alignment, ordered by decreasing contig size.

Appendix B Supplementary Tables and Figure for Chapter 3

Table B1. Species, accession, and MRE/non-MRE classification of all MRE non-MRE Mollicutes, and outgroup genomes used in 16S and genome-scale phylogenetic analyses.

Species	Accession	Group	Species	Accession	Group	Species	Accession	Group
DhMRE	DhMRE	MRE	Mycoplasma canadense	GCA_000828855.1	Non-MRE	Mycoplasma todarodis	GCA_004335995.1	Non-MRE
DeMREI-1	GvMREI-1	MRE	Mycoplasma yeatsii	GCA_000875755.1	Non-MRE	Mycoplasma testudineum	GCA_004362335.1	Non-MRE
DeMREI-2	GvMREI-2	MRE	Mycoplasma dispar	GCA_000941075.1	Non-MRE	Mycoplasma mustelae	GCA_004365095.1	Non-MRE
DeMREII	GvMREII	MRE	Acholeplasma oculi	GCA_000953195.1	Non-MRE	Spiroplasma gladiatoris	GCA_004379335.1	Non-MRE
CeMRE	CeMRE	MRE	Acholeplasma brassicae	GCA_000967915.1	Non-MRE	Spiroplasma melliferum	GCA_005222125.1	Non-MRE
RcMRE	RcMRE	MRE	Acholeplasma palmae	GCA_000968055.1	Non-MRE	Mycoplasma nasistruthionis	GCA_006228185.1	Non-MRE
RvMRE	RvMRE	MRE	Mycoplasma synoviae	GCA_000969765.1	Non-MRE	Mycoplasma equirhinis	GCA_006385185.1	Non-MRE
RhopMRE	RhopMRE	MRE	Spiroplasma atrichopogonis	GCA_001029245.1	Non-MRE	Mycoplasma falconis	GCA_006385795.1	Non-MRE
Mycoplasma hyopneumoniae	GCA_000008205.1	Non-MRE	Spiroplasma eriocheiris	GCA_001029265.1	Non-MRE	Mycoplasma neophronis	GCA_006491995.1	Non-MRE
Mesoplasma florum	GCA_000008305.1	Non-MRE	Spiroplasma turonicum	GCA_001262715.1	Non-MRE	Mycoplasma mucosicanis	GCA_006546935.1	Non-MRE
Mycoplasma mobile	GCA_000008365.1	Non-MRE	Spiroplasma litorale	GCA_001267155.1	Non-MRE	Mycoplasma anserisalpingitidis	GCA_007859615.1	Non-MRE
Mycoplasma penetrans	GCA_000011225.1	Non-MRE	Mycoplasma pneumoniae	GCA_001272835.1	Non-MRE	Cynodon dactylon phytoplasma	GCA_009268075.1	Non-MRE
Aster yellows witches-broom phytoplasma	GCA_000012225.1	Non-MRE	Spiroplasma kunkelii	GCA_001274875.1	Non-MRE	Saccharum officinarum phytoplasma SCGS	GCA_009268105.1	Non-MRE
Mycoplasma capricolum	GCA_000012765.1	Non-MRE	Spiroplasma cantharicola	GCA_001281045.1	Non-MRE	Spiroplasma tabanidicola	GCA_009730595.1	Non-MRE
Acholeplasma laidlawii	GCA_000018785.1	Non-MRE	Echinacea purpurea witches-broom phytoplasma	GCA_001307505.1	Non-MRE	Mycoplasma iowae	GCA_009883755.1	Non-MRE
Ureaplasma parvum	GCA_000019345.1	Non-MRE	Mycoplasma canis	GCA_001553195.1	Non-MRE	Mycoplasma felis	GCA_009936335.1	Non-MRE
Ureaplasma urealyticum	GCA_000021265.1	Non-MRE	Maize bushy stunt phytoplasma	GCA_001712875.1	Non-MRE	Mycoplasma verecundum	GCA_900167035.1	Non-MRE
Mycoplasma crocodyli	GCA_000025845.1	Non-MRE	Spiroplasma helicoides	GCA_001715535.1	Non-MRE	Mycoplasma agassizii	GCA_900176265.1	Non-MRE
Mycoplasma genitalium	GCA_000027325.1	Non-MRE	Rice orange leaf phytoplasma	GCA_001866375.1	Non-MRE	Mycoplasma edwardii	GCA_900476105.1	Non-MRE
Mycoplasma agalactiae	GCA_000063605.1	Non-MRE	Spiroplasma citri	GCA_001886855.1	Non-MRE	Mycoplasma alkalescens	GCA_900476125.1	Non-MRE
Mycoplasma hominis	GCA_000085865.1	Non-MRE	Mycoplasma pullorum	GCA_001900245.1	Non-MRE	Mycoplasma caviae	GCA_900631685.1	Non-MRE
Mycoplasma gallisepticum	GCA_000092585.1	Non-MRE	Mycoplasma hyosynoviae	GCA_002214445.1	Non-MRE	Mycoplasma orale	GCA_900660435.1	Non-MRE
Mycoplasma mycoides	GCA_000143865.1	Non-MRE	Spiroplasma corruscae	GCA_002237575.1	Non-MRE	Mycoplasma salivarium	GCA_900660445.1	Non-MRE
Mycoplasma alligatoris	GCA_000178375.1	Non-MRE	Mesoplasma chauliocola	GCA_002290085.1	Non-MRE	Mycoplasma neurolyticum	GCA_900660485.1	Non-MRE
Mycoplasma suis	GCA_000179035.2	Non-MRE	Mesoplasma lactucae	GCA_002441935.1	Non-MRE	Mycoplasma gallinaceum	GCA_900660495.1	Non-MRE
Mycoplasma leachii	GCA_000183365.1	Non-MRE	Mesoplasma entomophilum	GCA_002749675.1	Non-MRE	Mycoplasma bovirhinis	GCA_900660515.1	Non-MRE
Mycoplasma bovis	GCA_000183385.1	Non-MRE	Spiroplasma clarkii	GCA_002795265.1	Non-MRE	Mycoplasma bovigenitalium	GCA_900660525.1	Non-MRE
Mycoplasma fermentans	GCA_000186005.1	Non-MRE	Entomoplasma luminosum	GCA_002803985.1	Non-MRE	Mycoplasma cynos	GCA_900660545.1	Non-MRE
Mycoplasma haemofelis	GCA_000200735.1	Non-MRE	Entomoplasma somnilux	GCA_002804005.1	Non-MRE	Mycoplasma conjunctivae	GCA_900660555.1	Non-MRE
Mycoplasma putrefaciens	GCA_000224105.1	Non-MRE	Mesoplasma tabanidae	GCA_002804025.1	Non-MRE	Mycoplasma pulmonis	GCA_900660575.1	Non-MRE
Mycoplasma haemocanis	GCA_000238995.1	Non-MRE	Entomoplasma melaleucae	GCA_002804105.1	Non-MRE	Mycoplasma glycophilum	GCA_900660605.1	Non-MRE
Mycoplasma wenyonii	GCA_000277795.1	Non-MRE	Entomoplasma freundtii	GCA_002804205.1	Non-MRE	Mycoplasma gallopavonis	GCA_900660635.1	Non-MRE
Mycoplasma hyorhinis	GCA_000313635.1	Non-MRE	Mesoplasma coleopterae	GCA_002804245.1	Non-MRE	Mycoplasma citelli	GCA_900660645.1	Non-MRE
Mycoplasma feriruminatoris	GCA_000327395.1	Non-MRE	Spiroplasma floricola	GCA_002813555.1	Non-MRE	Mycoplasma anatis	GCA_900660655.1	Non-MRE
Peanut witches-broom phytoplasma	GCA_000364425.1	Non-MRE	Mesoplasma syrphidae	GCA_002843565.1	Non-MRE	Mycoplasma maculosum	GCA_900660665.1	Non-MRE
Spiroplasma chrysopicola	GCA_000400935.1	Non-MRE	Spiroplasma monobiae	GCA_002865545.1	Non-MRE	Mycoplasma columborale	GCA_900660675.1	Non-MRE
Spiroplasma syrphidicola	GCA_000400955.1	Non-MRE	Mesoplasma corruscae	GCA_002930145.1	Non-MRE	Mycoplasma columbinum	GCA_900660685.1	Non-MRE
Spiroplasma taiwanense	GCA_000439435.1	Non-MRE	Entomoplasma ellychniae	GCA_002930155.1	Non-MRE	Mycoplasma meleagridis	GCA_900660695.1	Non-MRE
Spiroplasma diminutum	GCA_000439455.1	Non-MRE	Mycoplasma auris	GCA_003253435.1	Non-MRE	Mycoplasma columbinasale	GCA_900660705.1	Non-MRE
Mycoplasma parvum	GCA_000477415.1	Non-MRE	Mycoplasma anseris	GCA_003285045.1	Non-MRE	Mycoplasma arthritidis	GCA_900660715.1	Non-MRE
Spiroplasma apis	GCA_000500935.1	Non-MRE	Mycoplasma phocidae	GCA_003332325.1	Non-MRE	Mycoplasma arginini	GCA_900660725.1	Non-MRE
Mycoplasma ovis	GCA_000508245.1	Non-MRE	Spiroplasma phoeniceum	GCA_003339775.1	Non-MRE	Mycoplasma cloacale	GCA_900660735.1	Non-MRE
Mycoplasma bovoculi	GCA_000524555.1	Non-MRE	Spiroplasma alleghenense	GCA_003363775.1	Non-MRE	Acholeplasma axanthum	GCA_900660745.1	Non-MRE
Spiroplasma culicicola	GCA_000565175.1	Non-MRE	Mycoplasma phocicerebrale	GCA_003383595.3	Non-MRE	Acholeplasma hippikon	GCA_900660755.1	Non-MRE
Spiroplasma mirum	GCA_000565195.1	Non-MRE	Anaeroplasma bactoclasticum	GCA_003550015.1	Non-MRE	Bacillus cereus	GCF_000007825.1	Outgroup
Spiroplasma sabaudiense	GCA_000565215.1	Non-MRE	Mycoplasma subdolum	GCA_003688445.1	Non-MRE	Staphylococcus aureus	GCF_000013425.1	Outgroup
Mycoplasma californicum	GCA_000695835.1	Non-MRE	Mycoplasma struthionis	GCA_003855455.1	Non-MRE	Lactobacillus paracasei	GCF_000014525.1	Outgroup
Ureaplasma diversum	GCA_000731915.1	Non-MRE	Spiroplasma endosymbiont of Megaselia nigra	GCA_003987485.1	Non-MRE			
Chrysanthemum coronarium phytoplasma	GCA_000744065.1	Non-MRE	Catharanthus roseus aster yellows phytoplasma	GCA_004214875.1	Non-MRE			
Mycoplasma flocculare	GCA_000815065.1	Non-MRE	Mycoplasma phocirhinis	GCA_004216495.1	Non-MRE			
Spiroplasma poulsonii	GCA_000820525.2	Non-MRE	Mycoplasma marinum	GCA_004335975.1	Non-MRE			

