AN ABSTRACT OF THE DISSERTATION OF

Zachary Stephen Longiaru Foster for the degree of Doctor of Philosophy in Molecular and Cellular Biology presented on June 18, 2020.

Title: Development of Computational, Visualization, and Molecular Tools for Fungal and Oomycete Community Ecology

Abstract approved: _____

Niklaus J. Grünwald

Oomycetes are an important group of organisms with a variety of ecological roles similar to fungi. Although many are well-studied plant pathogens known for their devastating effects on agricultural systems, most are little-studied saprobes and parasites of plants and animals in nearly every ecosystem on earth. The advent of affordable and high-throughput sequencing technologies have resulted in new opportunities for the study of microbial communities, including oomycetes, as well as new challenges associated with analyzing and visualizing the vast amount of data produced. This thesis describes new tools developed for the analysis and visualization of microbial communities, with an emphasis on oomycetes, and studies into the communities of oomycetes and fungi associated with *Rhododendron*.

The widespread adoption of high-throughput molecular community ecology methods is making large data sets classified by taxonomic information common, but addi-

tional tools to analyze and visualize these data are needed. Taxonomic classifications are hierarchical, making them much more difficult to analyze compared to typical tabular data. There are many R packages that use taxonomic data to varying degrees but there is currently no cross-package standard for how this information is encoded and manipulated. We developed the R package `taxa` to provide a robust and flexible solution to storing and manipulating taxonomic data in R and any application-specific information associated with it. It is meant to be a foundation for other packages to build on, so that diverse packages dealing with taxonomic information can be integrated seamlessly. One package that is built on top of `taxa` is `metacoder`. `Metacoder` is an R package for plotting and manipulating data classified by a taxonomy, like the abundance data associated with metabarcoding. Its primary feature is the novel tree-based visualization called "heat trees" that is used to depict data for every taxon in a taxonomy using color and size. Heat-trees provide a more informative alternative to pie charts or stacked bar charts for visualizing communities. `Metacoder` also provides various functions to do common tasks in microbiome research with data stored in the `taxmap` format supplied by the `taxa` package. Both of these packages are open source, version controlled, have unit tests that help detect bugs, and include extensive documentation.

Although metabarcoding methods for fungal and bacterial communities are well-developed at this point, no standard and reliable method for oomycete metabarcoding exists. Every currently proposed method for oomycete metabarcoding has at least one flaw; some produce too long an amplicon for Illumina sequencers, some target only a subset of oomycete diversity, some have unacceptable levels of non-target amplifi-

cation, and some have technical difficulties that make the PCR reactions unreliable. We developed a new method for oomycete metabarcoding targeting the *rps10* gene and an associated reference database. Compared to one of the more popular methods currently being used, our method has better taxonomic resolution, less non-target amplification, and a more reliable PCR reaction. A reference database of *rps10* sequences for many genera of oomycetes was developed for use in assigning taxonomic classifications to metabarcoding data. Finally, a website was created to host the database that supports searching the database and conducting BLAST searches.

Rhododendron is a major ornamental crop in the Pacific Northwest and is known to host both mycorrhizal symbionts and plant pathogens such as *Phytophthora ramorum*. The fungal and oomycete microbiome in the rhizosphere of rhododendrons from Oregon nurseries was sequenced and differences among cultivars, growth conditions, and nurseries were analyzed. Few oomycetes were found, but this might have been partially due to limitations of the metabarcoding method used. Fungal species found were mostly saprobes and mutualists. Nurseries that grew plants in containers and in-field had a significantly higher diversity of fungi than those that only grew plants in containers. Microbiome composition differed significantly among growth conditions and nurseries, but not among cultivars. This body of work provides novel insights into oomycete communities and novel tools for molecular community ecology that might be of broader interest.

Development of Computational, Visualization, and Molecular Tools for Fungal and Oomycete Community Ecology

by

Zachary Stephen Longiaru Foster

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented June 18, 2020
Commencement June 2021

Doctor of Philosophy dissertation of Zachary Stephen Longiaru Foster presented on June 18, 2020.

APPROVED:

Major Professor, representing Molecular and Cellular Biology

Director of the Molecular and Cellular Biology Program

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Zachary Stephen Longiaru Foster, Author

ACKNOWLEDGEMENTS

I would like to acknowledge many people who have helped make this work possible. I would like to thank my adviser Niklaus Grünwald, who gave much good advice for starting a successful research career and was instrumental in the development of many good ideas. Meg Larsen, Valerie Fieland, and Caroline Press provided much guidance and advice concerning molecular biology that would have been hard to replace. The advice of and ready helpfulness in matters of computational biology of Brian Knaus was similarly invaluable. I also had many productive and amusing discussions with my fellow students Zhian Kamvar, Javier Tabima, Nick Carleson, and Shankar Shakya. I would like to thank the many coauthors and collaborators who worked with me on these projects, including Thomas Sharpton, Scott Chamberlain, Frank Martin, Andrew Jones, Brett Tyler, Jerry Weiland, Carolyn Scagel, and Felipe Albornoz. The advice of my committee, which included Niklaus Grünwald, Aaron Liston, Joseph Spatafora, Jennifer Parke, and Kari Van Zee, was very useful in preparing this work. I would also like to thank the BPP and MCB departments for creating a productive and welcoming environment to work in. Most of all, I would like to thank my parents who always encouraged my interest in science and have helped every step of the way.

CONTRIBUTION OF AUTHORS

The following people contributed to this dissertation in addition to Niklaus J. Grünwald:

Chapter 2: Taxa: An R package implementing data standards and methods for taxonomic data

Scott Chamberlain assisted with conceptualization, methodology, project administration, software, and writing.

Chapter 3: Metacoder: An R package for visualization and manipulation of community taxonomic diversity data

Thomas J. Sharpton assited with conceptualization, methodology, project administration, and writing.

Chapter 4: The Composition of the Fungal and Oomycete Microbiome of Rhododendron Roots Under Varying Growth Conditions, Nurseries, and Cultivars

Jerry E. Weiland assisted with conceptualization, methodology, project administration, and writing. Carolyn F. Scagel assisted with conceptualization, methodology, project administration, and writing.

Chapter 5: Rps10: a new barcode for high throughput amplicon sequencing of Oomycete communities

Felipe E. Albornoz assited with conceptualization, methodology, lab work, data analysis, and writing. Frank N. Martin assited with conceptualization, methodology, project administration, and writing. Valerie J. Fieland assited with methodology, lab work, data analysis, and writing. Meredith M. Larsen assited with methodology, lab work, data analysis, and writing. Andrew F. Jones assited with conceptualization, methodology, and project administration. Brett M. Tyler assited with conceptualization, methodology, and project administration. Hai D. T. Nguyen, Carolyn Riddle, and Treena Burgess conributed sequences to the database.

TABLE OF CONTENTS

Page

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF TABLES

Chapter 1: Introduction

1.1 Oomycete ecology literature review

1.1.1 Introduction to oomycetes

Oomycetes are members of the Stramenopiles-Alveolata-Rhizaria (SAR) supergroup and are closely related to diatoms (McCarthy & Fitzpatrick, 2017). They inhabit nearly every environment on earth (Davis, 2016) and are important components of many ecosystems. They are characterized by sexual survival structures called oospores, the presence of cellulose in their cell walls, and by having biflagellate zoospores. Many oomycetes have coenocytic hyphae and others, particularly marine parasites, are holocarpic, meaning the entire organism is converted to reproductive prologues at the end of its life cycle. Oomycetes have ecological roles similar to fungi, ranging from saprophytic to obligate parasitic lifestyles, but with a greater proportion of pathogenic and parasitic species. They used to be placed in the kingdom fungi, but recent phylogenetic and ultrastructural studies have confirmed, as has long been suspected, that the two groups are not closely related and morphological similarities are due to convergent evolution.

Oomycetes are thought to have evolved from marine parasites and transitioned to land with their hosts. Many of the most basal clades of oomycetes are holocarpic parasites of marine algae, such as *Eurychasma* (Gachon, Strittmatter, Müller, Kleinteich,

& Küpper, 2009). The oldest fossils that are unambiguously oomcyetes are 400 million years old, but their origins are thought to be much earlier (Hansen, Reeser, & Sutton, 2012). Phylogenetic clock estimates place the divergence of oomycetes and diatoms around 500 million years ago and the divergence among major oomycete lineages around 100 million years ago (Matari & Blair, 2014). Although the oomycetes, as the group is currently delineated, seem to be monophyletic, there are many well-known oomycete families and genera that recent phylogenetics studies have demonstrated to be paraphyletic and the ooymcete taxonomy is likely to change much in the near future. Traditional oomycete taxonomy has grouped them into two "galaxies": The "perosporaleans", which include many well-known plant pathogens, and the "saprolegnians", which include animal pathogens and many poorly-characterized saprobes (Spring et al., 2018). Modern phylogenetic studies indicate that the order of divergence of major oomycete groups is as follows: the basal marine parasites, such as *Eurychasma*, the Saprolegniales (aquatic parasites of animals), the Albuginales (the white rust plant pathogens), and a clade containing *Pythium*, *Phytophthora* and the downy mildews, which are mostly plant pathogens (McCarthy & Fitzpatrick, 2017).

There are many excellent reviews on specific economically important oomycete pathogens and reviews on the phylogeny and taxonomy of oomycetes (Kamoun et al., 2015; Phillips, Anderson, Robertson, Secombes, & Van West, 2008; Thines, 2014; Tyler, 2007). Pathogenic species such as *Phytophthora infestans* (Akino, Takemoto, & Hosaka, 2014; Andrivon, 1996; Fry et al., 2015), the cause of the great potato famine, and *Plasmopara viticola* (Gessler, Pertot, & Perazzolli, 2011), the cause of grape downy mildew, have been studied extensively. There is also much literature,

some dating back to the mid 1800's, discussing the classification of oomycetes based on morphological and, more recently, genetic differences (Dick, 1969; McCarthy & Fitzpatrick, 2017; Sparrow, 1976). However, relatively little is known about the ecology of the vast majority of oomycetes that are minor pathogens in natural systems, aquatic saprobes, or marine parasites, although in some cases these organisms are the most influential members of their ecological niche (Leafio, Jones, & Vrijmoed, 2000). Therefore, this review will focus primarily on what relatively little is known about the ecology, diversity, and distribution of oomcyetes as a whole and how they relate to human activities.

1.1.2 Reproduction

Reproduction is as varied and complex as other traits of the oomycetes, but general similarities exist, particularly the formation of oospores, the sexual structure that give oomycetes their name, and flagellated zoospores, the water-dependent dispersal agent that most differentiate oomycetes morphologically from fungi. One or more oospores, which are thick-walled sexual structures, are produced inside oogonia when fertilized by a "male" hypha called an antheridium. Some oomycetes are homothallic, meaning that a pure culture can sexually recombine with itself and produce oospores, whereas other are heterothallic, meaning that two different strains are needed for sexual reproduction. Oospores function as both a mechanism of sexual reproduction and as a resistant resting spore. Oospores can germinate into either a hypha, which might produce a sporangium, or form a vesicle from which zoospores are formed. Sporangia

are typically less resistant and might detach and function as dispersal agents in air or water, particularly in terrestrial pathogens (Jung et al., 2017). The contents of the sporangia are cleaved into many smaller propagules called zoospores. In some species the zoospores mature inside the sporangium and swim out through an openening at the tip, such as in *Phytophthora*, but in others, they mature in a protoplasmic mass that is ejected outside the sporangium, as is typical for *Pythium* species (Rocha et al., 2014). Zoospores are short-lived mobile propagules that use flagella to find suitable hosts or substrates. They usually have two flagella, a smooth "whiplash" flagellum that faces away from the direction of movement and a hairy "tinsellated" flagellum that faces towards the direction of movement. The presence of zoospores with these two types of flagella is the defining feature of the Heterokonts, a group that includes oomycetes, diatoms, golden algae, and brown algae. Once a zoospore finds a suitable substrate, or after a fixed amount of time, it encysts and forms a hyphal-like projection. In pathogens, this is a structure called an appressorium that penetrates the cell wall of the host using a combination of cell wall degrading enzymes and physical pressure. In some species, such as those in *Saprolegnia*, zoospores can encyst and create a second generation of zoospores, a phenomenon called "polyplanetism" (Van West, 2006). Many basal aquatic groups of oomycetes are holocarpic, meaning that the entire thallus, usually restricted to a single cell and more globular than hyphal in shape, is converted to reproductive prologues at the end of its life cycle.

1.1.3 Dispersal

Oomycetes vary greatly in their strategies and potential for dispersal and dormancy. Pathogenic and parasitic species often depend on hosts with annual life cycles, such as annual plants, and thus must tolerate periods when a host is not available (Spring et al., 2018) and even saprophytic species in soil must tolerate reduced activity in winter in temperate climates. Oomycetes have three general strategies for dispersal: zoospores, sporangia or oospores, and vertical transmission of pathogens via seed. Zoospores of *Olpidiopsis* and *Pythium* were found to be infective for up to 7 days (Klochkova, Shim, Hwang, & Kim, 2012; Martin & Loper, 1999) in moist conditions, although zoospores of other oomycetes such as *Myzocytiopsis*, encyst almost immediately (Glockling & Beakes, 2000). Zoospores are attracted to their preferred substrate (Leafio et al., 2000) or host (Tyler, 2007) by chemical signals and actively swim towards it, a process known as chemotaxis. In some oomycetes, particularly the plant pathogenic downy mildews and some *Phytophthora* species, entire sporangia or oospores will detach at maturity and are dispersed by air (Bock, Jeger, Fitt, & Sherington, 1997; Hansen et al., 2012) or water (Misra, Sharma, & Mishra, 2008). In addition, some parasitic species, particularly biotrophs like the downy mildews, are transmitted between generations via resting structures in seeds (Lava, Heller, & Spring, 2013; Lebeda & Cohen, 2011). Oospores are particularly resistant structures that allow oomycetes to tolerate adverse conditions for a long time. Depending on the species, oospores have been recorded to be viable for up to 13 years, although durations of 2 to 10 years are more common (Davis, 2016; Martin & Loper, 1999;

Sakr & others, 2014). Marine oomycetes tend to have less resistant structures since their environment is much more consistent than terrestrial habitats (Klochkova, Shin, Moon, Motomura, & Kim, 2016).

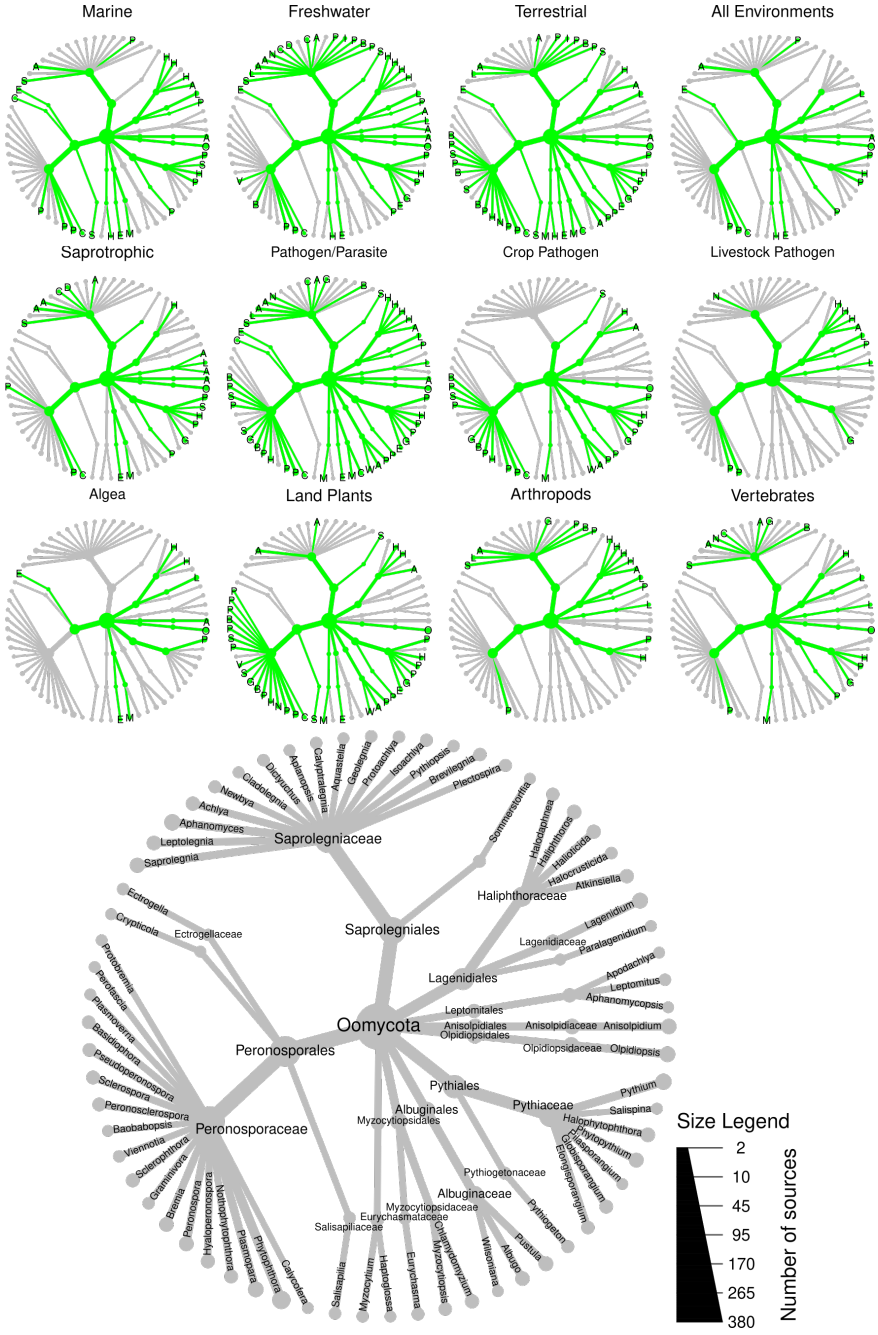

Figure 1.1: Taxonomic distribution of traits discussed in this review. Node size corresponds to the number of publications mentioning each taxon. Taxa in the smaller plots are green when the trait is present in at least one source. The larger tree functions as a legend for the smaller trees. Letters in smaller plots are the first letter of the taxon name. Only taxa mentioned in at least two publications are included.