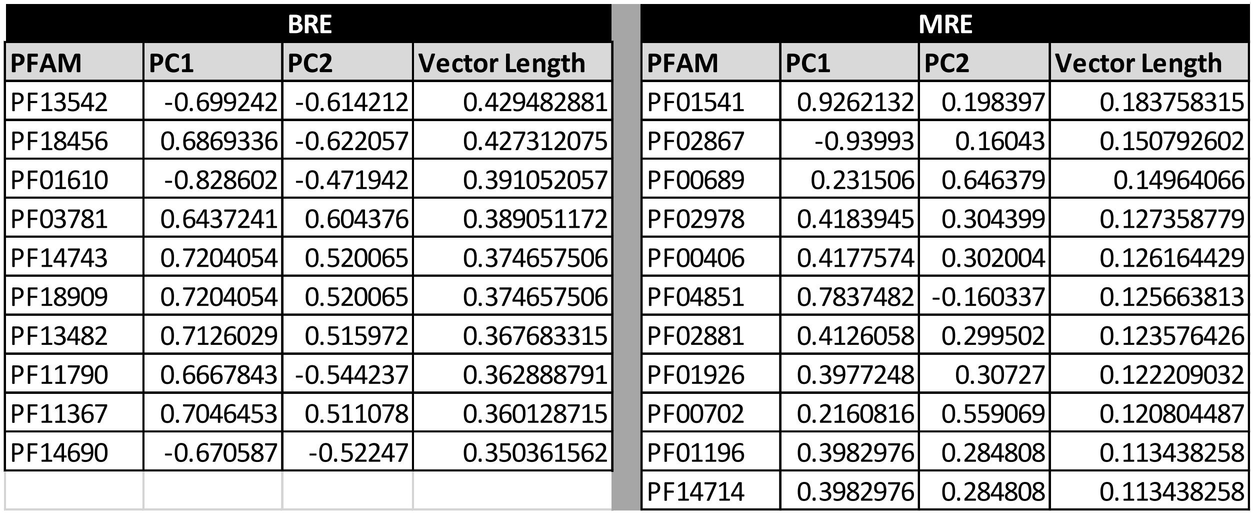
Table B2. Species, strain, accession, and BRE/non-BRE classification of all BRE and non-BRE Burkholderiaceae genomes used in 16S and genome-scale phylogenetic analyses.

Species	Strain	Accession	Clade	Species	Strain	Accession	Clade
Mycoavidus cysteinexigens	B1-EB	GCF_003966915.1	BRE	Paraburkholderia phenoliruptrix	BR3459a	GCF_000300095.1	Non-BRE
Mycoavidus sp.	B2-EB	GCF_014218255.1	BRE	Paraburkholderia fungorum	ATCC BAA-463	GCF_000961515.1	Non-BRE
Candidatus Glomeribacter gigasporarum	BEG34	GCA_000227585.1	BRE	Paraburkholderia sprentiae	WSM5005	GCF_001865575.1	Non-BRE
Candidatus Glomeribacter gigasporarum	BEG1	GCA_001684025.1	BRE	Paraburkholderia caledonica	PHRS4	GCF_003330745.1	Non-BRE
Candidatus Glomeribacter gigasporarum	IN211	GCA_001684175.1	BRE	Paraburkholderia graminis	PHS1	GCF_003330785.1	Non-BRE
Candidatus Glomeribacter gigasporarum	JA201A	GCA_001684155.1	BRE	Paraburkholderia terricola	mHS1	GCF_003330825.1	Non-BRE
Mycetohabitans rhizoxinica	HKI454	GCF_000198775.1	BRE	Paraburkholderia caribensis	852011	GCF_013378095.1	Non-BRE
Mycetohabitans endofungorum	HKI456	GCF_002927045.1	BRE	Paraburkholderia tropica	IAC135	GCF_014171495.1	Non-BRE
Mycoavidus cysteinexigens	AG77	PATRIC_224135.3	BRE	Paraburkholderia ginsengisoli	FDAARGOS_1049	GCF_016128195.1	Non-BRE
Mycoavidus sp.	SOG	N/A	BRE	Pandoraea pnomenusa	RB-44	GCF_000504585.2	Non-BRE
Burkholderia thailandensis	E264	GCF_000152285.1	Non-BRE	Pandoraea fibrosis	6399	GCF_000807775.2	Non-BRE
Burkholderia dolosa	1.0	GCF_000497165.1	Non-BRE	Pandoraea apista	FDAARGOS_126	GCF_002951195.1	Non-BRE
Burkholderia pseudomallei	BGK	GCF_000763555.1	Non-BRE	Ralstonia insidiosa	FC1138	GCF_001653935.1	Non-BRE
Burkholderia mallei	KC-1092	GCF_000959585.1	Non-BRE	Ralstonia pickettii	12J	GCF_000020205.1	Non-BRE
Burkholderia humptydooensis	MSMB122	GCF_001462435.1	Non-BRE	Ralstonia solanacearum	Po82	GCF_000215325.1	Non-BRE
Burkholderia sp.	MSMB1588	GCF_001546925.1	Non-BRE	Ralstonia sp.	UNCCL144	GCF_900099845.1	Non-BRE
Burkholderia ubonensis	MSMB1157	GCF_001546975.1	Non-BRE	Cupriavidus basilensis	4G11	GCF_000832305.1	Non-BRE
Burkholderia anthina	AZ-4-2-10-S1-D7	GCF_001547525.1	Non-BRE	Cupriavidus gilardii	FDAARGOS_639	GCF_013347325.1	Non-BRE
Burkholderia territorii	MSMB2203WGS	GCF_001636095.1	Non-BRE	Cupriavidus metallidurans	CH34	GCF_000196015.1	Non-BRE
Burkholderia cenocepacia	VC2307	GCF_001999805.1	Non-BRE	Cupriavidus necator	H16	GCF_000009285.1	Non-BRE
Burkholderia contaminans	170816	GCF_002924455.1	Non-BRE	Cupriavidus pauculus	FDAARGOS_664	GCF_008693385.1	Non-BRE
Burkholderia multivoran	FDAARGOS_624	GCF_012272655.1	Non-BRE	Cupriavidus pinatubonensis	JMP134	GCF_000203875.1	Non-BRE
Burkholderia humptydooensis	Bp5365	GCF_001513745.1	Non-BRE	Caballeronia insecticola	RPE64	GCF_000402035.1	Non-BRE
Burkholderia stabilis	NA	GCF_900240005.1	Non-BRE	Polynucleobacter asymbioticus	QLW-P1DMWA-1	GCF_000016345.1	Non-BRE
Paraburkholderia xenovorans	LB400	GCF_000013645.1	Non-BRE	Polynucleobacter difficilis	AM-8B5	GCF_003065365.1	Non-BRE
Paraburkholderia phytofirmans	PsJN	GCF_000020125.1	Non-BRE	Polynucleobacter necessarius	STIR1	GCF_000019745.1	Non-BRE
Paraburkholderia atlantica	CCGE1002	GCF_000092885.1	Non-BRE	Polynucleobacter paneuropaeus	MWH-Creno-4B4	GCF_003261295.1	Non-BRE

Table B3. HGT candidates identified by our AIS-based approach (i.e., AIS >= 20). HGT candidates determined to be strong cases of transfer from Eukaryota to RhopMRE are highlighted in green. HGT candidates whose gene trees seem to suggest transfer in the opposite direction (i.e., MRE to Fungi) are highlighted in yellow. HGT candidates that we excluded based on uninformative gene trees are highlighted in red.

Query Protein Identifier	Subject Protein Identifier	Subject Annotation	Alien Evalue	Domestic Evalue	AIS
NODE_14_length_30749_cov_1356.745710_32	UniRef90_D2VEB3	Predicted Protein	4.30E-117	3.89E-76	94.305783
NODE_1_length_115238_cov_2519.147526_6	UniRef90_A0A2P4QSF0	Uncharacterized Protein	1.13E-16	0.51	36.0457993
NODE_1_length_115238_cov_2519.147526_33	UniRef90_A0A397T5I1	Uncharacterized Protein	3.16E-84	1.12E-37	107.184256
NODE_1_length_115238_cov_2519.147526_35	UniRef90_A0A397UNT9	Jacalin-type lectin	3.63E-31	6.26E-08	53.5044047
NODE_1_length_115238_cov_2519.147526_76	UniRef90_A0A177END3	Formamidopyrimidine-DNA glcosylase	6.99E-92	3.27E-82	22.2661604
NODE_1_length_115238_cov_2519.147526_119	UniRef90_A0A397SCF0	Plasmid maintenance toxin/Cell growth inhibitor	9.52E-36	6.39E-19	38.745286
NODE_1_length_115238_cov_2519.147526_127	UniRef90_A0A397SD62	AIG1 family-domain-containing protein	5.33E-79	6.10E-70	20.8582034
NODE_1_length_115238_cov_2519.147526_131	UniRef90_B0DBP4	Predicted Protein	3.84E-19	0.04	39.1847686
NODE_1_length_115238_cov_2519.147526_135	UniRef90_A0A2Z6S9M1	Uncharacterized Protein	9.34E-11	0.27	21.7847965
NODE_3_length_78848_cov_1743.355255_28	UniRef90_A0A2I1F0M1	Psuedouridine synthase	6.00E-129	3.76E-64	149.200691
NODE_3_length_78848_cov_1743.355255_48	UniRef90_A0A397GDM2	AIG1-type G domain-containing protein	1.62E-89	5.66E-68	49.6052847
NODE_3_length_78848_cov_1743.355255_49	UniRef90_A0A2P4Q6G5	AIG1-type G domain-containing protein	2.06E-50	4.41E-41	21.4844345
NODE_7_length_43739_cov_1761.773511_3	UniRef90_UPI00026589F3	mexicain	4.13E-95	1	217.327306
NODE_7_length_43739_cov_1761.773511_10	UniRef90_A0A2I1GIU8	zf-3CxxC domain-containing protein	5.30E-48	0.87	108.717116
NODE_7_length_43739_cov_1761.773511_39	UniRef90_A0A397SJV3	P-loop containing nucleoside triphosphate hydrolase protein	5.89E-75	1.78E-49	58.6705698

Table B4. Top influential PFAMs as determined by our phylogenetically-scaled PCAs for BRE (left) and MRE (right).