1.1.4 Nutrition

Similar to fungi, nutritional modes range from free living saprobes to obligate parasites and representatives of oomycetes can be found that use every strategy for gaining nutrition except autotrophy, although their ancestors were likely at least partially photosynthetic (Figure 1.1). There are obligate parasites of terrestrial plants, aquatic algae, nematodes, diatoms, arthropods, and even other oomycetes. Others are necrotrophic pathogens, including some of the most damaging agricultural and aquacultural pests. Yet others, perhaps the majority, are opportunistic pathogens and saprobes in nearly every ecosystem.

1.1.4.1 Biotrophs and obligate pathogens

Well-known obligate pathogens and parasites include the white rusts, the downy mildews, and various holocarpic parasites of algae and nematodes. The two major groups of oomycete obligate pathogens are the white rusts, in the family Albuginaceae, and the downy mildews, which are distributed in three monophyletic groups containing at least 19 genera and 700 species (Jung et al., 2017; Thines & Choi, 2016). Obligate parasitism is thought to have evolved independently in the two groups (Ploch et al., 2011). Of the downy mildews, the genera *Peronospora* and *Pseudoperonospora* are the largest. Downy mildews are generally highly host-specific and usually can only reliably reproduce on a single host species, so much so that downy mildew taxonomy has traditionally been based largely on differences in host species. Thanks to molecular techniques, many downy mildews species that at first appeared to be

generalists, such as *Basidiophora entospora*, are now thought to be cryptic species complexes, with each species specific to a single host (Sökücü & Thines, 2014). The family Albuginaceae contains many white rust pathogens of Crassicaea, Asterales, Convolvulaceae, and grasses (Choi, Shin, & Thines, 2009; Thines, Telle, Choi, Tan, & Shivas, 2015). *Albugo candida* is a obligate pathogen of Brassicaceae, and is unusual in that it seems to be a generalist infecting more than 10 genera of Brassicaceae, whereas most oomycete obligate pathogens are highly-host specific (McMullan et al., 2015). In contrast to the downy mildews, the white rusts can reproduce without high levels of humidity since they produce spores below the epidermis of their hosts (Spring et al., 2018). Obligate parasites of marine and terrestrial nematodes and rotifers include the holocarpic genera *Chlamydomyzium* and *Haptoglossa*. *Haptoglossa* is particularly interesting due to its unique and complex "gun cell" that forcefully injects a needle-like structure into its host to began the infection process (Glockling & Beakes, 2000). Apart from a few rare exceptions, these obligate pathogens can not be collected in pure culture and many holocarpic species only infect a single cell at a time, so the diversity of these organisms is probably underestimated (Lebeda & Cohen, 2011; Spring et al., 2018).

1.1.4.2 Necrotrophs, hemibiotrophs, and opportunistic pathogens

The oomycetes also include many necrotrophic and hemibiotrophic pathogens, which weaken or kill their host and then live saprophytically on the remains while they reproduce or produce resting spores. Many of these are also weakly saprophytic

outside of a host (Jung et al., 2017). The best known examples of oomycetes with this lifestyle are species in the genus *Phytophthora* that primarily infect terrestrial plants (Jung et al., 2017). Unlike the obligate pathogens, necrotrophic oomycetes can often infect a wide range of hosts (Spring et al., 2018). For example, *Phytophthora palmivora* can infect at least 200 different plant species (Derevnina et al., 2016) and *Phytophthora ramorum* can infect at least 100 plant species (Grünwald, Garbelotto, Goss, Heungens, & Prospero, 2012). Many *Pythium* and some *Phytophthora* are known for killing seedlings as they emerge, a pathology known as "damping off" (Cuenca, 2016), or feeding on the fine roots of plants without necessarily killing them, exemplifying their intermediate nature between saprotroph and pathogen. Species of *Saprolegnia* are necrotrophic on various animals, including fish, crustaceans, and amphibians (Davis, 2016; Holt et al., 2018). Virulence of *Saprolegnia* infections varies greatly depending on the health of the host and the strain of *Saprolegnia*. For example, in some cases, *Saprolegnia parasitica* seems to be a mostly saprotrophic opportunistic pathogen of salmon, mostly affecting wounded or otherwise stressed fish, but can at other times cause death quickly on its own (Van West, 2006).

1.1.4.3 Saprobes

Perhaps the most under appreciated nutritional mode for oomycetes is saprotrophy, but in terms of diversity it could easily be the most common, considering the relatively little research that has been done on oomycete saprobes and saprobes in general (Blackwell, Letcher, & Powell, 2015). Most oomycetes in aquatic ecosystems

are thought to be saprotrophic (Jung et al., 2017; Nam & Choi, 2019). Species of the genera *Sapromyces* and *Pythiogeton* are freshwater saprobes (Blackwell et al., 2015; Jee, Ho, & Cho, 2000). In a survey of freshwater saprobes collected from decaying plant fragments, 119 oomycete species were found, with the species *Aphanomyces laevis*, *Saprolegnia litoralis*, and *Pythium rostratum* being the most common (Czeczuga, Mazalska, Godlewska, & Muszyńska, 2005). In mangrove swamps, members of the genus *Halophytophthora* are thought to be the most common colonizers of fallen leaves and the principle decomposers in this ecosystem (Leafio et al., 2000; Newell & Fell, 1992). Some species of *Pythium* such as *Pythium oligandrum* are dedicated saprobes and can often out-compete pathogenic species in agricultural systems (Martin & Loper, 1999). Various less well-characterized genera of free-living saprobes in soil are known, such as *Geolegnia*, but have not been studied extensively (Steciow et al., 2013). Recent DNA-based surveys have recovered many unknown oomycete sequences from environments with little to no apparent disease symptoms, suggesting oomycete saprobes might be more common than is currently thought (Hansen et al., 2012).

1.1.5 Host diversity

1.1.5.1 Terrestrial Plants

Representatives of oomycetes infect nearly every form of life on earth, ranging from algae to mammals, but are most well known for their ability to cause disease in terrestrial plants. Monocots such as grasses are infected by the graminicolous

downy mildews, including *Sclerospora graminicola* (Thines et al., 2015). Asteraceae, Caryophyllales, Convolvulaceae and Brassicales, are parasitized by the white rusts and the downy mildews (Choi et al., 2009; Rost & Thines, 2012; Wallace, Salgado-Salazar, Gregory, & Crouch, 2018) Gymnosperms, such as Japanese larch are infected by *Phytophthora* species like *Phytophthora ramorum* (Grünwald et al., 2012). Some mosses, such as *Physcomitrella patens*, can be infected with oomycetes and develop disease symptoms (Ponce de León, 2011).

1.1.5.2 Marine plants and algae

Like most other ancient groups of organisms, oomycetes evolved in the oceans and later migrated to land and, although their ancestors were likely photosynthetic, they lost their chloroplasts and became primarily parasites and saprobes, so it is not surprising that most forms of aquatic organisms are hosts for at least some oomycetes. Brown, green, and red algae are all infected by members of the genera *Olpidiopsis* (Klochkova et al., 2016; Sekimoto, Klochkova, West, Beakes, & Honda, 2009; West, Klochkova, Kim, & Loiseaux-de Goër, 2006) and *Atkinsiella* (Nakamura & Hatai, 1994). Additionally, brown algae are infected by the genera *Anisolpidium* (Gachon et al., 2017; Garvetto, Perrineau, Dressler-Allame, Bresnan, & Gachon, 2020) and *Eurychasma* (Gachon et al., 2009; Tsirigoti, Beakes, Hervé, Gachon, & Katsaros, 2015). Diatoms are infected by *Olpidiopsis* (Klochkova et al., 2016; Sekimoto et al., 2009), *Lagenidium* (Spies et al., 2016), and *Ectrogella* (Garvetto et al., 2020). There is even a record of a species of oomycete, *Lagenidium nodosum*, infecting cyanobacte-

ria (Dick, 2001). These examples listed above include every major group of primary producers in the world's oceans and emphasizes the vast influence oomycetes have on the largest ecosystem on earth.

1.1.5.3 Aquatic animals

Like aquatic plants, aquatic animals ranging from fish to rotifers are also infected by diverse oomycetes. Infections of crustaceans are particularly well known and, oddly enough, they are infected by some of the same genera that infect algae. Shrimp and lobsters are infected by members of the genera *Haliphthoros* (Fisher, Nilson, & Shleser, 1975; Tharp & Bland, 1977), *Halioticida* (Hatai, 2012; Holt et al., 2018), and *Lagenidium* (Holt et al., 2018). Additionally, shrimp are infected by some *Pythium* species (Hatai, 2012). Crabs are known to host species of *Plectospira* (Atkins, 1954) and *Atkinsiella* (Holt et al., 2018; Nakamura & Hatai, 1995). Opportunistic infections in fish, particularly of wounds, are caused by species in the genera *Saprolegnia* (Davis, 2016; Van West, 2006), *Pythium* (Martin & Loper, 1999), *Aphanomyces* (Blazer et al., 2002; Derevnina et al., 2016), and *Dictyuchus* (Rattan, Muhsin, & Ismail, 1978). Rotifers, which are common microscopic multicellular filter feeders, are infected by the genera *Aquastella* (Molloy et al., 2014) and *Atkinsiella* (Nakamura & Hatai, 1995). There is also some evidence that gastropods can be infected by some oomycetes, such as infection of abalone by *Halioticida noduliformans* (Derevnina et al., 2016). Considering that most of these hosts for which oomycete infections are known have been studied primarily due to their relevance to humans as common foods, it is likely

that many other aquatic animals also are infected by oomycetes.

1.1.5.4 Terrestrial Animals

Many terrestrial animals are also infected by oomycetes, including some mammals. Various mammals are infected by at least six species of oomycetes, including *Pythium insidiosum* (Spies et al., 2016) and *Lagenidium giganteum* (Vilela, Humber, Taylor, & Mendoza, 2019). *Pythium insidiosum* is known to infect humans, causing a subcutaneous vascular disease called pythiosis insidiosii (Mendoza, Hernandez, & Ajello, 1993). There are many records of amphibians being infected with oomycetes. Amphibian eggs and young are infected and often killed by *Saprolegnia infections* (Fernández-Benéitez, Ortiz-Santaliestra, Lizana, & Diéguez-Uribeondo, 2008). In some cases, adults stressed by pollution or other causes are also affected, such as infection of salamanders by *Saprolegnia parasitica* (Ruthig, 2009). Nematodes also are infected by oomcyetes, although relatively little research has been done on the subject, considering the abundance of nematodes and their large impact on most ecosystems. Nematodes are known to be infected by species of the holocarpic genera *Myzocytiopsis* (Glockling & Dick, 1997), *Gonimochaete*, *Haptoglossa* (Glockling & Beakes, 2000), *Lagenidium* (Spies et al., 2016), and *Chlamydomyzium* (Beakes, Glockling, & James, 2014). These oomycetes are particularly abundant in wet soils or in water near land (Glockling & Beakes, 2000). Finally, insects are known to be infected by *Lagenidium* and *Leptolegnia* (Pelizza, LASTRA, Becnel, Bisaro, & Garcia, 2007). It seems the same pathogens that infect mammals can also infect nematodes, suggesting that com-

plex multi-lost life cycles are possible (Spies et al., 2016). Like oomycete infections of aquatic animals, most the research on terrestrial oomycete infections has been done on hosts relevant to humans, so the overall diversity of animal hosts as a whole is probably much greater than is currently appreciated.

1.1.6 Effects on agriculture and aquaculture

1.1.6.1 Terrestrial pathogens

Oomycetes are most well known for their devastating effects on agriculture. Both the pathogens and the hosts they infect are highly diverse. Some pathogens that kill young seedlings as they emerge, an effect known as "damping off", are various species of *Pythium* on corn and soybean (Radmer et al., 2017), *Aphanomyces euteiches* on various legumes (Gaulin, Jacquet, Bottin, & Dumas, 2007), and *Globisporangium* on ornamental crops in greenhouses (Cuenca, 2016). Downy mildews are a particularly diverse group of related obligate pathogens including *Plasmopara viticola* on grape (Gessler et al., 2011), *Plasmopara halstedii* on sunflower (Sakr & others, 2014), *Sclerospora graminicola* on maize (Spring et al., 2018), *Peronosclerospora sorghi* on sorghum (Spring et al., 2018), *Pseudoperonospora cubensis* on cucumber (Savory et al., 2011), *Peronospora belbahriion* on basil, and *Hyaloperonospora* species on brassicaceous crops (Thines & Choi, 2016). The genus *Phytophthora* is particularly destructive to agriculture and includes the potato late blight pathogen *Phytophthora infestans* (Andrivon, 1996) and the pathogens *Phytophthora colocasiae* and *Phytoph-*

thora palmivora that cause devastating losses on many tropical crops in developing countries. Other pathogens include *Albugo candida*, the cause of white rust on *Brassica* crops (McMullan et al., 2015), and *Plectospira* species pathogenic on tomato and sugarcane (Jeronimo, Jesus, Rocha, Goncalves, & Pires-Zottarelli, 2017).

Many oomycete pathogens of woody plants have been known to devastate nurseries and threaten long-lived managed forest ecosystems. *Phytophthora ramorum* is an emerging pathogen causing sudden oak death in the US and sudden larch death in the United Kingdom and has also severely damaged some parts of the nursery and timber industry there (Brasier & Webber, 2010; Grünwald, LeBoldus, & Hamelin, 2019). It has also affected parts of the nursery industry in the western United States (Grünwald et al., 2012). *Rhododenron* and other woody ericaceous plants are particularly susceptible and act as vectors when asymptomatic plants are moved between nurseries. Holm oak decline in the traditional dehesa silvopastoral ecosystem of Portugal has been attributed to *Phytophthora cinnamomi*, although other factors are likely important contributors (Clara, Almeida Ribeiro, & others, 2013). Black pod disease of *Theobroma cacao*, the plant used to make chocolate, is caused by *Phytophthora megakarya* or *Phytophthora palmivora* and is currently one of the factors limiting production in some regions (Akrofi, 2015).

Most of the damage caused by terrestrial oomycetes is due to pathogens introduced from distant places on the globe (Hansen et al., 2012). These pathogens are thought to do relatively little harm to ecosystems they are native to, when the source of such pathogens is known (Studholme et al., 2019). The pathogen that causes grape downy mildew in Europe, *Plasmopara viticola*, is native to North America, where

is infects wild *Vitis* species. Ironically, it was probably introduced to Europe when cuttings from American grape were used to replant vineyards destroyed by phylloxera, an insect pathogen of grape (Gessler et al., 2011). A related pathogen that causes downy mildew of cultivated sunflower, *Plasmopara halstedii*, was also moved from the United States to France in the 1960's, where is causes losses of up to 50% (Sakr & others, 2014). Perhaps the best known example of the devastating effects an exotic pathogen can have on agriculture is *Phytophthora infestans*, which was moved from its native range in Central Mexico to Ireland, where it caused the Great Potato Famine, leading to the death and displacement of over a million people (Yoshida et al., 2013).

1.1.6.2 Aquatic pathogens

Although terrestrial oomycete pathogens are the most well studied, much of oomycete diversity is aquatic and some of these organisms cause damage to aquaculture. The recently characterized oomycete *Halioticida noduliformans* was found to cause up to 90% mortality in young cultured abalone (Muraosa, Morimoto, Sano, Nishimura, & Hatai, 2009). *Halioticida noduliformans* has also been found to infect the eggs of European lobster (Holt et al., 2018) and *Haliphthoros milfordensis* has been linked with the death of farmed American lobster (Fisher et al., 1975). Mortality in commercially harvested crustaceans, including crab, lobster, and shrimp have been linked to *Lagenidium* (Holt et al., 2018). *Saprolegnia* and *Aphanomyces* have caused large die-offs of freshwater crayfish in Europe (Holt et al., 2018). Various oomcyetes infect farmed seaweed and can cause losses of up to 30% (Tsirigoti et al., 2015). *Olidiopsis* causes

the disease "red rot" in the farmed red algae *Porphyra*, one of the seaweeds used for wrapping sushi, among other uses (Klochkova et al., 2012). The disease "winter kill" of catfish caused by *Saprolegnia parasitica* can reduce yields by up to 50% and can kill up to 22% of salmon returning to rivers to spawn, further threatening these fish already suffering from habitat loss (Van West, 2006). With the increasing demand for fish, shellfish, and seaweed, aquaculture has become one of the fastest growing food industries, greatly increasing the relevance of aquatic oomycete pathogens (Derevnina et al., 2016; Gachon et al., 2017).