BRE				MRE			
PFAM	PC1	PC2	Vector Length	PFAM	PC1	PC2	Vector Length
PF13542	-0.699242	-0.614212	0.429482881	PF01541	0.9262132	0.198397	0.183758315
PF18456	0.6869336	-0.622057	0.427312075	PF02867	-0.93993	0.16043	0.150792602
PF01610	-0.828602	-0.471942	0.391052057	PF00689	0.231506	0.646379	0.14964066
PF03781	0.6437241	0.604376	0.389051172	PF02978	0.4183945	0.304399	0.127358779
PF14743	0.7204054	0.520065	0.374657506	PF00406	0.4177574	0.302004	0.126164429
PF18909	0.7204054	0.520065	0.374657506	PF04851	0.7837482	-0.160337	0.125663813
PF13482	0.7126029	0.515972	0.367683315	PF02881	0.4126058	0.299502	0.123576426
PF11790	0.6667843	-0.544237	0.362888791	PF01926	0.3977248	0.30727	0.122209032
PF11367	0.7046453	0.511078	0.360128715	PF00702	0.2160816	0.559069	0.120804487
PF14690	-0.670587	-0.52247	0.350361562	PF01196	0.3982976	0.284808	0.113438258
				PF14714	0.3982976	0.284808	0.113438258

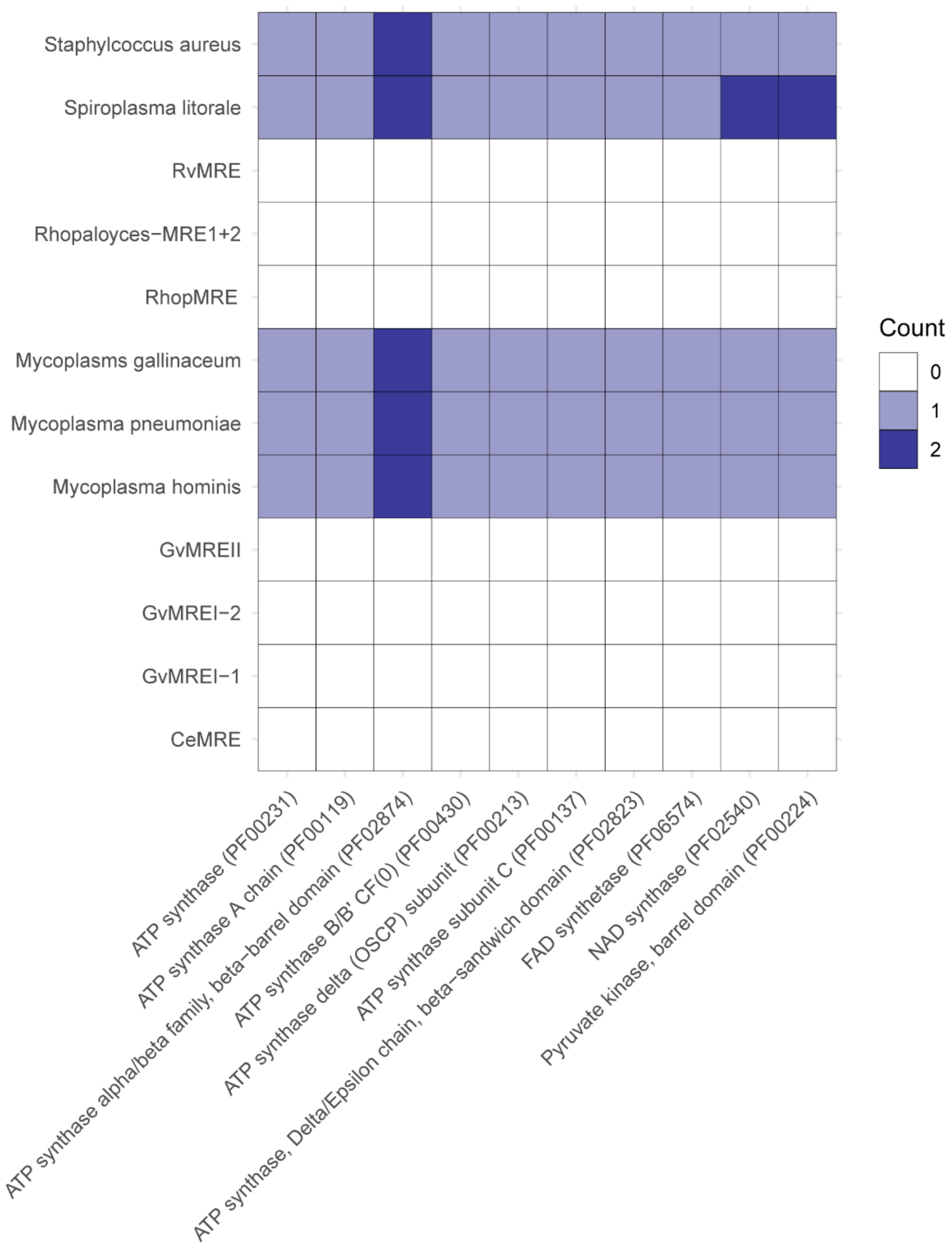


Figure B1. Heat map showing occupancy of a selection of PFAM domains involved in cellular respiration in MRE and non-MRE predicted proteomes. Annotations conducted with *interproscan*.

Appendix C Supplementary Table and Figures for Chapter 4

Table C1. Summary of taxa included in phylogenomic analyses with accompanying assembly statistics, numbers of annotated genes, and *BUSCO* (protein mode, *fungi_odb10*) statistics. Rows with values for Isolate_ID indicate genomes that were sequenced as part of this study. Missing assembly statistics indicate taxa where genes were not annotated as part of this study; that is, predicted proteomes were downloaded and used as is from NCBI GenBank.