1.1.6.3 Potential for biocontrol

Although most well-known oomycetes are dreaded pathogens, a few have potential applications for biocontrol of other pests. Some species of the genus *Lagenidium* are deadly pathogens specific to mosquito larva and have long been studied in the hopes of creating an effective biocontrol for mosquitos, which transmit deadly human diseases (Kerwin, 2007). A commercial biocontrol made with the species *Lagenidium giganteum* marketed as Laginex was available until several cases of infection of another strain of *Lagenidium giganteum* in dogs prompted the Environmental Protection Agency to deregister the product (Vilela et al., 2019). *Leptolegnia chapmanii* is also thought to have potential to control mosquitos (Pelizza et al., 2007; Seymour, 1984). Oddly enough, some oomycetes have potential as biocontrols for other closely related oomycetes. Some studies have shown that treatments with *Pythium oligandrum* reduce damping off of tomato seedlings caused by *Pythium ultimum* as much as

treatments by metalaxyl, a leading treatment for oomycete pathogens in agriculture (Martin & Loper, 1999). Considering almost all major groups of life are infected by oomycetes and many oomycetes are highly host-specific, there might be other undiscovered opportunities for oomycete-based biocontrol.

1.1.7 Effects on natural ecosystems

Although oomycetes are mostly known for the great damage they cause to agriculture, they also have a significant impact on natural ecosystems. Where oomycetes are native, they are often minor pathogens and parasites of a large variety of plants and animals. In contrast, exotic oomycetes can cause extensive damage to forested ecosystems (Hansen et al., 2012; Jung et al., 2018) and can significantly alter ecosystem composition and functioning, which is a particular concern for iconic landscapes, such as New Zealand's Kauri trees which are threatened by *Phytophthora agathidicida* (Davis, 2016). Some are also harmless saprobes that contribute to the mineralization of detritus and in some environments, such as mangrove swamps, where they are the principal decomposers (Bennett & Thines, 2019). Saprobes and minor pathogens likely have a larger cumulative effect on ecosystems, although much less is known about these organisms than their more destructive relatives.

1.1.7.1 Terrestrial pathogens

The impact of oomycetes on terrestrial ecosystems, particularly forests, are the most well studied among impacts on natural systems. In the last few decades, the pathogen *Phytophthora ramorum* has caused extensive death of oak species in the Western United States and of Japanese larch in the United Kingdom (Grünwald et al., 2012). Since the 1990's *Phytophthora ramorum* has been causing massive die-offs of *Notholithocarpus densiflorus* (tanoak) and *Quercus agrifolia* (coast live oak) in California (Rizzo, Garbelotto, & Hansen, 2005). *Phytophthora cinnamomi* is associated with widespread die-back of *Eucalyptus marginata* (jarrah) trees and associated understory vegetation, particularly on waterlogged soils after logging (Shearer & Tippett, 1989). Some researchers have questioned the evidence that die-back of *Eucalyptus marginata* is in fact caused by *Phytophthora cinnamomi*, even though the pathogen is often found in association with dying trees, but the death of associated understory vegetation at least seems attributable to the pathogen (Davison, 2015). *Phytophthora agathidicida* has caused the death of many of the iconic and culturally-significant Kauri trees of New Zealand (Davis, 2016). Kauri trees are large and impressive trees once extensively logged for lumber and to clear land for farming and the relatively few ancient kauri trees that remain are now under threat due to *Phytophthora agathidicida*. Although agricultural diseases of annual crops often get the greatest attention, invasive diseases of long-lived trees in natural systems are much harder to control and have the potential to permanently alter entire landscapes.

1.1.7.2 Aquatic pathogens

Oomycetes also include many important aquatic pathogens of amphibians, fish, and arthropods in both marine and freshwater ecosystems, although these are less studied than terrestrial pathogens (Rasconi, Jobard, & Sime-Ngando, 2011). These pathogens usually have the greatest effect on the eggs and larval stages of their host, but sometimes can also affect adults. For example, the eggs of the American bullfrog, *Rana catesbeiana* can be killed by members of *Saprolegnia* and related oomycetes (Ruthig, 2009). These pathogens are particularly damaging to hosts weakened by pollution or warmer temperatures and such combinations of factors could be contributing to the drastic decline in amphibian populations around the world (Ruthig, 2009). Lobster and crayfish, along with other invertebrates, are hosts to a variety of oomycete pathogens including species of *Lagenidium*, *Haliphthoros*, *Halocrusticida*, and *Atkinsiella*, among others (Holt et al., 2018). For some hosts, such as the American lobster (*Homarus americanus*), oomycetes are thought to be one of the leading causes of death of larva (Holt et al., 2018). Freshwater crayfish are known to be infected by species of *Saprolegnia* and *Aphanomyces*, and outbreaks of disease caused by these pathogens have devastated natural populations of crayfish in Europe (Holt et al., 2018). *Saprolegnia* is also known to infect fish. *Saprolegnia parasitica* infections of minor wounds in salmon have been known to kill up to 22% of fish returning to rivers to spawn and might be contributing to the widespread decline of salmon (Van West, 2006). Oomycetes, such as *Eurychasma dicksonii*, are also well-known pathogens of brown algae, which make up the majority of biomass in some in temperate shores (Ga-

chon et al., 2009). In contrast to the many oomycetes that have rather narrow host ranges, *Eurychasma dicksonii* infects almost all species of brown algae that have been tested (at least 45 species (Tsirigoti et al., 2015)), making this species particularly important for ecosystems where brown algae are abundant (Gachon et al., 2009). Finally, some oomycetes, such as *Ectrogella* species infect diatoms, which are important primary produces in many aquatic ecosystems (Garvetto et al., 2020). Outbreaks of disease caused by *Ectrogella perforans* has caused up to 99% mortality in populations of the diatom *Licmophora* in the United States (Garvetto et al., 2020).

1.1.7.3 Saprobes

Perhaps the least well-characterized ecological group of oomycetes is the saprobes of aquatic and terrestrial ecosystems, although what is known suggests that they play a major part in the recycling of nutrients in some ecosystems (Czeczuga et al., 2005). In mangroves and salt marshes, species of *Halophytophthora* play a major role as decomposers of fallen leaves (Bennett & Thines, 2019). In estuaries and salt marshes, the genera *Phytopythium*, *Salisapilia*, *Salispina*, and *Calycofera* are also abundant (Bennett & Thines, 2019). Oomycetes of the genera *Aphanomyces*, *Saprolegnia*, and *Pythium*, have also been found to be common decomposers of plant debris in freshwater ecosystems (Czeczuga et al., 2005). In soil, species of *Pythium* are the primary saprobes for some soils with high water contents (Martin & Loper, 1999). Considering the recent findings of many unknown oomycetes not associated with disease (Jung et al., 2018), the diversity and importance of saprotrophic oomycetes

is likely greatly underestimated.

1.1.8 Priorities for future research

Considering the abundance, diversity, and impact of oomycetes on ecosystems and human welfare, relatively little is known about them compared with other groups of organisms. This is partially due to their microscopic nature, our inability to readily culture them, and, for some groups, a lack of distinguishing morphological traits. Because of these challenges, the taxonomy of many groups, which is still largely based on inconsistent morphological features, has been in constant flux and in some cases remains quite uncertain (Hatai, 2012; Muraosa et al., 2009; Telle, Shivas, Ryley, & Thines, 2011). The same organisms might have three different names in three different studies, making it so only experts in that group of organisms have a hope of reconciling that information. This is particularly problematic since the only definitive descriptions of some oomycetes are quite old and use outdated concepts and names. Additionally, many oomycetes cannot be grown in pure culture, since they rely on a host for reproduction. This makes studying oomycetes and consolidating findings from multiple studies particularly challenging (Choi, Thines, Tek, & Shin, 2012). However, modern molecular techniques might make it possible to overcome some of these difficulties. In the era of whole genome sequencing, rapid progress on the molecular taxonomy of oomycetes is to be expected (Baxter et al., 2010; Haas et al., 2009; Tyler, 2001).

In order to facilitate the study of these important organisms, a comprehensive

phylogenetics-based taxonomy for oomycetes and culture-independent tools for identification, such as barcoding and metabarcoding, are needed. Studies revising the taxonomy of oomycetes using phylogenetics are becoming increasingly common, but some findings will require drastic changes if each taxon is to correspond to a monophyletic evolutionary clade. For example, all of the 19 genera and at least 700 species of downy mildews, including many well-known pathogens, are in the clade containing all *Phytophthora* species, another group containing many important pathogens (Jung et al., 2017). Renaming taxa in either group would be a major change, requiring an international consensus. Many species of *Pythium*, another important and well-studied group, have recently been redistributed to four new genera: *Globisporangium*, *Ovatisporangium*, *Elongisporangium*, and *Pilasporangium* (Uzuhashi, Hata, Matsuura, & Tojo, 2017). Several other taxa were recently found to be non-monophyletic, including *Halophytophthora* (Yang & Hong, 2014), *Achlya* (Spencer & others, 2002), and *Myzocytiopsis* (Glockling & Beakes, 2006). Such changes will likely cause confusion in the short term, but as molecular identification techniques become increasingly, they will facilitate the study of oomycetes into the future.

Molecular techniques for species identification, both for individual samples and communities as a whole, have revolutionized the study of fungal and bacterial ecology and biodiversity, but such techniques have been applied much less to oomycetes and are less refined where they have been applied. One of the main benefits of molecular identification techniques is the ability to detect pathogens without culturing, which is important for obligate pathogens that are impossible to culture (West et al., 2006) or don't always cause symptoms (Lava et al., 2013). These techniques also are usable by

people who are not experts at morphology-based identification, greatly expanding the proportion of researchers able to study oomycetes. Techniques like metabarcoding are particularly powerful since they allow for an entire community of related organisms to be identified at once. This has the potential to finally uncover the extent of oomycete diversity in natural ecosystems that has been only hinted at by previous research. However, in order for techniques like metabarcoding to yield useful results, public databases of reference sequences assigned to a reliable taxonomy are required. Currently reference databases for oomycetes lag far behind those for bacteria and fungi. If reliable taxonomic and molecular methods for identification are developed, the resulting increase in information could parallel the remarkable advancement of fungal and bacterial ecology in recent years.

1.2 Metabarcoding of oomycetes: tools and analysis

1.2.1 Introduction to metabarcoding

Metabarcoding is a powerful high-throughput sequencing technique used to identify multiple organisms at once using the DNA sequence of a particular gene (Taberlet, Coissac, Pompanon, Brochmann, & Willerslev, 2012). The first step is extracting DNA from a sample containing an unknown mixture of organisms, such as soil, water, or tissue (Figure 1.2). A gene is chosen with a sequence variable enough in the taxon of interest to distinguish species-level differences (ideally) and flanked by conserved regions for which PCR primers can be designed. PCR is then used to amplify the

gene using primers specially designed to match all organisms in the taxonomic group of interest. This produces a mixture of amplicons intended to represent the diversity of organisms in the community. The amplicons are then sequenced using a high-throughput sequencing instrument like the Illumina MiSeq. One or more FASTQ files are produced by the sequencer containing the sequences of a subset of the amplicons. These sequences and associated quality scores are the starting point for computational analysis.

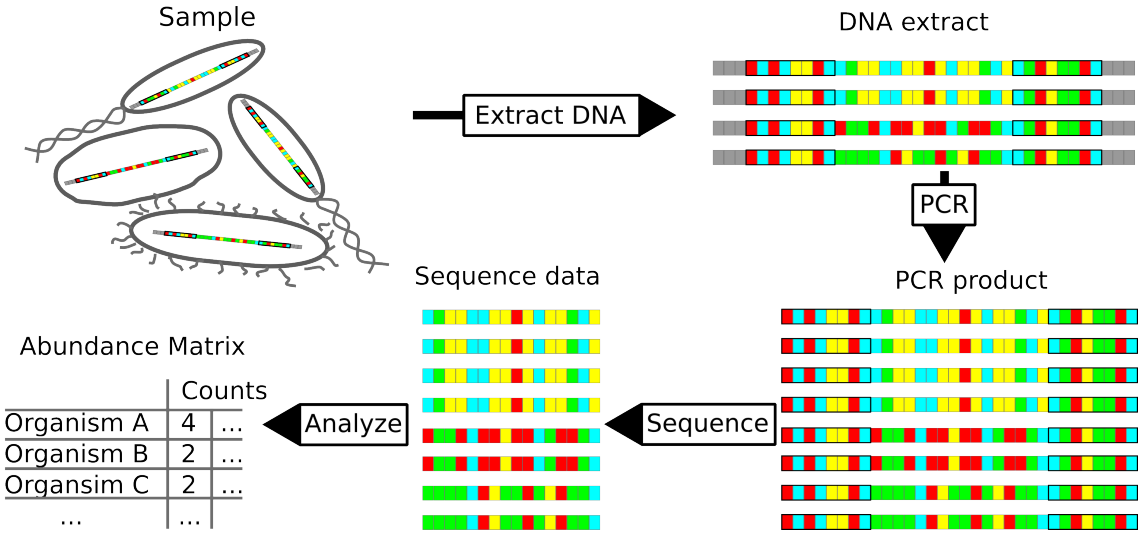

Figure 1.2: Visual summary of the metabarcoding method.

1.2.2 Computational analysis of metabarcoding data

The first step in the computational analysis of metabarcoding data typically involves clustering similar sequences together, assigning taxonomic classifications to clusters, and compiling the counts of reads in each cluster for each sample in an abundance

matrix. First, primer sequences and low-quality regions of the reads are filtered out. If paired-end sequencing is used, the pairs of reads are merged into a single sequence and any that cannot be merged are filtered out. Then some form of clustering is used to group similar sequences together. The main two approaches are clustering into Operational Taxonomic Units (OTUs) and Amplified Sequence Variants (ASVs).

Table 1.1: Types of algorithms used in OTU clustering

Clustering method	Cluster grouping criteria	Cluster size
Complete-linkage	All sequences are more similar than the threshold	Smallest
Average-linkage	The mean pairwise dissimilarity is less than the threshold	Medium
Single-linkage	At least one pair of sequences are more similar than the threshold	Largest

Clustering into OTUs is the older of the two methods and is intended to cluster sequences into groups that approximate species-level differences. A clustering threshold in terms of percent sequence similarity is chosen, based on the taxon and gene used in the study, and sequences more similar to each other than this clustering threshold are grouped together into clusters called OTUs (Taberlet et al., 2012). There are three main clustering approaches that vary in how the clustering threshold is applied: complete-linkage (a.k.a., farthest neighbor clustering), average-linkage (a.k.a., unweighted pair group method with arithmetic mean), and single-linkage (a.k.a.,

nearest neighbor clustering) (Table 1.1). The OTU method is useful for approximating diversity of communities in terms of number of species and helps to mitigate the effects of sequencing error by clustering erroneous sequences with the "real" sequence they are derivatives of. However, some errors are large enough not to be nullified by OTU clustering, so OTU-based methods often overestimate diversity (Edgar, 2017). Another problem is that OTU-based methods produce results that cannot be compared across studies, since OTU clusters are emergent properties of the data set and method used (Callahan et al., 2016). The ASV-based method is meant to address these shortcomings of OTU-based methods.

ASVs are a new approach that uses a model of sequence mutation to cluster sequences based on their abundance and the abundance of similar sequences (Callahan et al., 2016). The likelihood of specific transformations needed to convert more abundant sequences to less abundant sequences is calculated and if the resulting p-value is below a threshold the lower abundance sequences are clustered with the higher abundance sequence. The intent of this method is to correct errors accumulated during the PCR and sequencing steps, yielding only the original templates amplified by the PCR. ASVs therefore, unlike OTUs, attempt to represent only real biological sequences, making ASVs comparable across studies, although uncorrected errors might still exist. It also does a much better job avoiding the error-induced inflation of apparent diversity that is typical of OTU-based methods. However, these sequences do not represent species-level differences as OTUs do, so diversity statistics derived from ASVs must be interpreted somewhat differently. Although OTU-based methods are still much more popular, the advantages of the ASV methods will likely make them

the standard method in the future.

Once similar sequences are grouped together into OTUs or ASVs, they are typically used to calculate diversity statistics to characterize samples and the differences between samples. Diversity in the ecological sense is intuitively understood as the complexity of one or more samples of a community of organisms. There are many ways to quantify this complexity so that communities can be compared objectively. The two main categories of methods are known as alpha diversity and beta diversity (Whittaker, 1960). Alpha diversity measures the diversity within a single sample and is generally based on the number and relative abundance of taxa at some rank (e.g., species or OTUs). Beta diversity also uses the number or relative abundance of taxa at some rank, but measures variation between samples. In other words, an alpha diversity statistic describes a single sample and a beta diversity statistic describes how two samples compare. There are numerous statistics commonly used for both alpha and beta diversity, some of which incorporate abundance information or phylogenetic relatedness of the species analysed, or both (Figures 1.3 and 1.4). These statistics are often the basis for other analyses and visualizations, like ordination, and are often used to present the primary findings of metabarcoding research.