SPECIES.TREE.LABEL	Isolate_ID	Assembly Length	GC Content	N50	L50	Proteins	BUSCO Completeness	BUSCO Duplication
Allomyces_arbuscula_Burma_1F.LCG	Burma_1_F	26482826	0.617947798	317448	26	8889	0.868073879	0.03343465
Allomyces_javanicus_California_12.LCG	California_1	29927322	0.623939088	104684	85	9981	0.817941953	0.101612903
Allomyces_macrogynus_ATCC_38327		57060573	0.568014292	1114524	17	19446	0.86939314	0.729893778
Anaeromyces_robustus_v1.0		71685009	0.162882633	141798	158	12832	0.828496042	0.036624204
Armillaria_ostoyae_C18_9		60106801	0.483284396	2283935	9	22299	0.972295515	0.010854817
Arthrobotrys_oligospora_ATCC_24927		40072829	0.443335483	2037373	8	11479	0.981530343	0.004032258
Backusella_circina_FSU_941.Bacci1.v1		48648288	0.333665411	185938	78	17039	0.947229551	0.215877437
Basidiobolus_heterosporus_B8920.N168.v1		44053790	0.397463964	2204	3776	8992	0.629287599	0.111111111
Basidiobolus_meristosporus_CBS_931.73		89489060	0.427634518	106019	272	16110	0.973614776	0.284552846
Batrachochytrium_dendrobatidis_JAM81_v1.0		24315081	0.389550913	1484462	6	8732	0.897097625	0.020588235
Batrachochytrium_dendrobatidis_JEL423	JEL0423	23897668	0.387428179	1707251	5	8819	0.802110818	0.024671053
Batrachochytrium_salamandrivorans_BS		32636440	0.416632451	10956	814	10135	0.845646438	0.014040562
Bifiguratus_adelaidae_AZ0501		19478880	0.476637004	102371	44	5719	0.709762533	0.018587361
Blastocladiella_britannica_v1.0		19075314	0.58605992	23470	254	9431	0.722955145	0.009124088
Blyttiomyces_helicus_single-cell_v1.0		46472760	0.537513438	6669	1974	12167	0.386543536	0.017064846
Blyttiomyces_sp._JEL0837.LCG	JEL0837	46468325	0.412203496	12527	1047	13891	0.704485488	0.016853933
Boothiomyces_macroporosum_PLAUS21.LCG	PLAUS21	15700521	0.385833693	109397	44	7605	0.692612137	0.041904762
Boothiomyces_sp._JEL0838.LCG	JEL0838	12849375	0.381256754	116125	34	6264	0.588390501	0.062780269
Boothiomyces_sp._JEL0866.LCG	JEL0866	14241334	0.377537455	114768	36	7120	0.64116095	0.061728395
Borealophlyctis_nickersoniae_WJD170.LCG	WJD170	38361113	0.50632178	16159	668	11278	0.828496042	0.036624204
Calcarisporiella_thermophila_CBS279.70.v1						11703	0.973614776	0.075880759
Candida_arabinofermentans_NRRL_YB-2248		13233932	0.34387263	701640	6	5826	0.935356201	0.002820874
Capsaspora_owczarzaki_ATCC_30864.v2		27967784	0.527986665	1617775	6	8621	0.726912929	0.01814882
Catenaria_anguillulae_PL171_v2.0	PL171	41337528	0.560449043	217825	41	12804	0.742744063	0.053285968
Caulochytrium_protostelioides_ATCC_52028_v1.0		21796879	0.664080624	92490	68	6168	0.676781003	0.007797271
Chytridium_lagenaria_Arg66_v1.0	ARG066	42380874	0.45331236	216171	61	14275	0.666226913	0.033663366
Chytriomyces_confervae_CBS_675.73		35973405	0.477797556	44341	205	10712	0.836411609	0.178233438
Chytriomyces_hyalinus_ARG085.LCG	ARG085	29777168	0.478981917	50756	175	11560	0.803430079	0.114942529
Chytriomyces_hyalinus_ARG121.LCG	ARG121	29515939	0.478794186	52956	159	11575	0.836411609	0.124605678
Chytriomyces_hyalinus_JEL0176.LCG	JEL0176	29545400	0.479372593	28716	303	11511	0.798153034	0.100826446
Chytriomyces_hyalinus_JEL0345.LCG	JEL0345	28404701	0.476569283	41827	196	11171	0.786279683	0.092281879
Chytriomyces_hyalinus_JEL632_v1.0	JEL0632	38103441	0.478851949	312457	37	15516	0.885224274	0.244411326
Chytriomyces_sp._MP_71_v1.0		36383760	0.490719458	45725	234	16054	0.835092348	0.109004739
Cladochytrium_replicatum_JEL714_v1.0	JEL0714	50665862	0.479122668	394289	42	16307	0.920844327	0.41260745
Cladochytrium_tenue_CCIBt4013.v0.LCG	GHJ CCIBt 40	48957730	0.564285027	5135	2739	15890	0.439313984	0.165165165
Clydaea_vesicula_JEL0476.LCG	JEL0476	24650228	0.28040978	24845	277	8647	0.733509235	0.026978417
Coelomomyces_lativittatus_CIRM-AVA-1-Meiospore.LCG		21922045	0.322942089	6695	990	7631	0.486807388	0.008130081
Coemansia_reversa_NRRL_1564		21837878	0.427022168	347177	21	7338	0.872031662	0.009077156
Conidiobolus_coronatus_NRRL_28638		39903661	0.219761415	102411	113	10568	0.798153034	0.016528926
Conidiobolus_thromboides_FSU_785.Conth1.v1		24635576	0.265406987	90842	82	8867	0.835092348	0.007898894
Coprinopsis_cinerea_okayama7_130						13355	0.964379947	0.004103967
Dinochytrium_kinnereticum_KLL_TL_06062013.LCG	KLL_TL_0606	31510542	0.476494946	39098	220	10332	0.699208443	0.013207547
Drosophila_melanogaster.v6						13994	0.660949868	0.035928144
Entophlyctis_helioformis_JEL805_v1.0	JEL0805	30915201	0.600270656	103262	89	10118	0.856200528	0.043143297
Entophlyctis_luteolus_JEL0120.LCG	JEL0120	25900143	0.481682553	16966	435	9099	0.662269129	0.099601594