	Uses relative abundance	**Ignores relative abundance**
Uses phylogenetic relatedness	Rao's quadratic entropy "Hp" from Allen *et al.* 2009 "Iq" from Pavione *et al.* 2009 Adapted Hill numbers from Chao *et al.* 2010	Cladistic diversity (CD) phylogenetic diversity (PD)
Ignores phylogenetic relatedness	Simposon Shannon Chao1 ACE Hill numbers	Richness

Figure 1.3: Alpha diversity statistics used in community ecology.

	Uses relative abundance	**Ignores relative abundance**
Uses phylogenetic relatedness	Weighted Unifrac Generalized UniFrac	Unweighted Unifrac
Ignores phylogenetic relatedness	Bray-Curtis Canberra ...	Sørensen Jaccard ...

Figure 1.4: Beta diversity statistics used in community ecology.

After sequences are grouped into OTUs or ASVs, the representative sequence for each cluster is assigned a taxonomic classification by comparing them to reference database sequences. Two of the more popular methods used to assign a taxonomy to sequences are BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990) and the Naive Bayesian Classifier (Wang, Garrity, Tiedje, & Cole, 2007). When BLAST is used, the taxonomy of the top BLAST hit is used as the taxonomy for the representative sequence. Determining which hit is the best is often done using the e-value or a combination of coverage and percent identity. One downside of this method is that there is no per-rank confidence measure, making it difficult to determine how much

confidence to have in taxonomic classifications. This can be addressed by using somewhat arbitrary thresholds for e-value, coverage, or percent-identity for each rank or picking multiple top BLAST hits and and only reporting information for the ranks for which all the hits agree for a given query sequence. This problem is better handled by the Naive Bayesian Classifier, which assigns a bootstrap value to each rank in the taxonomic classification for each sequence. The sequence is broken up into K-mers and the number of shared K-mers with reference sequences are used to pick the closest reference sequence. This process is repeated many times with random sub samples of the K-mers to produce the bootstrap values for each rank. Although this method produces more robust taxonomic classifications, it is often so conservative relative to BLAST-based methods that it has not been adopted extensively outside of bacterial metabarcoding, where the reference databases are more well-developed (Tedersoo, Drenkhan, Anslan, Morales-Rodriguez, & Cleary, 2019).

The taxonomy of the sequence clusters is the primary end-product of most metabarcoding research, so robust tools for analyzing data in a taxonomic context are needed. The hierarchical natural of taxonomic classifications combined with the scale of metabarcoding data makes taxonomic information difficult to analyze and visualize. Although the R programming language has extensive support for metabarcoding and ecological research, it has no packages dedicated to analyzing taxonomic information. We created the R packages `taxa` (Foster, Chamberlain, & Grünwald, 2018) and `metacoder` (Foster, Sharpton, & Grünwald, 2017) to fill this gap. The `taxa` package provides all-purpose classes and methods for manipulating taxonomic data. It is meant to provide a robust basis for other more specialized

packages to build upon, hopefully encouraging the creation of an ecosystem of compatible packages. One such package is `metacoder`, which provides functions for visualizing data in a taxonomic context and functions to do commonly-needed tasks in metabarcoding research using the classes provided by the `taxa` package. The primary contribution of the `metacoder` package is a visualization method we call "heat trees" that uses color and size to plot data on a taxonomic tree. Heat trees are uniquely well-suited for depicting data from metabarcoding research, such as taxon abundance and differential taxon abundance between experimental factors. This provides an alternative to stacked bar charts, which are commonly used for the same purpose, but are not well suited for hierarchical data with many classes. Both packages can also be used for any type of hierarchical data, such as geography and gene ontology (Foster et al., 2017). These packages provide additional functionality to the already very flexible R ecosystem.

1.2.3 Oomycete metabarcoding

We used a range of approaches to characterize oomycete biodiversity associated with terrestrial plants including an ITS-based method and a novel *rps10*-based method. The composition of plant microbiomes influences important agricultural processes such as nutrient absorption and plant health. Plant genotype and environment affect the microbiome, but the nature and relative importance of these effects are not well understood. We evaluated the effect of host genotype, nursery, and production system (potted versus in-ground planting) on the composition of the fungal and oomycete rhi-

zosphere microbiome of rhododendrons in Oregon nurseries (Foster, Weiland, Scagel, & Grünwald, 2020). Rhizosphere and roots were sampled from randomly selected, potted and in-ground plants of 3 host cultivars at 4 nurseries. ITS1 amplicons were sequenced using the Illumina MiSeq. We found fewer oomycetes than expected and the ITS1-based method proved to have numerous shortfalls including an unreliable PCR reaction that required extensive optimization to avoid non-target amplification. To address this issue, our succeeding studies that used oomycete metabarcoding employed a new *rps10*-based method.

Although metabarcoding methods for fungi and bacteria are well-developed and often used, oomycete metabarcoding methods are still quite experimental and not yet widely used. Traditionally, the detection and identification of oomycetes has relied on culturing from baits or infected plant material. For other groups of microorganisms, such as fungi and bacteria, the culture-independent high-throughput sequencing technique metabarcoding has replaced such techniques in many cases, but metabarcoding has only rarely been applied to oomycete communities due to a lack of an effective locus, primers, and a reference database designed for the purpose. Here we present work demonstrating that the mitochondrial gene *rps10* could be used to allow for improved metabarcoding of oomycete communities compared with the current ITS-based technique. The protocols and resources presented here should allow for a more effective and less biased way of characterizing oomycete communities compared with currently available ITS-based methods, potentially improving the detection and control of the many damaging pathogens in this group of organisms.

Chapter 2: Taxa: An R package implementing data standards and methods for taxonomic data

Zachary S.L. Foster, Scott Chamberlain, and Niklaus J. Grünwald

Published in:

F1000Research, 2018, 7:272

DOI: 10.12688/f1000research.14013.2

2.1 Abstract

The `taxa` R package provides a set of tools for defining and manipulating taxonomic data. The recent and widespread application of DNA sequencing to community composition studies is making large data sets with taxonomic information commonplace. However, compared to typical tabular data, this information is encoded in many different ways and the hierarchical nature of taxonomic classifications makes it difficult to work with. There are many R packages that use taxonomic data to varying degrees but there is currently no cross-package standard for how this information is encoded and manipulated. We developed the R package `taxa` to provide a robust and flexible solution to storing and manipulating taxonomic data in R and any application-specific information associated with it. `Taxa` provides parsers that can read common sources of taxonomic information (taxon IDs, sequence IDs, taxon names, and classifications) from nearly any format while preserving associated data. Once parsed, the taxonomic data and any associated data can be manipulated using a cohesive set of functions modeled after the popular R package `dplyr`. These functions take into account the hierarchical nature of taxa and can modify the taxonomy or associated data in such a way that both are kept in sync. `Taxa` is currently being used by the `metacoder` and `taxize` packages, which provide broadly useful functionality that we hope will speed adoption by users and developers.

2.2 Introduction

The R statistical computing language is rapidly becoming the leading tool for scientific data analysis in academic research programs (Tippmann, 2015). One of the reasons for R's popularity is how easy it is to develop and install extensions called R packages, relative to other programming languages. There are now more than 10,000 packages on the Comprehensive R Archive Network (CRAN), over 1,300 packages on Bioconductor (Gentleman et al., 2004), and countless more on GitHub.

The recent increases in the affordability and effectiveness of high-throughput sequencing has led to a large number of ecological datasets of unprecedented size and complexity. The R community has responded with the creation of numerous packages for ecological data analysis and visualization, such as `vegan` (Oksanen et al., 2013), `phyloseq` (McMurdie & Holmes, 2013), `taxize` (Chamberlain & Szöcs, 2013), and `metacoder` (Foster et al., 2017). Taxonomic information is often associated with these large data sets and each package encodes this information differently. Some store taxonomic classification as a table with ranks as columns (e.g. `phyloseq`), some store it as simple character vectors (i.e. plain text) or column/row names, leaving it up to the user to decide on the details on how taxa in the classification are distinguished (e.g. `vegan`), and some store it as a list of tables with one classification in each table (e.g. `taxize`). Since each package tends to have a unique focus, it is common to use multiple packages on the same data set but converting between formats can be difficult. Considering how recently these large taxonomic data sets have become commonplace, it is likely that many more packages that use taxonomic information

will be created.

Without a common data standard, using multiple packages with the same data set requires constant reformatting, which complicates analyses and increases the chance of errors. Package maintainers often add functions to convert between the formats of other popular packages, but this practice will become unsustainable as the number of packages dealing with taxonomic data increases. Even if a conversion function exists, doing the conversion can significantly increase the time needed to analyze very large data sets, like those generated by high-throughput sequencing. In addition, not all formats accommodate the same types of information, so conversion can force a loss of information.

The sources of taxonomic data, typically online databases, also vary in how they are encoded. Reference sequence databases used in ecology research often have taxon names in the headers separated by some character, but the details differ. For example, the popular Greengenes database (McDonald et al., 2012) for prokaryotic 16S sequences encodes classifications as follows:

```
k__Bacteria; p__Cyanobacteria; c__Synechococcophycideae...
```

In contrast, the SILVA database (Yilmaz et al., 2014) uses:

```
Bacteria;Proteobacteria;Gammaproteobacteria...
```

And the Ribosomal Database Project (RDP) (Cole et al., 2014) has the ranks and taxon names intermixed with the same separator:

```
Root;rootrank;Fungi;domain;Ascomycota...
```

These minor differences, while not a problem for humans to understand, mean that different code must be used to read each type. Also, this information is often intermixed with other information in the same header, like the sequence ID or description of the organism further complicating parsing. In other cases, a classification might not be supplied at all, but just a taxon name (e.g. *Homo sapiens*), sequence ID, or taxon ID, as is done in sequences downloaded from GenBank:

```
>AC005336.1 Homo sapien chromosome 19
```

In this case the classification must be looked up using tools like the `taxize` package, but to do that the relevant information must be extracted from the rest of the header.

`Taxa` is a new R package that defines classes and functions for storing and manipulating taxonomic data. It is meant to provide a solid foundation on which to build an ecosystem of packages that will be able to interact seamlessly with minimal hassle for developers and users. It also provides highly flexible functions to read in data (i.e. parsers) from diverse formats, allowing it to be used with the ever-changing and proliferating selection of file formats used by biologists. The classes in taxa are designed to be as flexible as possible so they can be used in all cases involving taxonomic information. Complexity ranges from low level classes used to store the names of taxa, ranks, and databases to high-level classes that can store multiple data sets associated with a taxonomy. In particular, the taxmap class is designed to hold any type of arbitrary, user-defined data associated with taxonomic information, making its applications limitless. In addition to the classes, there are associated functions for manipulating data based on the `dplyr` philosophy (Wickham, Francois, Henry,

Müller, & others, 2015). These functions provide an intuitive way of filtering and manipulating both taxonomic and user-defined data simultaneously. In combination with flexible parsers and classes, this allows for `taxa` to be used to subset complicated data/files based on their associated taxonomic information.

2.3 Methods

2.3.1 Implementation

The basic classes. `Taxa` defines some basic taxonomic classes and functions to manipulate them (Figure 2.1). The goal is to use these as low-level building blocks that other R packages can use. The database class stores the name of a database and any associated information, such as a description, its URL, and a regular expression matching the format of valid taxon identifiers (IDs):

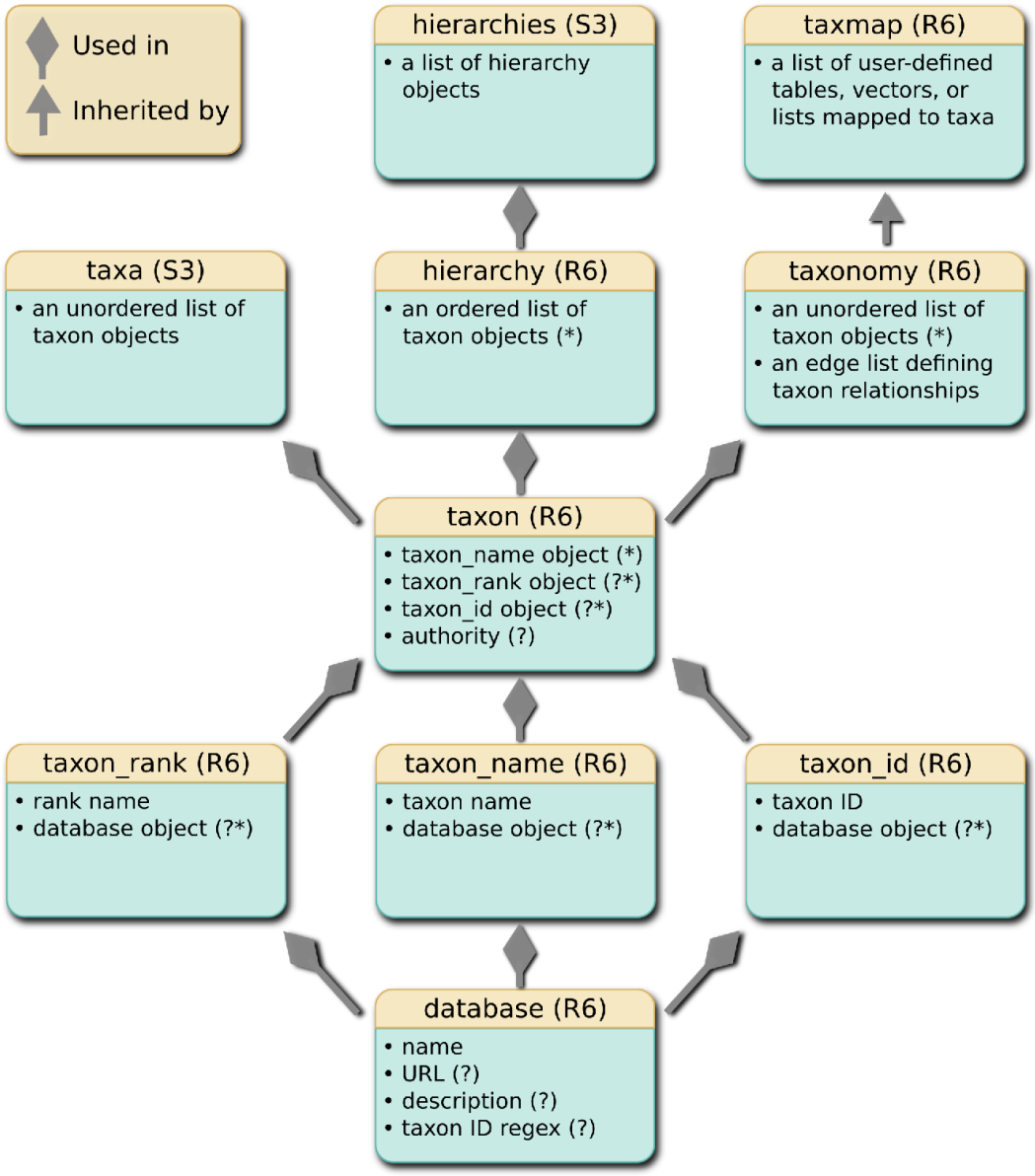

Figure 2.1: A class diagram representing the relationship between classes implemented in the `taxa` package. Diamond-tipped arrows indicate that objects of a lower class are used in a higher class. For example, a database object can be stored in the `taxon_rank`, `taxon_name`, or `taxon_id` objects. A standard arrow indicates that the lower class is inherited by the higher class. For example, the `taxmap` class inherits the `taxonomy` class. An asterisk indicates that an object (e.g. a database object) can be replaced by a simple character vector. A question mark indicates that the information is optional.

```
taxon_database(
  name = "ncbi",
  url = "http://www.ncbi.nlm.nih.gov/taxonomy",
  description = "NCBI Taxonomy Database",
  id_regex = "*")
#> <database> ncbi
#>   url: http://www.ncbi.nlm.nih.gov/taxonomy
#>   description: NCBI Taxonomy Database
#>   id regex: *
```

The classes `taxon_name`, `taxon_id`, and `taxon_rank` store the names, IDs, and ranks of taxa and can include a database object indicating their source:

```
taxon_name("Poa", database = "ncbi")
#> <TaxonName> Poa
#>   database: ncbi
taxon_rank(name = "species", database = "ncbi")
#> <TaxonRank> species
#>   database: ncbi
taxon_id(12345, database = "ncbi")
#> <TaxonId> 12345
#>   database: ncbi
```

All of the classes mentioned so far can be replaced with character vectors in the higher-level classes that use them. This is convenient for users who do not have or need database information. However, using these classes allows for greater flexibility and rigor as the taxa develops; new kinds of information can be added to these classes without affecting backwards compatibility and the database objects stored in the `taxon_name`, `taxon_id`, and `taxon_rank` classes can be used to verify the integrity of data, even if data from multiple databases are combined. These classes are used to create the taxon class, which is the main building block of the package. It stores the

name, ID, and rank of a taxon using the `taxon_name`, `taxon_id`, and `taxon_rank` classes. The taxa class is simply a list of taxon objects with a custom print method (i.e. the function controlling how it is displayed when printed to the console).

The hierarchy and taxonomy classes. The `taxon` class is used in the `hierarchy` and `taxonomy` classes, which store multiple taxa (Figure 2.1). The `hierarchy` class stores a taxonomic classification composed of nested taxa of different ranks (e.g. Animalia, Chordata, Mammalia, Primates, Hominidae, *Homo*, *sapiens*). Each taxon is stored as a `taxon` object in a list in the order they appear in the classification, from most inclusive to most specific. The `hierarchies` class is simply a list of `hierarchy` objects with a custom print method. The `hierarchies` class has the convenience of each `hierarchy` being independent, making it easy to subset by index or name, but it could also waste memory by storing multiple copies of the more coarse taxa (e.g. Animalia) that are likely to appear in many `hierarchy` objects. The `taxonomy` class is a more memory-efficient alternative that can store the same information.

The `taxonomy` class stores multiple taxa in a tree structure representing a taxonomy. The individual taxa are stored as a list of `taxon` objects and the tree structure is stored as an edge list representing subtaxa-supertaxa relationships. The edge list is a two-column table of taxon IDs that are automatically generated for each taxon. Using automatically generated taxon IDs, as opposed to taxon names, allows for multiple taxa with identical names. For example, *Achlya* is the name of an oomycete genus as well as a moth genus. It is also preferable to using taxon IDs from particular databases, since users might combine data from multiple databases and the same ID

might correspond to different taxa in different databases. For example, "180092" is the ID for Homo sapiens in the Integrated Taxonomic Information System, but is the ID for *Acianthera teres* (an orchid) in the NCBI taxonomy database. The tree structure of the taxonomy class uses less memory than the same information saved as a table of ranks by taxa, since the information for each taxon occurs in only one instance. It also does not require explicit rank information (e.g. "genus" or "family").

The taxmap class. The `taxmap` class inherits the `taxonomy` class and is used to store any number of data sets associated with taxa in a taxonomy (Figure 2.1). A list called "data" stores any number of lists, tables, or vectors that are mapped to all or a subset of the taxa at any rank in the taxonomy. Therefore, the raw data used to make the object (and any other data associated with it) can be included in the `taxmap` object itself in its original form. In the case of tables, the presence of a "taxon_id" column containing unique taxon IDs indicates which rows correspond to which taxa. Lists and vectors can be named by taxon IDs to indicate which taxa their elements correspond to. When a taxmap object is subset or otherwise manipulated, these IDs allow for the taxonomy and associated data to remain in sync. The taxmap also contains a list called "funcs" that stores functions that return information based on the content of the taxmap object. In most functions that operate on taxmap objects, the results of built-in functions (e.g. `n_obs`), user-defined functions, and the user-defined content of lists, vectors, or columns of tables can be referenced as if they are variables on their own, using non-standard evaluation (NSE). NSE is a technique used to make functions more convenient to use by interpreting things like variable names in a function call differently than they would be outside the function call or in

other functions not using NSE. Any value returned by the `all_names` function can be used in this way. This greatly reduces the amount of typing needed and makes the code easier to read.

Manipulation functions. The `hierarchy`, `hierarchies`, and `taxa` classes have a relatively simple structure that is easily manipulated using standard indexing (i.e. using `[`, `[[`, or `$`), but the `taxonomy` and `taxmap` classes are hierarchical, making them much harder to modify. To make manipulating these classes easier, we have developed a set of functions based on the `dplyr` data manipulation philosophy. The `dplyr` framework provides a consistent, intuitive, and chain-able set of commands that is easy for new users to understand. For example, `filter_taxa` and `filter_obs` are analogs of the `dplyr` `filter` function used to subset tables.

One aspect that makes `dplyr` convenient is the use of NSE to allow users to refer to column names as if they are variables on their own. The taxa package builds on this idea. Since `taxmap` objects can store any number of user-defined tables, vectors, lists, and functions, the values accessible by NSE are more diverse. All columns from any table and the contents of lists/vectors are available. There are also built-in and user-defined functions whose results are available via NSE. Referring to the name of the function as if it were an independent variable will run the function and return its results. This is useful for data that is dependent on the characteristics of other data and allows for convenient use of the `magrittr` `%>%` piping operator. For example, the built-in `n_subtaxa` function returns the number of subtaxa for each taxon. If this was run once and the result was stored in a static column, it would have to be updated each time taxa are filtered. If there are multiple filtering steps piped together using

`%>%`, a static "n_subtaxa" column would have to be recalculated after each filtering to keep it up to date. Using a function that is automatically called when needed eliminates this hassle. The user still has the option of using a static column if it is preferable to avoid redundant calculations with large data sets.

Unlike `dplyr`'s filter function, `filter_taxa` works on a hierarchical structure and, optionally, on associated data simultaneously. By default, the hierarchical nature of the data is not considered; taxa that meet some criterion are preserved regardless of their place in the hierarchy. When the `subtaxa` option is `TRUE`, all of the subtaxa of taxa that pass the filter are also preserved and when `supertaxa` is `TRUE`, all of the supertaxa are likewise preserved. For example,

```
filter_taxa(my_taxmap, taxon_names == 'Fungi', subtaxa = TRUE)
```

would remove any taxa that are not named "Fungi" or are not a subtaxon of a taxon named "Fungi". By default, steps are taken to ensure that the hierarchy remains intact when taxa are removed and that user-defined data are remapped to remaining taxa. When the `reassign_taxa` option is `TRUE` (the default), the subtaxa of removed taxa are reassigned to any supertaxa that were not removed, keeping the tree intact. When the `reassign_obs` option is `TRUE` (the default), any user-defined data assigned to removed taxa are reassigned to the closest supertaxa that passed the filter if such a taxon exists. This makes it easy to remove parts of the taxonomy without losing associated information. Finally, if the `drop_obs` option is `TRUE` (the default), any user-defined data assigned to removed taxa are also removed, allowing for subsetting of user-defined data based on taxon characteristics. The many combinations of these powerful options make `filter_taxa` a flexible tool and make

it easier for new users to deal with the hierarchical nature of taxonomic data. For example, if the `drop_obs` option is `TRUE` (the default) and the `reassign_obs` option is `FALSE`, then any user-defined data assigned to taxa are removed even if a supertaxon is preserved. If the `drop_obs` option is `FALSE`, and the `reassign_obs` option is `FALSE`, then data associated with removed taxa is assigned a taxon ID placeholder of NA, but not removed. The function `sample_n_taxa` is a wrapper for `filter_taxa` that randomly samples some number of taxa. All of the options of `filter_taxa` can also be used for `sample_n_taxa`, in addition to options that influence the relative probability of each taxon being sampled.

Other `dplyr` analogs that help users manipulate their data include `filter_obs`, `sample_n_obs`, and `mutate_obs`. `filter_obs` is similar to running the `dplyr` function `filter` on a tabular, user-defined dataset, except that there are more values available to NSE and lists and vectors can also be subset. The `drop_taxa` option can be used to remove any taxa whose only observations have been removed during the filtering. The `sample_n_obs` function is a wrapper for `filter_obs` that randomly samples some number of observations. Like `sample_n_taxa`, there are options to weight the relative probability that each observation will be sampled. The `mutate_obs` function simply adds columns to tables of user-defined data.

Mapping functions. There are also a few functions that create mappings between different parts of the data contained in `taxmap` or `taxonomy` objects. These are heavily used internally in the functions described already, but are also useful for the user. The `subtaxa` and `supertaxa` functions return the taxon IDs (or other values) associated with all subtaxa or supertaxa of each taxon. They return one value per

taxon. The `recursive` option controls how many ranks below or above each taxon are traversed. For example, `subtaxa(obj, recursive = 3)` will return information for all subtaxa and their immediate subtaxa for each taxon. The recursive option also accepts a simple `TRUE`/`FALSE`, with `TRUE` indicating all subtaxa of subtaxa, etc., and `FALSE` only returning immediate subtaxa, but not their descendants. By default, `subtaxa` and `supertaxa` return taxon IDs, but the value option allows the user to choose what information to return for each taxon. For example, `subtaxa(obj, value = "taxon_names")` will return the names of taxa instead of their IDs. Any data available to NSE (i.e. in the result of `all_names(obj)`) can be returned in this way.

The functions `roots`, `stems`, `branches`, and `leaves` are a conceptual set of functions that return different subsets of a taxonomy. A "root" is any taxon that does not have a supertaxon. A "stem" is a root plus all subtaxa before the first split in the tree. A "branch" is any taxon that has only one subtaxon and one supertaxon. Stems and branches are useful to identify since they can be removed without losing information on the relative relationship among the remaining taxa. "Leaves" are taxa with no subtaxa. By default, these options return taxon IDs, but also have the value option like subtaxa and supertaxa, so they can return other information as well. For example, `leaves(obj, value = "taxon_names")` will return the names of taxa on the tips of the tree.

In the case of `taxmap` objects, the `obs` function returns information for observations associated with each taxon and its subtaxa. The observations could be rows in a table or elements in a list/vector that are named by taxon IDs. This is used

to easily map between user-supplied information and taxa. For example, assuming a taxonomy with a single root, the value returned by `obs` for the root taxon will contain information for all observations, since they will all be assigned to a subtaxon of the root taxon. By default, row/element indices of observations will be returned, but the `obs` function also accepts the value option, so the contents of any column or other information associated with taxa can be returned as well.

The parsers. Taxonomic data appear in many different forms depending on the source of the data, making parsing a challenge. There are two main sources of variation in how taxonomic data are typically stored: the type of information supplied (e.g. a taxon name vs. a taxon ID) and how it is encoded (e.g. in a table vs. as part of a string). In addition, there might be additional user-specific data associated with the taxa that need to be parsed. These data might be associated with each taxon in a classification (e.g the taxon ranks) or might be associated with each classification (e.g. a sequence ID). In many cases, both types are present. This complexity makes implementing a generic parser for all types of taxonomic data difficult, so parsers are typically only available for specific formats. The taxa package introduces a set of three parsing functions that can parse the vast majority of taxonomic data as well as any associated data and return a `taxmap` object.

The `parse_tax_data` function is used to parse taxonomic classifications stored as vectors in tables that have already been read into R. In the case of tables, the classification can be spread over multiple columns or in a single column with character separators (e.g. “Primates; Hominidae; Homo; sapiens”) or a combination of the two. Other columns are preserved in the output and the rows are mapped to

the taxon IDs (e.g. the ID assigned to "sapiens" in the above example). For both tables and vectors, additional lists, vectors or tables can be included and are assigned taxon IDs based on some shared attribute with the source of the taxonomic data (e.g. a shared element ID or the same order). This makes it possible to parse many data sets at once and have them all mapped to the same taxonomy in the resultant taxmap object. Data associated with each taxon in each classification can also be parsed and included in the output using regular expressions with capture groups identifying the information to be stored and a key corresponding to the capture groups that identifies what each piece of information is. For example, `"Hominidae_f_2;Homo_g_3;sapiens_s_4"` would use the separator `";"`, the regular expression `"(.+)_(.+)_(.+)"`, and the key `c(my_taxon = "taxon_name", my_rank = "taxon_rank", my_id = "info")`. The values of the key indicate what the information is (a taxon name and two arbitrary pieces of information) and the names of the key (e.g. `"my_rank"`) determine the names of columns in the output.

If only a taxon name (e.g. Primates) or a taxon ID for a reference database (e.g. the NCBI taxon ID for Homo sapiens is 180092) is available in a table or vector, then the classification information must be queried from online databases and the function `lookup_tax_data` is used. `lookup_tax_data` has all the same functionality of `parse_taxa_data` in addition to being able to look up taxonomic classifications associated with taxon names, taxon IDs, and NCBI sequence IDs. If the data are embedded in a string (e.g. a FASTA header), then the function `extract_tax_data` is used instead. `extract_tax_data` has the functionality of `parse_tax_data` and `lookup_tax_data`, except that the information is extracted from raw strings using a

regular expression and a corresponding key, the same way that data for each taxon in a classification is extracted by `parse_tax_data`. Together, these three parsing functions can handle every combination of data type and format presented in Figure 2.2 and many variations of those formats.

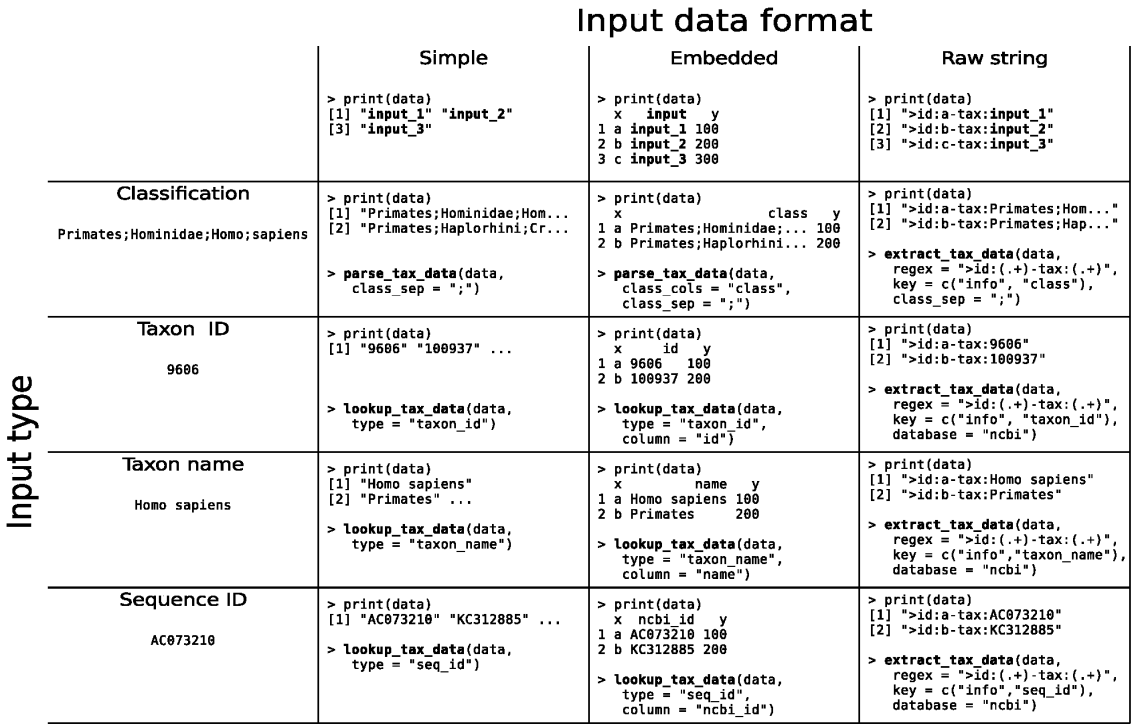

Figure 2.2: A table for determining how to parse different sources of taxonomic information using the `taxa` package. The rows correspond to the common sources of taxonomic information: full taxonomic classifications encoded in text, taxon IDs from a database, taxon names (a single rank), and NCBI sequence IDs. The columns correspond to the different formats the information can be encoded in: as a simple vector, as columns in a table, and as a piece of a complex string (e.g. a FASTA header). In the case of tables and complex strings, other information associated with the taxa can be preserved in the parsed result, as is done in the "use cases" example below. Any one cell in the table shows how to parse a given taxonomic information source in a given format using one of the three parsing functions: `parse_tax_data`, `lookup_tax_data`, `extract_tax_data`.

2.3.2 Operation

`Taxa` is an R package hosted on CRAN, so only an R installation and internet connection are needed to install and use `taxa`. Once installed, most of the functionality of the package can be used without an internet connection. R can be installed on nearly any operating system, including most UNIX systems, MacOS, and Windows. The minimum system requirements of R and the taxa package are easily met by most personal computers. The amount of resources needed will depend on the size of data being used and the complexity of analyses being conducted. The package can be installed by entering `install.packages("taxa")` in an interactive R session. The development version can be installed from GitHub using the devtools package:

```
library(devtools)
install_github("ropensci/taxa")
```

For users, the typical operation of the software will involve parsing some kind of input data into a taxmap object using a method demonstrated in Figure 2.2. Alternatively, a dependent package, such as `metacoder`, might provide a parser that wraps one of the taxa parsers or otherwise returns a `taxmap` object. Once the data is in a `taxmap` object, the majority of a user's interaction with the `taxa` package would typically involve filtering and manipulating the data using functions described in Table 2.1 and applying application-specific functions in other packages, such as `metacoder` (Figure 2.3).