Table C1 continued on next 2 pages…

SPECIES.TREE.LABEL	Isolate_ID	Assembly Length	GC Content	N50	L50	Proteins	BUSCO Completeness	BUSCO Duplication
Entophlyctis_luteolus_JEL0129.LCG	JEL0129	27694625	0.48028594	68043	122	9698	0.733509235	0.111510791
Entophlyctis_sp._JEL0112.LCG	JEL0112	26962970	0.481095295	48027	154	9626	0.726912929	0.110707804
Fimicolochytrium_jonesii_JEL569_v1.0	JEL0569	30619153	0.538454868	799348	12	10067	0.940633245	0.03085554
Fonticula_alba_ATCC_38817.v2		31296464	0.551258538	2529562	5	5901	0.432717678	0.012195122
Gaertneriomyces_semiglobifer_Barr_43_v1.0	Barr043	20918219	0.495286716	575746	9	8714	0.899736148	0.039589443
Gaertneriomyces_sp._JEL0708.LCG	JEL0708	20557201	0.494743229	558018	12	7636	0.915567282	0.03314121
Geranomyces_michiganensis_JEL0563.LCG	JEL0563	23465858	0.531615635	269724	29	8389	0.89182058	0.022189349
Geranomyces_variabilis_JEL0379.LCG	JEL0379	23394538	0.544132353	225234	31	8699	0.927440633	0.028449502
Geranomyces_variabilis_JEL0389.LCG	JEL0389	24136728	0.542660753	475954	17	8866	0.941952507	0.023809524
Geranomyces_variabilis_JEL0566.LCG	JEL0566	23613044	0.543517092	397283	20	8803	0.936675462	0.025352113
Geranomyces_variabilis_JEL0567.LCG	JEL0567	23750802	0.543271339	473289	16	9029	0.934036939	0.026836158
Geranomyces_variabilis_JEL559_v1.0	JEL0559	23695997	0.542939215	231969	30	9411	0.935356201	0.026798307
Gigaspora_rosea_DAOM_194757		567860885	0.265219232	232087	734	31243	0.936675462	0.029577465
Globomyces_pollinis-pini_Arg68_v1.0	ARG068	21646612	0.349838718	50517	125	11537	0.897097625	0.010294118
Gonapodya_prolifera_v1.0		48794828	0.518232342	347324	42	13902	0.843007916	0.03286385
Gonapodya_sp._JEL0774.LCG	JEL0774	35595249	0.510834971	71677	113	10034	0.573878628	0.013793103
Gorgonomyces_haynaldii_MP57_v1.0	MP57	13983420	0.457747604	597596	8	7898	0.860158311	0.013803681
Hesseltinella_vesiculosa_NRRL_3301		27216191	0.462540552	571097	14	11139	0.968337731	0.051771117
Homolaphlyctis_polyrhiza_JEL142_v1.0	JEL0142	21324754	0.428583326	10789	577	7123	0.639841689	0.01443299
Hortaea_werneckii_EXF-2000		49942992	0.534781817	153735	100	15620	0.974934037	0.857916103
Hyaloraphidium_curvatum_SAG235-1_v1.0	JEL0383	31926619	0.651156391	722379	18	15197	0.852242744	0.007739938
Irineochytrium_annulatum_JEL0729.LCG	JEL0729	36841590	0.541227455	20474	470	11905	0.687335092	0.032629559
Kappamyces_sp._JEL0680.LCG	JEL0680	13243986	0.522749798	7589	433	7477	0.410290237	0.006430868
Kappamyces_sp._JEL0829.LCG	JEL0829	11696060	0.524292625	100276	39	5512	0.565963061	0.023310023
Lichtheimia_corymbifera_FSU_9682		33531723	0.401212339	367562	25	12282	0.873350923	0.113293051
Linderina_pennispora_ATCC_12442		26202545	0.541441642	908848	9	9350	0.80474934	0.029508197
Lobosporangium_transversale_NRRL_3116		42768949	0.415735865	672590	22	11818	0.964379947	0.082079343
Lobulomyces_angularis_JEL0522.LCG	JEL0522	24944113	0.279083886	46169	165	9112	0.850923483	0.023255814
Martensiomyces_pterosporus_CBS_209.56.Marpt1.v1		19815802	0.543248716	117925	51	8435	0.903693931	0.011678832
Mitosporidium_daphniae_UGP3		5635072	0.429960434	32179	50	3322	0.27176781	0.029126214
Mixia_osmundae		13634488	0.551991171	1194905	5	6858	0.94591029	0.008368201
Monosiga_brevicolis_MX1.v1						9203	0.518469657	0.007633588
Mortierella_elongata_AG-77		49851634	0.479057036	517143	31	14959	0.985488127	0.107095047
Mucor_circinelloides_f._circinelloides_1006PhL		36348485	0.370830889	140649	82	12227	0.986807388	0.124331551
Neocallimastix_californiae_G1_v1.0		193032486	0.181670784	443414	134	20290	0.854881266	0.430555556
Neurospora_crassa_OR74A						9757	1	0
Nowakowskiella_sp._JEL0078.LCG	JEL0078	33052143	0.355192854	2655	3774	11645	0.518469657	0.091603053
Nowakowskiella_sp._JEL0407.LCG	JEL0407	22617538	0.388697214	52255	138	8328	0.662269129	0.057768924
Obelidium_mucronatum_JEL802_v1.0	JEL0802	49458483	0.448617743	189008	68	15468	0.852242744	0.078947368
Olpidium_bornovanus_UCB_F19785.Olpbor1		38674623	0.56903024	2083	6356	8477	0.158311346	0.05
Olpidium_sp._PSC023		16557855	0.489697126	18523	260	7901	0.715039578	0.009225092
Paramicrosporidium_saccamoebae_KSL3						3766	0.463060686	0.005698006
Paraphelidium_tribonemae_X-108.Trinity		53288649	0.51222432	1343	11381	42481	0.824538259	0.5888
Paraphysoderma_sedebokerense_JEL821_v1.0	JEL0821	27876074	0.412220638	239846	31	10859	0.866754617	0.068493151
Pecoramyces_ruminatium_C1A		100954185	0.169985989	3373	10167	18936	0.393139842	0.077181208
Phlyctochytrium_bullatum_JEL0754.LCG	JEL0754	40192336	0.539482104	33519	362	11036	0.497361478	0.021220159
Phlyctochytrium_planicorne_JEL0388.LCG	JEL0388	30010325	0.472941496	79884	124	10408	0.721635884	0.036563071
Phycomyces_blakesleeanus_NRRL_1555		53939167	0.354032738	1515579	11	16542	0.939313984	0.092696629
Physocladia_obscura_JEL0513.LCG	JEL0513	44089900	0.388477248	11689	1143	12792	0.692612137	0.102857143
Piptocephalis_cylindrospora_RSA_2659		10748482	0.512047748	11086	282	4301	0.443271768	0
Piromyces_finnis_v3.0		56455805	0.211785945	749539	25	10992	0.865435356	0.035060976
Piromyces_sp._E2_v1.0		71019055	0.131451721	144455	143	14648	0.477572559	0.022099448
Podochytrium_sp._JEL0797.LCG	JEL0797	32060568	0.507521139	20738	411	12043	0.766490765	0.092943201
Polychytrium_aggregatum_JEL109_v1.0	JEL0109	64917104	0.573040011	389128	41	10690	0.949868074	0.029166667
Powellomyces_hirtus_BR81_v1.0	Barr081	29428253	0.514962985	1016081	10	9359	0.949868074	0.029166667
Powellomyces_hirtus_CBS_809.83		26238698	0.513727892	157542	47	6536	0.936675462	0.021126761
Puccinia_graminis_f._sp._tritici_CRL_75-36-700-3		88724376	0.398680471	964966	30	15800	0.90237467	0.092105263
Quaeritorhiza_haematococci_JEL0916.LCG	JEL0916	48216321	0.498862097	11770	1101	13723	0.68469657	0.013487476
Ramicandelaber_brevisporus_CBS_109374.Rambr1.v1		25531049	0.422553966	41156	190	9281	0.711081794	0.055658627
Rhizoclosmatium_globosum_JEL800_v1.0	JEL0800	57018351	0.448997534	292246	51	15991	0.819261214	0.05958132
Rhizoclosmatium_hyalinum_JEL0917.LCG	JEL0917	23450088	0.445341826	2546	2086	10776	0.343007916	0.023076923
Rhizoclosmatium_sp._JEL0117.LCG	JEL0117	30544921	0.448869486	40205	220	11857	0.837730871	0.083464567
Rhizophagus_irregularis_DAOM_181602		149750837	0.278833447	2308146	23	26143	0.953825858	0.024896266
Rhizophlyctis_rosea_JEL0318.LCG	JEL0318	38859322	0.490091155	21549	498	11459	0.759894459	0.010416667
Rhizophlyctis_rosea_JEL0764.LCG	JEL0764	48149512	0.501009792	104982	135	12571	0.781002639	0.013513514
Rhizopus_delemar_RA_99-880		46148878	0.349539029	3104119	6	17459	0.808707124	0.212071778
Rhizopus_microsporus_var._microsporus_ATCC_52814		24950816	0.373366947	105542	71	11496	0.927440633	0.06401138
Rozella_allomycis_CSF55_v1.0		13461086	0.348153782	7173	524	6350	0.625329815	0.002109705
Rozella_multimorpha		13558553	0.396279234	5938	703	7336	0.443271768	0.00297619
Rozella_sp._PSC023		14727194	0.401597209	17387	235	7708	0.598944591	0.00660793
Saccharomyces_cerevisiae_S288C						6008	0.965699208	0.020491803
Saccharomycopsis_capsularis_NRRL_Y-17638		17823225	0.442349799	289111	20	6736	0.934036939	0.012711864
Saksenaea_vasiformis_B4078.G233.v1		42502389	0.426548729	79983	143	9656	0.936675462	0.087323944
Schizosaccharomyces_pombe_972h-		12591251	0.360491821	4539804	2	5130	0.961741425	0.027434842
Siphonaria_sp._JEL0065.LCG	JEL0065	37595823	0.437062516	21877	522	12648	0.703166227	0.178236398
Spizellomyces_punctatus_DAOM_BR117		24131112	0.471598532	1465700	7	9424	0.95646438	0.03862069
Spizellomyces_sp._palustris_CBS_455.65		22937368	0.477960375	219277	32	8518	0.94591029	0.012552301
Syncephalastrum_racemosum_NRRL_2496.Synrac1.v1		30745403	0.467722085	2374188	5	11124	0.964379947	0.073871409