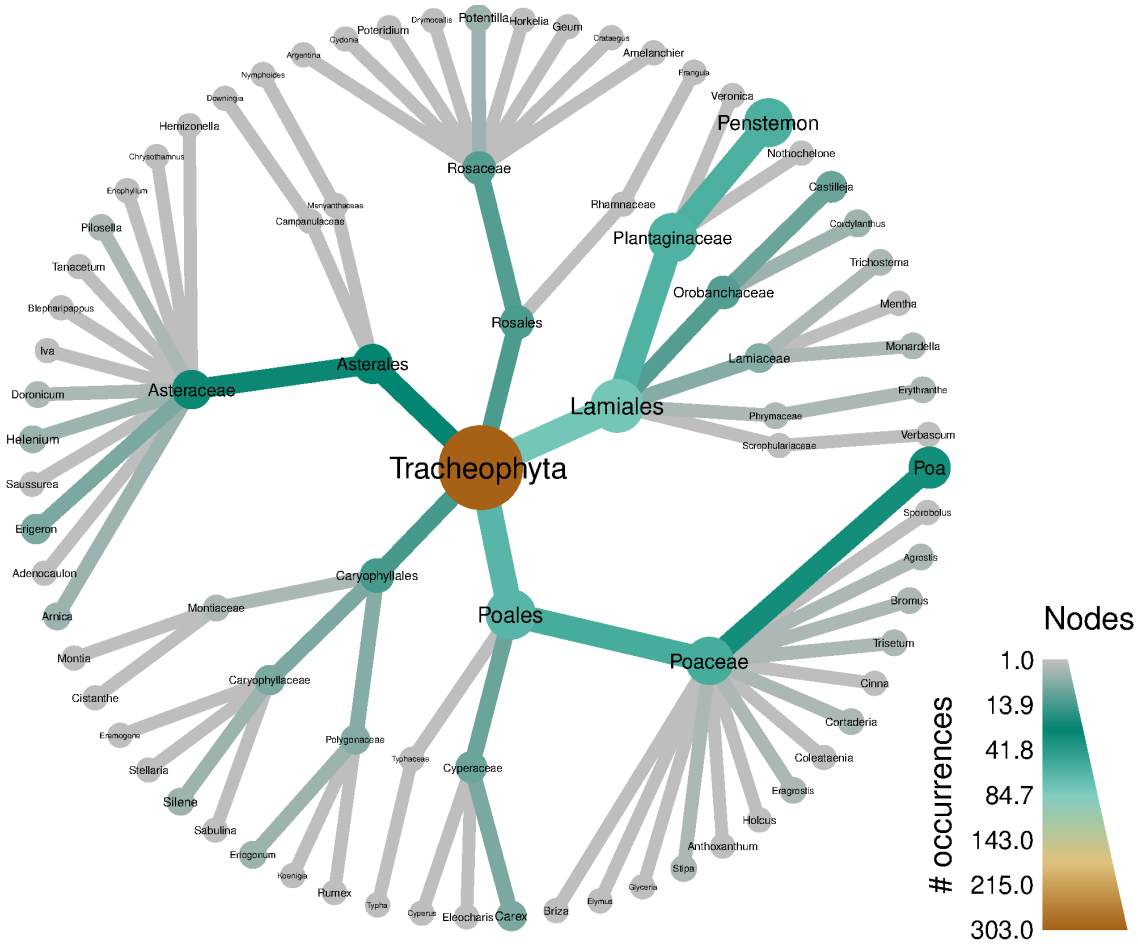

Figure 2.3: The result of the example analysis shown in the text. Records of plant species occurrences in Oregon are downloaded from the Global Biodiversity Information Facility (GBIF) using the `rgbif` package (Chamberlain, 2017). Then a taxa parser is used to parse the table of GBIF data into a `taxmap` object. A series of filters are then applied. First, all occurrences that are not from preserved specimens as well any taxa that have no occurrences from preserved specimens are removed. Then, all taxa at the species level are removed, but their occurrences are reassigned to the genus level. All taxa without names are then removed. In the final two filters, only orders within Tracheophyta with greater than 10 subtaxa are preserved. The `metacoder` package is then used to create a heat tree (i.e. taxonomic tree) with color and size used to display the number of occurrences associated with each taxon at each level of the hierarchy.

Table 2.1: Primary classes and functions found in `taxa`.

Function	Description
`taxon`	A class that combines the classes containing the name, rank, and ID for a taxon.
`taxa`	A simple list of taxon objects in an arbitrary order.
`hierarchy`	A class that stores a list of nested taxa constituting a classification.
`hierarchies`	A simple list of hierarchy objects in an arbitrary order.
`taxonomy`	A class that stores a list of unique taxon objects and a tree structure.
`taxmap`	A class that combines a taxonomy with user-defined, tables, lists, or vectors associated with taxa in the taxonomy. The taxonomic tree and the associated data can then be manipulated such that the two remain in sync.

Function	Description
`supertaxa`, `subtaxa`	A "supertaxon" is a taxon of a coarser rank that encompasses the taxon of interest (e.g. *Homo* is a supertaxon of *Homo sapiens*). The "subtaxa" of a taxon are all those of a finer rank encompassed by that taxon. For example, *Homo sapiens* is a subtaxon of *Homo*. The `supertaxa`/`subtaxa` function returns the supertaxa/subtaxa of all or a subset of the taxa in a taxonomy object. By default, these functions return taxon IDs, but they can also return any data associated with taxa.
`roots`, `leaves`, `stems`, `branches`	Roots are taxa that lack a supertaxon. Likewise, leaves are taxa that lack a subtaxon. Stems are those taxa from the roots to the first split in the tree. Branches are taxa with exactly one supertaxon and one subtaxon. In general, stems and branches can be filtered out without changing the relative relationship between the remaining taxa. By default, these functions return taxon IDs, but they can also return any data associated with taxa.

Function	Description
`obs`	Returns the information about every observation from an user-defined data set for each taxon and their subtaxa. By default, indices of a list, vector, or table mapped to taxa are returned.
`filter_taxa`, `filter_obs`	Subset taxa or associated data in taxmap objects based on arbitrary conditions. Hierarchical relationships among taxa and mappings between taxa and observations are taken into account.
`arrange_taxa`, `arrange_obs`	Order taxon or observation data in taxmap objects.
`sample_n_taxa`, `sample_n_obs`, `sample_frac_taxa`, `sample_frac_obs`	Randomly sample taxa or observation data in taxmap objects. Weights can be applied that take into account the taxonomic hierarchy and associated data. Hierarchical relationships among taxa and mappings between taxa and associated data are taken into account.

Since `taxa` provides highly flexible parsers, it is usually possible to convert data from other packages to `taxa` classes, enabling manipulation of that data by `taxa` functions or packages that build upon `taxa`, like `metacoder`. For example, using the general-use parsers provided by the `taxa` package, `metacoder` supplies specialized and easy to use parsers for the following formats: taxonomy files produced by mothur,

biom files produced by QIIME and MEGAN, newick files, objects from the `phyloseq` package, `phylo` objects from the `ape` package, and fasta files from the Greengenes (McDonald et al., 2012), RDP (Cole et al., 2014), SILVA (Yilmaz et al., 2014), and UNITE databases (Kõljalg et al., 2013). We have not encountered any text-based file format containing taxonomic information that can be described using regular expressions that the `taxa` parsers cannot read. For classes from other packages that inherit `list`, `vector`, or `data.frame`, conversion is not needed to include that information in a `taxmap` object, since the manipulation functions such as `filter_taxa` will handle them correctly as is.

2.4 Use case

`Taxa` is currently being used by `metacoder` and we are working on refactoring parts of `taxize` to work seamlessly with `taxa` as well. Both `taxize` and `metacoder` provide broadly useful functions such as querying databases with taxonomic information and plotting taxonomic information, respectively. We hope that having these two packages adopt the `taxa` framework will encourage developers of new packages to do so as well. Regardless, the flexible parsers implemented in `taxa` (Figure 2.2) allow for data from nearly any source to be used. The example analysis below uses data from the package `rgbif` (S. A. Chamberlain & Boettiger, 2017; S. Chamberlain et al., 2017), even though `rgbif` was not designed to work with `taxa`. This example shows a few of the benefits of using `taxa`. The function `occ_data` from the `rgbif` package returns a `data.frame` (i.e. table) of occurrence data for species from the Global Biodiversity

Information Facility (GBIF) with one row per occurrence. The table has one column per taxonomic rank from kingdom to species.

```
# Look up plant occurrence data for Oregon
library(rgbif)
occ <- rgbif::occ_data(stateProvince = "Oregon",
                       scientificName = "Plantae")
```

This format returned by `rgbif::occ_data` is a variant on the format described in Figure 2.2, row 1, column 2, except that there is only one rank per column instead of all ranks being concatenated in the same column (the parser accepts any number of columns, each of which could contain multiple ranks delineated by a separator).

```
# Parse data with taxa
library(taxa)
obj <- parse_tax_data(occ$data, class_cols = c(22:26, 28),
                      named_by_rank = TRUE)
```

In the `taxmap` object returned by `parse_tax_data`, the original table returned by `occ_data` is stored as `obj$data$tax_data`, but an extra column with taxon IDs for each row is prepended.

```
> print(obj)
<Taxmap>
626 taxa: aab. Plantae ... ayc. NA
626 edges: NA->aab, aab->aac ... aml->ayc
1 data sets:
  tax_data: # A tibble: 500 x 103
    taxon_id name          key    decimalLatitude
    <chr>  <chr>         <int>  <dbl>
    1 amm    Racomitriu... 1.70e9 44.2
    2 amn    Orthotrich... 1.68e9 NA
    3 amo    Didymodon ... 1.67e9 45.7
    # ... with 497 more rows, and 99 more
    # <<< List of additional columns ommited >>>
```

The data are then passed through a series of filters piped together. The `filter_obs` command removes rows from the occurrence data table not corresponding to preserved specimens, as well as any corresponding taxa that no longer have occurrences due to this filtering. The multiple calls to `filter_taxa` that follow demonstrate some of the different parameterizations of this powerful function. By default, `taxa` that don't pass the filter are simply removed and any occurrences assigned to them are reassigned to supertaxa that did pass the filter (e.g. occurrences for a deleted species would be assigned to the species' genus). When the `supertaxa` option is set to `TRUE`, all the supertaxa of taxa that pass the filter will also be preserved. The `subtaxa` option works the same way. Finally, the filtered data are passed to a plotting function from the `metacoder` package that accepts the `taxmap` format. The plot is a taxonomic tree with color and size used to display the number of occurrences associated with each taxon (Figure 2.3).

```
# Plot number of occurrences for each taxon
library(metacoder)
obj %>%
   filter_obs("tax_data", basisOfRecord == "PRESERVED_SPECIMEN",
              drop_taxa = TRUE) %>%
   filter_taxa(taxon_ranks != "specificEpithet") %>%
   filter_taxa(! is.na(taxon_names)) %>%
   filter_taxa(taxon_names == "Tracheophyta", subtaxa = TRUE) %>%
   filter_taxa(taxon_ranks == "order", n_subtaxa > 10,
               subtaxa = TRUE, supertaxa = TRUE) %>%
   heat_tree(node_label = taxon_names,
             node_color = n_obs, node_size = n_obs,
             node_color_axis_label = "# occurrences")
```

Note the use of columns in the original input table like "basisOfRecord" being used as if they were independent variables. This is implemented by NSE as a convenience

to users, but they could also have been included by typing the full path to the variable (e.g. `obj$data$tax_data$basisOfRecord` or `occ$data$basisOfRecord`). This is similar to the use of `taxon_ranks` and `taxon_names`, which are actually functions included in the class (e.g. `obj$taxon_ranks()`). The benefit of using NSE is that they are reevaluated each time their name is referenced. This means that the first time `taxon_ranks` is referenced in the example code it returns a different value than the second time it is referenced, because some taxa were filtered out. If `obj$taxon_ranks()` is used instead, it would fail on the second call because it would return information for taxa that have been filtered out already.

2.5 Conclusions

While `taxa` is useful on its own, its full potential will be realized after being adopted by the community as a standard for interacting with taxonomic information in R. A robust standard for the commonplace problems of data parsing and manipulation will free developers to focus on specific novel functionality. The `taxa` package already serves as the foundation of another package called `metacoder`, which provides functions for plotting taxonomic information and parsing common file formats used in metagenomics research. `Taxize`, the primary package for querying taxonomic information from internet sources, is also being refactored to be compatible with `taxa`. We hope the broadly useful functionality of these two packages will jump start adoption of `taxa` as the standard for taxonomic data manipulation in R.

2.6 Data and software availability

Install in R as `install.packages("taxa")`

Software available from: https://cran.r-project.org/web/packages/taxa/index.html

Source code available from: https://github.com/ropensci/taxa

Archived source code available from: https://doi.org/10.5281/zenodo.1183667

License: MIT

2.7 Funding statement

This work was supported in part by funds from USDA Agricultural Research Service Projects 2027-22000-039-00 and 2072-22000-039-15-S to NG and an rOpenSci grant to ZF.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Chapter 3: Metacoder: An R package for visualization and manipulation of community taxonomic diversity data

Zachary S. L. Foster, Thomas J. Sharpton, and Niklaus J. Grünwald

Published in:

PLoS computational biology, 2017, 13:2

DOI: 10.1371/journal.pcbi.1005404

3.1 Abstract

Community-level data, the type generated by an increasing number of metabarcoding studies, is often graphed as stacked bar charts or pie graphs that use color to represent taxa. These graph types do not convey the hierarchical structure of taxonomic classifications and are limited by the use of color for categories. As an alternative, we developed `metacoder`, an R package for easily parsing, manipulating, and graphing publication-ready plots of hierarchical data. `Metacoder` includes a dynamic and flexible function that can parse most text-based formats that contain taxonomic classifications, taxon names, taxon identifiers, or sequence identifiers. `Metacoder` can then subset, sample, and order this parsed data using a set of intuitive functions that take into account the hierarchical nature of the data. Finally, an extremely flexible plotting function enables quantitative representation of up to 4 arbitrary statistics simultaneously in a tree format by mapping statistics to the color and size of tree nodes and edges. `Metacoder` also allows exploration of barcode primer bias by integrating functions to run digital PCR. Although it has been designed for data from metabarcoding research, `metacoder` can easily be applied to any data that has a hierarchical component such as gene ontology or geographic location data. Our package complements currently available tools for community analysis and is provided open source with an extensive online user manual.

3.2 Introduction

Metabarcoding is revolutionizing our understanding of complex ecosystems by circumventing the traditional limits of microbial diversity assessment, which include the need and bias of culturability, the effects of cryptic diversity, and the reliance on expert identification. Metabarcoding is a technique for determining community composition that typically involves extracting environmental DNA, amplifying a gene shared by a taxonomic group of interest using PCR, sequencing the amplicons, and comparing the sequences to reference databases (Cristescu, 2014). It has been used extensively to explore communities inhabiting diverse environments, including oceans (De Vargas et al., 2015), plants (Coleman-Derr et al., 2016), animals (Douglas et al., 2012), humans (Huttenhower et al., 2012), and soil (Gilbert, Jansson, & Knight, 2014).

The complex community data produced by metabarcoding is challenging conventional graphing techniques. Most often, bar charts, stacked bar charts, or pie graphs are employed that use color to represent a small number of taxa at the same rank (e.g. phylum, class, etc). This reliance on color for categorical information limits the number of taxa that can be effectively displayed, so most published figures only show results at a coarse taxonomic rank (e.g. class) or for only the most abundant taxa. These graphing techniques do not convey the hierarchical nature of taxonomic classifications, potentially obscuring patterns in unexplored taxonomic ranks that might be more biologically important. More recently, tree-based visualizations are becoming available as exemplified by the python-based MetaPhlAn and the corresponding

graphing software GraPhlAn (Segata et al., 2012). This tool allows visualization of high-quality circular representations of taxonomic trees.

Here, we introduce the R package `metacoder` that is specifically designed to address some of these problems in metabarcoding-based community ecology, focusing on parsing and manipulation of hierarchical data and community visualization in R. `Metacoder` provides a visualization that we call "heat trees" which quantitatively depicts statistics associated with taxa, such as abundance, using the color and size of nodes and edges in a taxonomic tree. These heat trees are useful for evaluating taxonomic coverage, barcode bias, or displaying differences in taxon abundance between communities. To import and manipulate data, `metacoder` provides a means of extracting and parsing taxonomic information from text-based formats (e.g. reference database FASTA headers) and an intuitive set of functions for subsetting, sampling, and rearranging taxonomic data. `Metacoder` also allows exploration of barcode primer bias by integrating digital PCR, which simulates PCR success using alignments between reference sequences and primers. All this functionality is made intuitive and user-friendly while still allowing extensive customization and flexibility. `Metacoder` can be applied to any data that can be organized hierarchically such as gene ontology or geographic location. `Metacoder` is an open source project available on CRAN and is provided with comprehensive online documentation including examples.

3.3 Design and implementation

The R package `metacoder` provides a set of novel tools designed to parse, manipulate, and visualize community diversity data in a tree format using any taxonomic classification (Figure 3.1). Figure 3.1 illustrates the ease of use and flexibility of `metacoder`. It shows an example analysis extracting taxonomy from the 16S Ribosomal Database Project (RDP) training set for mothur (Schloss et al., 2009), filtering and sampling the data by both taxon and sequence characteristics, running digital PCR, and graphing the proportion of sequences amplified for each taxon. Table 3.1 provides an overview of the core functions available in `metacoder`.