SPECIES.TREE.LABEL	Isolate_ID	Assembly Length	GC Content	N50	L50	Proteins	BUSCO Completeness	BUSCO Duplication
Syncephalis_fuscata_S228.Synfus1.v1		29358178	0.366449682	474633	20	8846	0.887862797	0.016344725
Synchytrium_endobioticum_MB42		21483073	0.467534789	44081	153	8031	0.841688654	0.028213166
Synchytrium_microbalum_JEL517	JEL0517	26244175	0.436715195	518601	16	6304	0.90237467	0.010233918
Terramyces_sp._JEL0728.LCG	JEL0728	15585454	0.393807585	101760	45	7848	0.724274406	0.063752277
Thoreauomyces_humboldtii_JEL0095.LCG	JEL0095	26328562	0.548837115	23424	325	9403	0.846965699	0.048286604
Triparticalcar_arcticum_BR59_v1.0	Barr059	31548172	0.484569027	855893	12	10963	0.924802111	0.02853067
Umbelopsis_ramanniana_AG_#.Umbra1.v1		23077072	0.43126849	294116	26	9931	0.948548813	0.041724618
Unknown_Chytridiales_sp._JEL0842.LCG	JEL0842	26948520	0.477582554	72085	105	8731	0.77176781	0.017094017
Unknown_Rhizophydiales_sp._JEL0801.LCG	JEL0801	16104292	0.352042921	42882	105	6698	0.414248021	0.044585987
Unknown_unknown_JEL0888.LCG	JEL0888	23308959	0.628337241	11652	575	8331	0.774406332	0.010221465
Ustilago_maydis_521		19664356	0.539692528	884984	7	6764	0.978891821	0.001347709
Yarrow_lipolytica_CLIB122						6471	0.973614776	0
Zoophthora_radicans_ATCC_208865		655199646	0.318707054	544305	307	14479	0.870712401	0.065151515
Zopfochytrium_polystomum_WB228_v1.0		81192468	0.532549128	222082	105	16599	0.882585752	0.07922272

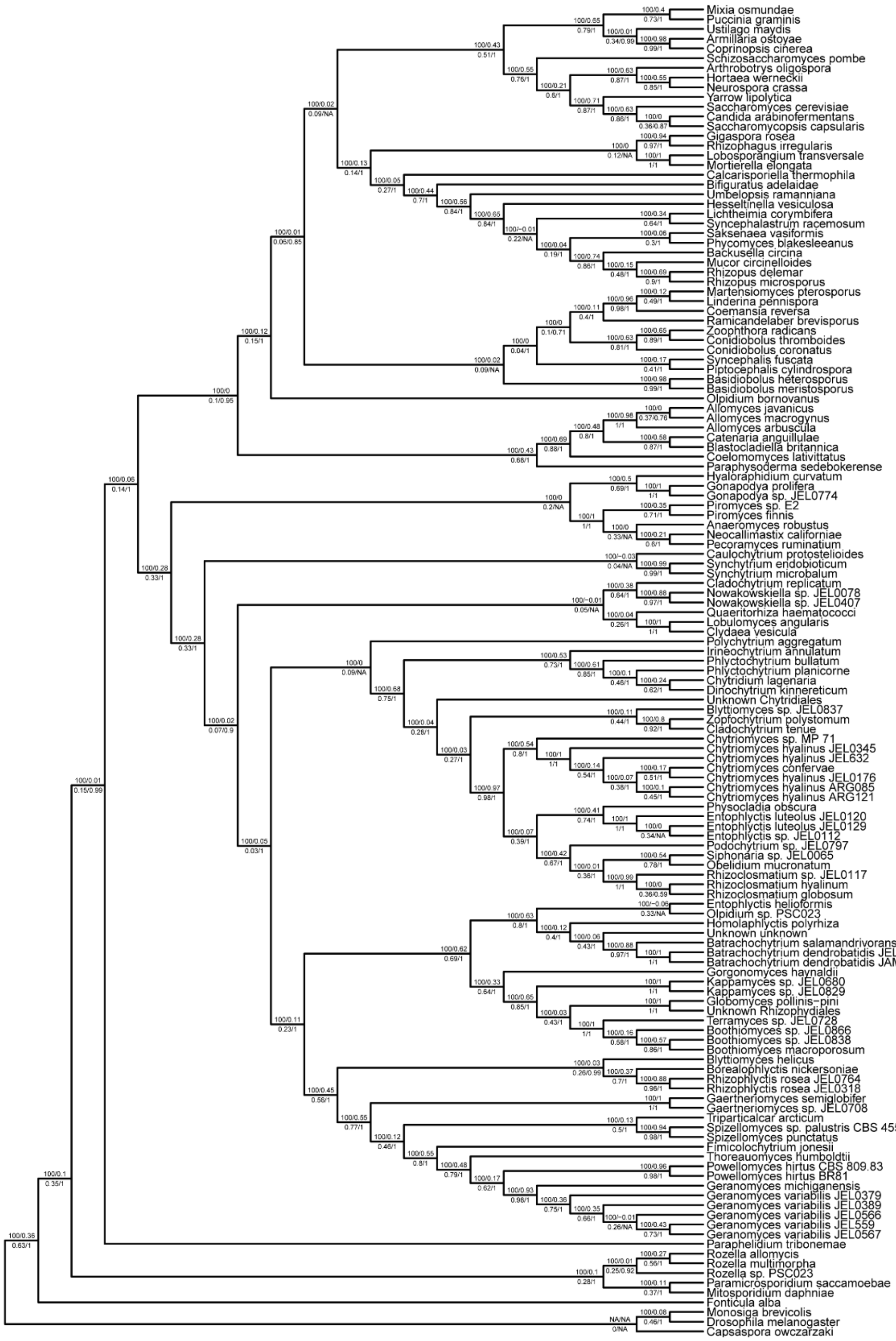


Figure C1. Concatenated ML tree identical to that shown in Figure 4.1, except will all nodes expanded, and all support values listed on nodes. Support values listed in the following format. Above Node: <bootstrap> / <qpic>, Below Node: <gCF> / <ASTRAL LPP (where applicable)>