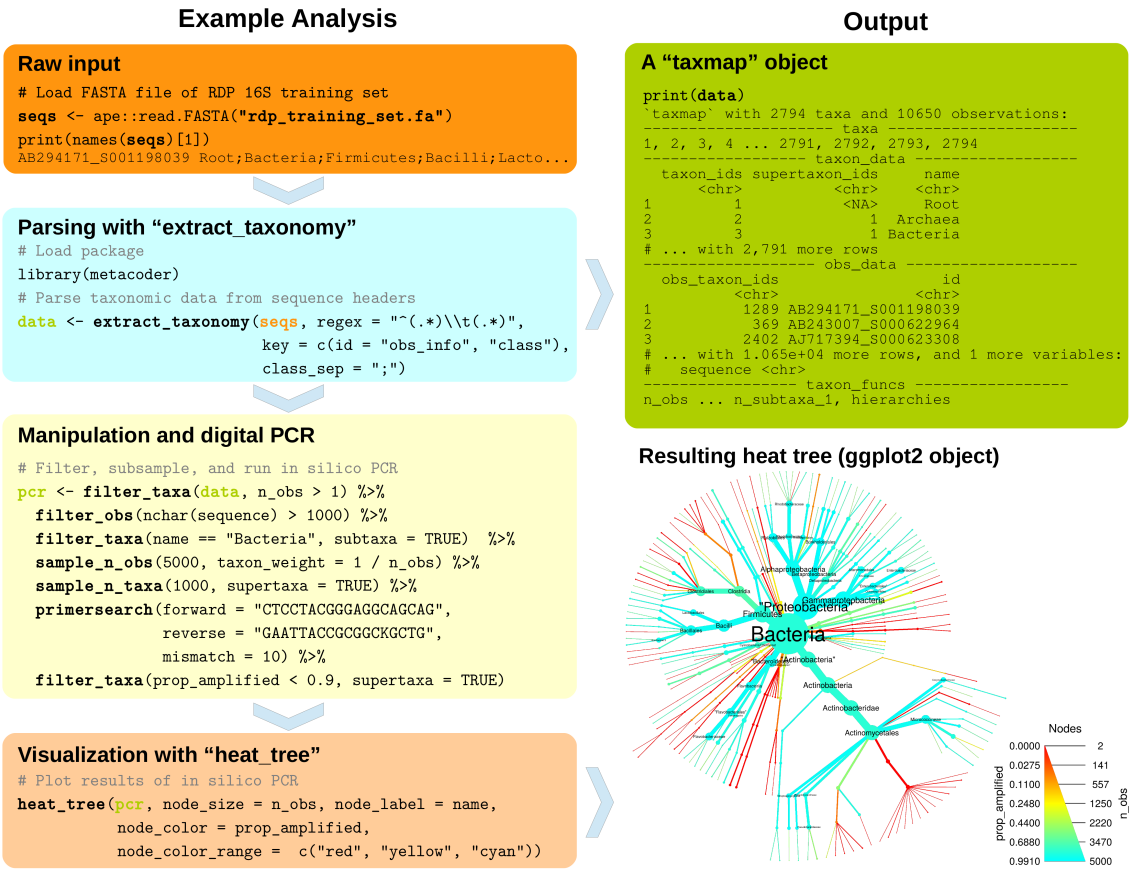

Figure 3.1: `Metacoder` has an intuitive and easy to use syntax. The code in this example analysis parses the taxonomic data associated with sequences from the Ribosomal Database Project 16S training set, filters and subsamples the data by sequence and taxon characteristics, conducts digital PCR, and displays the results as a heat tree. All functions in bold are from the `metacoder` package. Note how columns and functions in the taxmap object (green box) can be referenced within functions as if they were independent variables.

Table 3.1: Primary functions found in `metacoder`.

Function	Description
`extract_taxonomy`	Parses taxonomic data from arbitrary text and returns a `taxmap` object containing a table with rows corresponding to inputs (i.e. observations) and a table with rows corresponding to taxa.
`heat_tree`	Makes tree-based plots of data stored in `taxmap` objects. Color, size, and labels of tree components can be mapped to arbitrary data. The output is a `ggplot2` object.
`primersearch`	Executes the EMBOSS program primersearch on sequence data stored in a taxmap object. Results are parsed, added to the input taxmap object and returned.
`mutate_taxa`, `mutate_obs`, `transmute_taxa`, `transmute_obs`	Modify or add columns of taxon or observation data in `taxmap` objects. mutate * adds columns and transmute * returns only new columns.
`select_taxa`, `select_obs`	Subset columns of taxon or observation data in `taxmap` objects.

Function	Description
`filter_taxa`, `filter_obs`	Subset rows of taxon or observation data in `taxmap` objects based on arbitrary conditions. Hierarchical relationships among taxa and mappings between taxa and observations are taken into account.
`arrange_taxa`, `arrange_obs`	Order rows of taxon or observation data in `taxmap` objects.
`sample_n_taxa`, `sample_n_obs`, `sample_frac_taxa`, `sample_frac_obs`	Randomly subsample rows of taxon or observation data in `taxmap` objects. Weights can be applied that take into account the taxonomic hierarchy and associated observations. Hierarchical relationships among taxa and mappings between taxa and observations are taken into account.
`subtaxa`, `supertaxa`, `observations`, `roots`	Returns the indices of rows in taxon or observation data in `taxmap` objects. Used to map taxa to related taxa and observations.

3.3.1 The `taxmap` data object

To store the taxonomic hierarchy and associated observations (e.g. sequences) we developed a new data object class called `taxmap`. The `taxmap` class is designed to

be as flexible and easily manipulated as possible. The only assumption made about the user's data is that it can be represented as a set of observations assigned to a hierarchy; the hierarchy and the observations do not need to be biological. The class contains two tables in which user data is stored: a taxonomic hierarchy stored as an edge list of unique IDs and a set of observations mapped to that hierarchy (Figure 3.1). Users can add, remove, or reorder both columns and rows in either `taxmap` table using convenient functions included in the package (Table 3.1). For each table, there is also a list of included functions that create a temporary column with the same name when referenced by one of the manipulation or plotting functions. These are useful for attributes that must be updated when the data is subset or otherwise modified, such as the number of observations for each taxon (see `n_obs` in Figure 3.1). If this kind of derived information was stored in a static column, the user would have to update the column each time the data set is subset, potentially leading to mistakes if this is not done. There are many of these column-generating functions included by default, but the user can easily add their own by adding a function that takes a taxmap object. The names of columns or column-generating functions in either table of a taxmap object can be referenced as if they were independent variables in most `metacoder` functions in the style of popular R packages like `ggplot2` and `dplyr`. This makes the code much easier to read and write.

3.3.2 Universal parsing and retrieval of taxonomic information

`Metacoder` provides a way to extract taxonomic information from text-based formats so it can be manipulated within R. One of the most inefficient steps in bioinformatics can be loading and parsing data into a standardized form that is usable for computational analysis. Many databases have unique taxonomy formats with differing types of taxonomic information. The taxonomic structure and nomenclature used can be unique to the database or reference another database such as GenBank (Benson, Cavanaugh, Clark, Karsch, & DJ, 2013). Rather than creating a parser for each data format, `metacoder` provides a single function to parse any format definable by regular expressions that contains taxonomic information (Figure 3.1). This makes it easier to use multiple data sources with the same downstream analysis.

The `extract_taxonomy` function can parse hierarchical classifications or retrieve classifications from online databases using taxon names, taxon IDs, or Genbank sequence IDs. The user supplies a regular expression with capture groups (parentheses) and a corresponding key to define what parts of the input can provide classification information. The `extract_taxonomy` function has been used successfully to parse several major database formats including Genbank (Benson et al., 2013), UNITE (Kõljalg et al., 2013), Protist Ribosomal Reference Database (PR2) (Guillou et al., 2012), Greengenes (DeSantis et al., 2006), Silva (Quast et al., 2012), and, as illustrated in Figure 3.1, the RDP (Maidak et al., 1996). Examples for each database are provided in the user manuals.

3.3.3 Intuitive manipulation of taxonomic data

`Metacoder` makes it easy to subset and sample large data sets composed of thousands of observations (e.g. sequences) assigned to thousands of taxa, while taking into account hierarchical relationships. This allows for exploration and analysis of manageable subsets of a large data set. Taxonomies are inherently hierarchical, making them difficult to subset and sample intuitively compared with typical tabular data. In addition to the taxonomy itself, there is usually also data assigned to taxa in the taxonomy, which we refer to as "observations". Subsetting either the taxonomy or the associated observations, depending on the goal, might require subsetting both to keep them in sync. For example, if a set of taxa are removed or left out of a random subsample, should the subtaxa and associated observations also be removed, left as is, or reassigned to a supertaxon? If observations are removed, should the taxa they were assigned to also be removed? The functions provided by `metacoder` gives the user control over these details and simplifies their implementation.

`Metacoder` allows users to intuitively and efficiently subset complex hierarchical data sets using a cohesive set of functions inspired by the popular `dplyr` data-manipulation philosophy. `Dplyr` is an R package for providing a conceptually consistent set of operations for manipulating tabular information (Wickham et al., 2015). Whereas `dplyr` functions each act on a single table, `metacoder`'s analogous functions act on both the taxon and observation tables in a `taxmap` object (Table 3.1). For each major `dplyr` function there are two analogous `metacoder` functions: one that manipulates the taxon table and one that manipulates the observations table. The

functions take into account the relationship between the two tables and can modify both depending on parameterization, allowing for operations on taxa to affect their corresponding observations and vice versa. They also take into account the hierarchical nature of the taxon table. For example, the `metacoder` functions `filter_taxa` and `filter_obs` are based on the `dplyr` function filter and are used to remove rows in the taxon and observation tables corresponding to some criterion. Unlike simply applying a filter to these tables directly, these functions allow the subtaxa, supertaxa, and/or observations of taxa passing the filter to be preserved or discarded, making it easy to subset the data in diverse ways (Figure 3.1). There are also functions for ordering rows (`arrange_taxa`, `arrange_obs`), subsetting columns (`select_taxa`, `select_obs`), and adding columns (`mutate_taxa`, `mutate_obs`).

`Metacoder` also provides functions for random sampling of taxa and corresponding observations. The function `taxonomic_sample` is used to randomly sub-sample items such that all taxa of one or more given ranks have some specified number of observations representing them. Taxa with too few sequences are excluded and taxa with too many are randomly subsampled. Whole taxa can also be sampled based on the number of sub-taxa they have. Alternatively, there are `dplyr` analogues called `sample_n_taxa` and `sample_n_obs`, which can sample some number of taxa or observations. In both functions, weights can be assigned to taxa or observations, influencing how likely each is to be sampled. For example, the probability of sampling a given observation can be determined by a taxon characteristic, such as the number of observations assigned to that taxon, or it could be determined by an observation characteristic, like sequence length. Similar to the `filter_*` functions, there are pa-

rameters controlling whether selected taxa's subtaxa, supertaxa, or observations are included or not in the sample (Figure 3.1).

3.3.4 Heat tree plotting of taxonomic data

Visualizing the massive data sets being generated by modern sequencing of complex ecosystems is typically done using traditional stacked barcharts or pie graphs, but these ignore the hierarchical nature of taxonomic classifications and their reliance on colors for categories limits the number of taxa that can be distinguished (Figure 3.2). Generic trees can convey a taxonomic hierarchy, but displaying how statistics are distributed throughout the tree, including internal taxa, is difficult. `Metacoder` provides a function that plots up to 4 statistics on a tree with quantitative legends by automatically mapping any set of numbers to the color and width of nodes and edges. The size and content of edge and node labels can also be mapped to custom values. These publication-quality graphs provide a method for visualizing community data that is richer than is currently possible with stacked bar charts. Although there are other R packages that can plot variables on trees, like `phyloseq` (McMurdie & Holmes, 2013), these have been designed for phylogenetic rather than taxonomic trees and are therefore optimized for plotting information on the tips of the tree and not on internal nodes. There is also a set of python scripts called GraPhlAn that can make similar tree-based visualizations. GraPhlAn has better annotation abilities than `metacoder`, supports edge length for phylogenetic trees, and can plot a variety of node shapes. However, `metacoder`'s heat tree function can plot multiple trees per graph,

use different layout algorithms, automatically transform raw data to color/size for quantitative display with a scale bar, and optimize the size range of nodes to avoid crowded or sparse graphs.

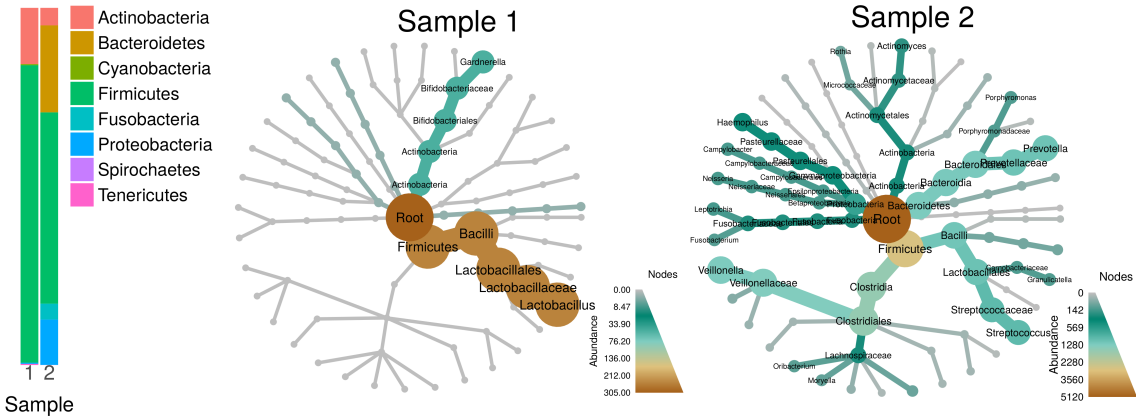

Figure 3.2: Heat trees allow for a better understanding of community structure than stacked bar charts. The stacked bar chart on the left represents the abundance of organisms in two samples from the Human Microbiome Project. The same data are displayed as heat trees on the right. In the heat trees, size and color of nodes and edges are correlated with the abundance of organisms in each community. Both visualizations show communities dominated by firmicutes, but the heat trees reveal that the two samples share no families within firmicutes and are thus much more different than suggested by the stacked bar chart.

The function `heat_tree` creates a tree utilizing color and size to display taxon statistics (e.g., sequence abundance) for many taxa and ranks in one intuitive graph (Figure 3.2). Taxa are represented as nodes and both color and size are used to represent any statistic associated with taxa, such as abundance. Although the `heat_tree` function has many options to customize the appearance of the graph, it is designed to minimize the amount of user-defined parameters necessary to create an effective

visualization. The size range of graph elements is optimized for each graph to minimize overlap and maximize size range. Raw statistics are automatically translated to size and color and a legend is added to display the relationship. Unlike most other plotting functions in R, the plot looks the same regardless of output size, allowing the graph to be saved at any size or used in complex, composite figures without changing parameters. These characteristics allow `heat_tree` to be used effectively in pipelines and with minimal parameterization since a small set of parameters displays diverse taxonomy data. The output of the `heat_tree` function is a `ggplot2` object, making it compatible with many existing R tools. Another novel feature of heat trees is the automatic plotting of multiple trees when there are multiple "roots" to the hierarchy. This can happen when, for example, there are "Bacteria" and "Eukaryota" taxa without a unifying "Life" taxon, or when coarse taxonomic ranks are removed to aid in the visualization of large data sets (Figure 3.3).

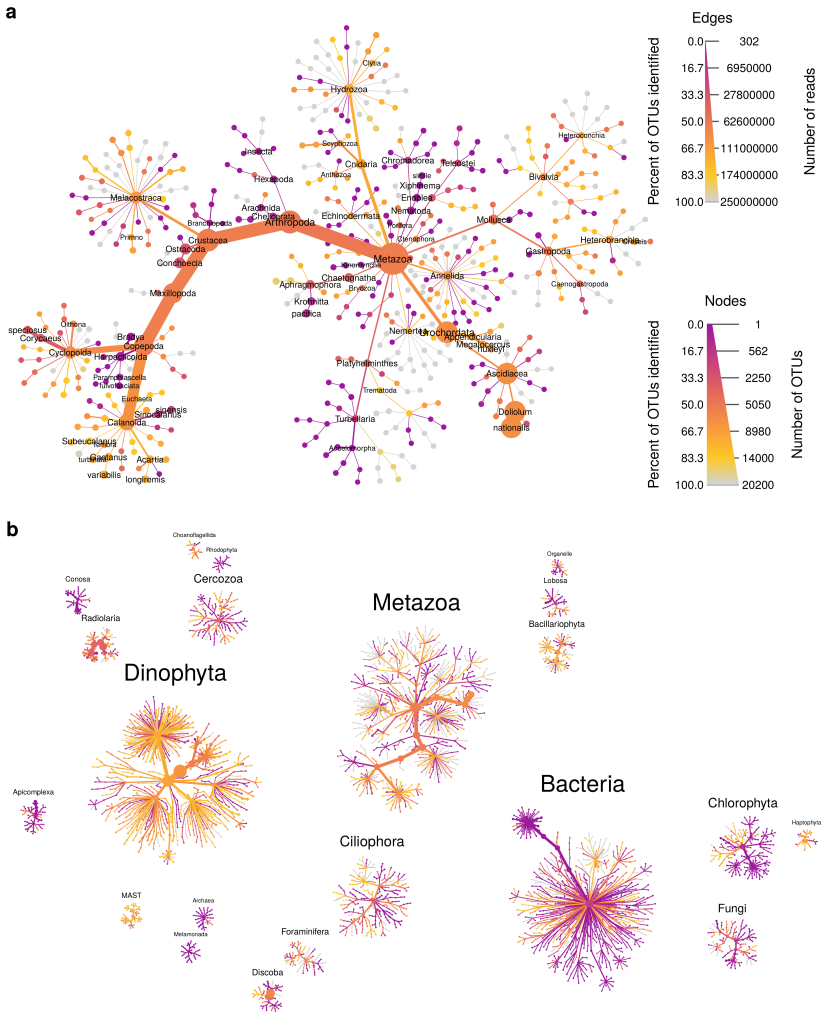

Figure 3.3: Heat trees display up to four metrics in a taxonomic context and can plot multiple trees per graph. Most graph components, such as the size and color of text, nodes, and edges, can be automatically mapped to arbitrary numbers, allowing for a quantitative representation of multiple statistics simultaneously. This graph depicts the uncertainty of OTU classifications from the TARA global oceans survey. Each node represents a taxon used to classify OTUs and the edges determine where it fits in the overall taxonomic hierarchy. Node diameter is proportional to the number of OTUs classified as that taxon and edge width is proportional to the number of reads. Color represents the percent of OTUs assigned to each taxon that are somewhat similar to their closest reference sequence (>90% sequence identity). a. Metazoan diversity in detail. b. All taxonomic diversity found. Note that multiple trees are automatically created and arranged when there are multiple roots to the taxonomy.