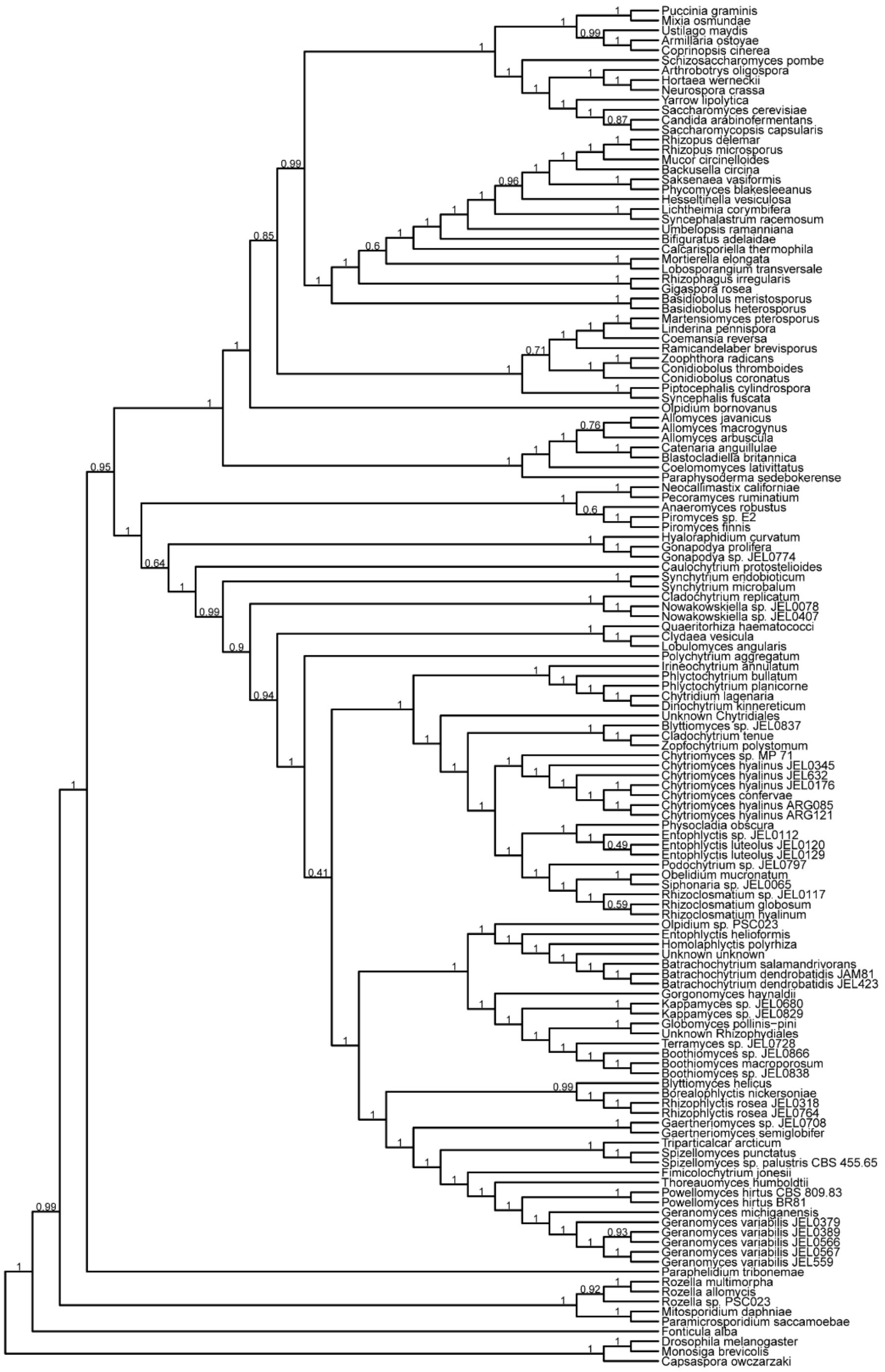


Figure C2. ASTRAL coalescent tree based on 487 gene trees of each of 487 markers used to compute concatenated ML tree in Figure 4.1. ASTRAL local posterior probabilities shown on nodes.

Appendix D Supplementary Table for Chapter 5

Table D1. Genomes included in the phylogenomic reconstructions conducted in this study. Genomes generated in this study are highlighted in yellow. Second column indicates either the JGI Genome Portal where publicly data is available or the publication within which genome data was generated.

Species	JGI Genome Portal ID or Citation	Species	JGI Genome Portal ID or Citation
Allomyces macrogynys	Allma1	Martensiomyces pterosporus	Marpt1
Amanita muscaria	Amamu1	Mitosporidium daphineae	Mdap
Anaeromyces sp.	Anasp1	Mixia osmundae	Mixos1
Armillaria gallica	Armga1	Monosiga brevicolus	Monbr1
Neozygites floridana ARSEF 5376	This Study	Morchella conica	Morco1
Arthrobotrys oligospora	Artol1	Mortierella elongata	Morel2
Aspergillus flavus	Aspfl1	Malassezia restricta	MRES
Acaulopage tetraceros	Davis et al. 2019	Mucor circinelloides	Mucci2
Basidiobolus meristosporus	Basme2finSC	Neozygites sp. Neo_30	This Study
Batrachochytrirum dendrobaditis	BDET	Neozygites sp. Neo_Co_SC	This Study
Blastocladiella britannica	Blabri1	Neozygites parvispora ARSEF 5620	This Study
Catenaria anguillulae	Catan2	Neocallimastix californiae	Neosp1
Caulochytrium protostelioides	Caupr1	Orpinomyces sp.	Orpsp1_1
Chytriomyces sp. MP71	Chytri1	Phycomyces blakeseeanus	Phybla1
Cocholomyces odontomsperma	Davis et al. 2019	Piptocephalus cylindrospora	Pipcy3_1
Coemansia reverse	Coere1	Piromyces finis	Pirfi3
Cokeromyces recurvatus	Cokrec1	Pleurotus ostreatus	PleosPC15_2
Conidiobolus coronatus	Conco1	Ramicandelaber brevisporus	Rambr1
Conidiobolus thromboides	Conth1	Rhizopus oryzae	Rhior3
Coprinus cinerea	Copci_AmutBmut1	Rozella allomycis	Rozal1_1
Cordyceps militaris	Cormi1	Saccharomyces ceriviseae	SACCE
Dichotomocladium elegans	Dicele1	Stylopage hadra	Davis et al. 2019
Dimargaris cristalligena	DimcrSC1	Smittium culicis	SmicuMNP_2
Drosophila melanogaster	Dmel	Smittium mucronatum	Smimuc2
Entophlyctis helioformis	Enthel1	Spizellomyces punctatus	Spipu1
Fomitopsis pinicola	Fompi3	Syncephalis fuscata	Synfus1
Fusarium oxysporum	Fusox2	Syncephalis plumigaleata	Synplu1
Gaertneriomyces semiglobifer	Gaesem1	Syncephalis pseudoplumigaleata	Synps1
Gonapodya prolifera	Ganpr1	Syncephalastrum racemosum	Synrac1
Rhizophagus irregularis	Gloin1	Thamnocephalis sphaerospora	Thasp1
Globomyces pollinis-pini	Glopol1	Umbelopsis isabellina	Umbisa1
Hesseltinella vesiculosa	Hesve2finisherSC	Ustilago maydis	Ustma2_2
Hypoxylon sp.	HypEC38_3	Wallemia mellicola	Walse1
Hypholoma sublateritium	Hypsu1	Zancudomyces culisetae	Zancul2
Lacaria bicolor	Lacbi2	Zoophagus insidians	Davis et al. 2019
Linderina pennispora	Linpe1	Zoophthora radicans	Zoorad1
Lobosporangium transversale	Lobtra1	Zoopage sp.	Davis et al. 2019