3.4 Results

3.4.1 Heat trees allow quantitative visualization of community diversity data

We developed heat trees to allow visualization of community data in a taxonomic context by mapping any statistic to the color or size of tree components. Here, we reanalyzed data set 5 from the TARA oceans eukaryotic plankton diversity study to visualize the similarity between OTUs observed in the data set and their closest match to a sequence in a reference database (De Vargas et al., 2015). The TARA ocean expedition analyzed DNA extracted from ocean water throughout the world. Even though a custom reference database was made using curated 18S sequences spanning all known eukaryotic diversity, many of the OTUs observed had no close match. Figure 3.3 shows a heat tree that illustrates the proportion of OTUs that were well characterized in each taxon (at least 90% identical to a reference sequence). Color indicates the percentage of OTUs that are well characterized, node width indicates the number of OTUs assigned to each taxon, and edge width indicates the number of reads. Taxa with ambiguous names and those with less than 300 reads have been filtered out for clarity. This figure illustrates one of the principal advantages of heat trees, as it reveals many clades in the tree that contain only purple and orange lineages, which indicate that the entire taxonomic group is poorly represented in the reference sequence database. Of particular interest are those clades with predominantly purple and orange lineages that also have relatively large nodes, such as Harpacticoida (in Copepoda on the left). These represent taxonomic groups that were found to have high amounts of diversity in the oceans, but for which we have

a paucity of genomic information. Investigators interested in improving the genomic resolution of the biosphere can thus use these approaches to rapidly assess which taxa should be prioritized for focused investigations.

3.4.2 Flexible parsing allows for similar use of diverse data

Metabarcoding studies often rely on techniques or data that may introduce bias into an investigation. For example, the specific set of PCR primers used to amplify genomic DNA and the taxonomic annotation database can both have an effect on the study results. A quick and inexpensive way to estimate biases caused by primers is to use digital PCR. `Metacoder` can be used to explore different databases or primer combinations to assess these effects since it supplies functions to parse divserse data sources, conduct digital PCR, and plot the results. Figure 3.4 shows a series of heat tree comparisons that were produced using a common 16S rRNA metabarcoding primer set and digital PCR against the full-length 16S sequences found in three taxonomic annotation databases: Greengenes (DeSantis et al., 2006), RDP (Maidak et al., 1996), and SILVA (Yilmaz et al., 2014). These heat trees reveal subsets of the full taxonomies for these three databases that poorly amplify by digital PCR using the selected primers. As a result, they indicate which lineages within each of the taxonomies may be challenging to detect in a metabarcoding study that uses these primers. Importantly, different sets of primers likely amplify different sets of taxa, so investigators interested in specific lineages can use this approach in conjunction with various primer sets to identify those that maximize the likelihood of discovery and re-

duce wasted sequencing resources on non-target organisms. However, these heat trees do not indicate whether one database is necessarily preferable over another, as they differ in the structure of their taxonomies, as well as the number and phylogenetic diversity of their reference sequences. For example, most of the bacterial clades that do not amplify well in the SILVA lineages are unnamed lineages that are not found in the other databases, indicating that they warrant further exploration.

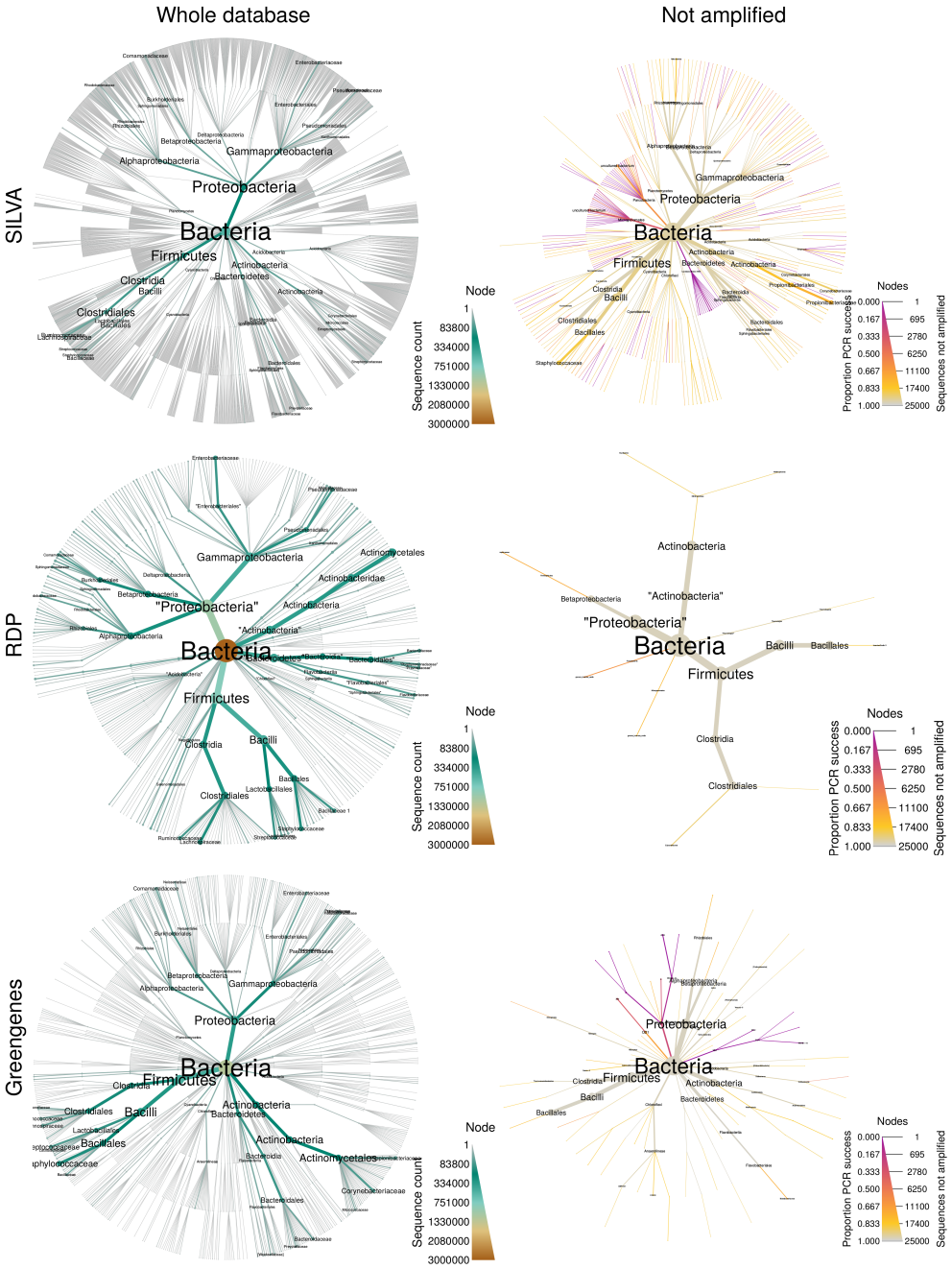

Figure 3.4: Flexible parsing and digital PCR allows for comparisons of primers and databases. Shown is a comparison of digital PCR results for three 16S reference databases. The plots on the left display abundance of all bacterial 16S sequences. Plots on the right display all taxa with subtaxa not entirely amplified by digital PCR using universal 16S primers. Node color and size display the proportion and number of sequences not amplified respectively.

3.4.3 Heat trees can show pairwise comparisons of communities across treatments

One challenge in metabarcoding studies is visually determining how specific sub-sets of samples vary in their taxonomic composition. Unlike most other graphing software in R, `metacoder` produces graphs that look the same at any output size or aspect ratio, allowing heat trees to be easily integrated into larger composite figures without changing the code for individual subplots. Using color to depict the difference in read or OTU abundance between two treatments can result in particularly effective visualizations, especially when the presence of color is made dependent on a statistical test. To examine more than two treatments at once, a matrix of these kind of heat trees can be combined with a labeled "guide" tree. Figure 3.5 shows application of this idea to human microbiome data showing pairwise differences between body sites. Coloring indicates significant differences between the median proportion of reads for samples from different body sites as determined using a Wilcox rank-sum test followed by a Benjamini-Hochberg (FDR) correction for multiple testing. The intensity of the color is relative to the log-2 ratio of difference in median proportions. Brown taxa indicate an enrichment in body sites listed on the top of the graph and green is the opposite. While the original study (Huttenhower et al., 2012) showed abundance plots, our visualization provides the taxonomic context. For example, *Haemophilus*, *Streptococcus*, and *Prevotella* spp. are enriched in saliva (brown) relative to stool where *Bacteroides* is enriched (green). We also see that in the Lachnospiraceae clade several genera shown in both green and brown taxa are differentially abundant. These observations are consistent with known differences in the human-associated microbiome

across body sites, but heat trees uniquely provide an integrated view of how all levels of a taxonomy vary for all pairs of body sites.

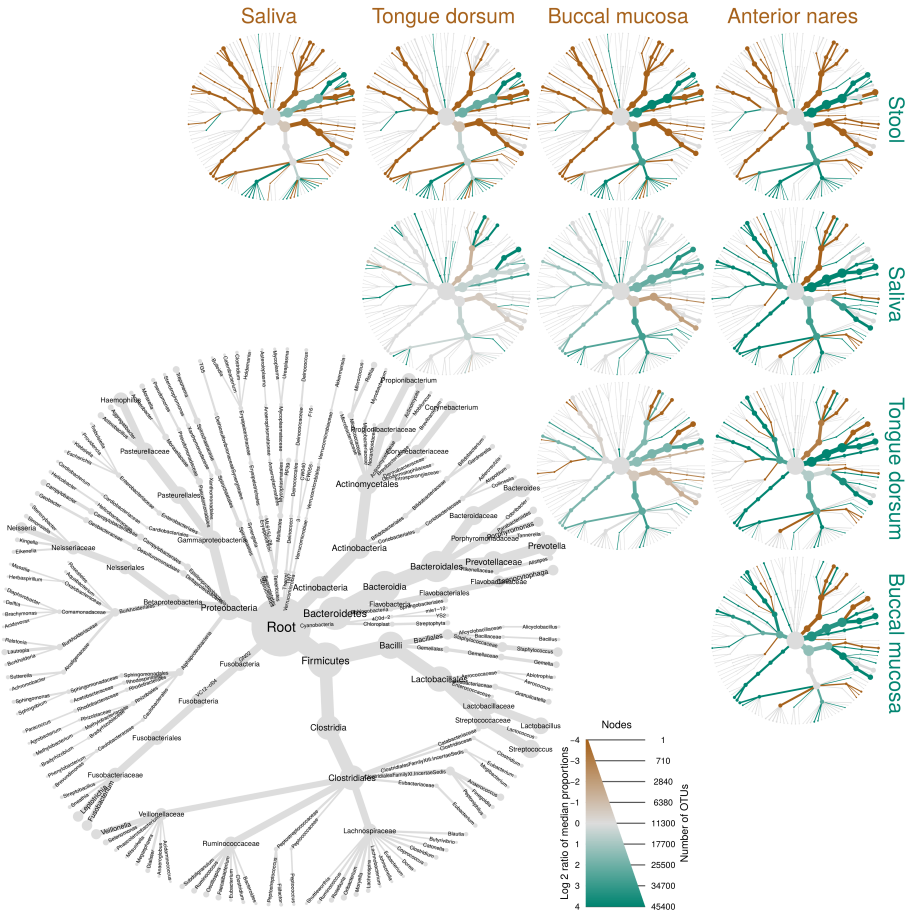

Figure 3.5: Scale-independent appearance facilitates complex, composite figures. This graph uses 16S metabarcoding data from the human microbiome project study to show pairwise comparisons of microbiome communities in different parts of the human body. All graph components, including text, have the same relative sizes independent of output size, unlike most graphical packages in R, making it easier to create composite figures entirely within R. The gray tree on the lower left functions as a key for the smaller unlabeled trees. The color of each taxon represents the log-2 ratio of median proportions of reads observed at each body site. Only significant differences are colored, determined using a Wilcox rank-sum test followed by a Benjamini-Hochberg (FDR) correction for multiple comparisons. Taxa colored green are enriched in the part of the body shown in the row and those colored brown are enriched in the part of the body shown in the column. For example, *Haemophilus*, *Streptococcus*, *Prevotella* are enriched in saliva (brown) relative to stool where *Bacteroides* is enriched (green).

3.4.4 Other applications

The `taxmap` data object defined in `metacoder` can be used for any data that can be classified by a hierarchy. Figure 3.6, for example, shows an analysis of votes cast in the 2016 US Democratic party national primaries organized by geography. The heat tree reveals distinct patterns such as a sweep by Clinton in the South and a split on the West coast, with California predominantly voting for Clinton while Washington and Oregon predominantly voted for Sanders. Another potential application is displaying the results of gene expression studies by associating differential expression with gene ontology (GO) annotations. Figure 3.7 shows the results of a RNA-seq study on the effect of glucocorticoids on smooth muscle tissue (Himes et al., 2014). All biological processes influenced by at least one gene with a significant change in expression are plotted. The authors of the study find that genes involved in immune response are influenced by the glucocorticoid treatment. Viewing these results in a heat tree shows not only the specific immune process affected (the branch on the middle right), but also the more general phenomena they constitute; regulation of high level phenomena, like immune system function, can be explained by specific processes like "lymphocyte homeostasis" and these specific processes are put into the context of the phenomena they contribute to. This is more informative than simply reporting the results for a single level of the GO annotation hierarchy or discussing the effects of genes one at a time.

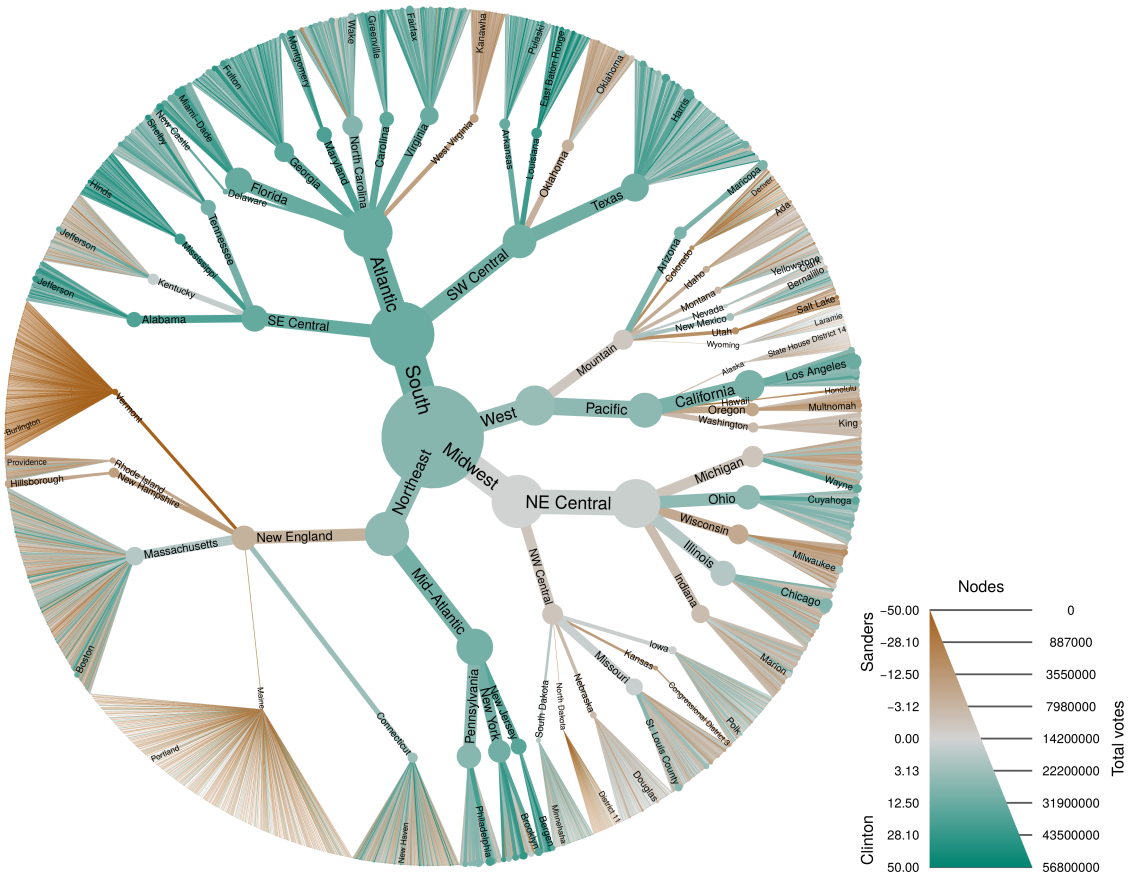

Figure 3.6: Metacoder can be used with any type of data that can be organized hierarchically. This plot shows the results of the 2016 Democratic primary election organized by region, division, state, and county. The regions and divisions are those defined by the United States census bureau. Color corresponds to the difference in the percentage of votes for candidates Hillary Clinton (green) and Bernie Sanders (brown). Size corresponds to the total number of votes cast. Data was downloaded from https://www.kaggle.com/benhamner/2016-us-election/.

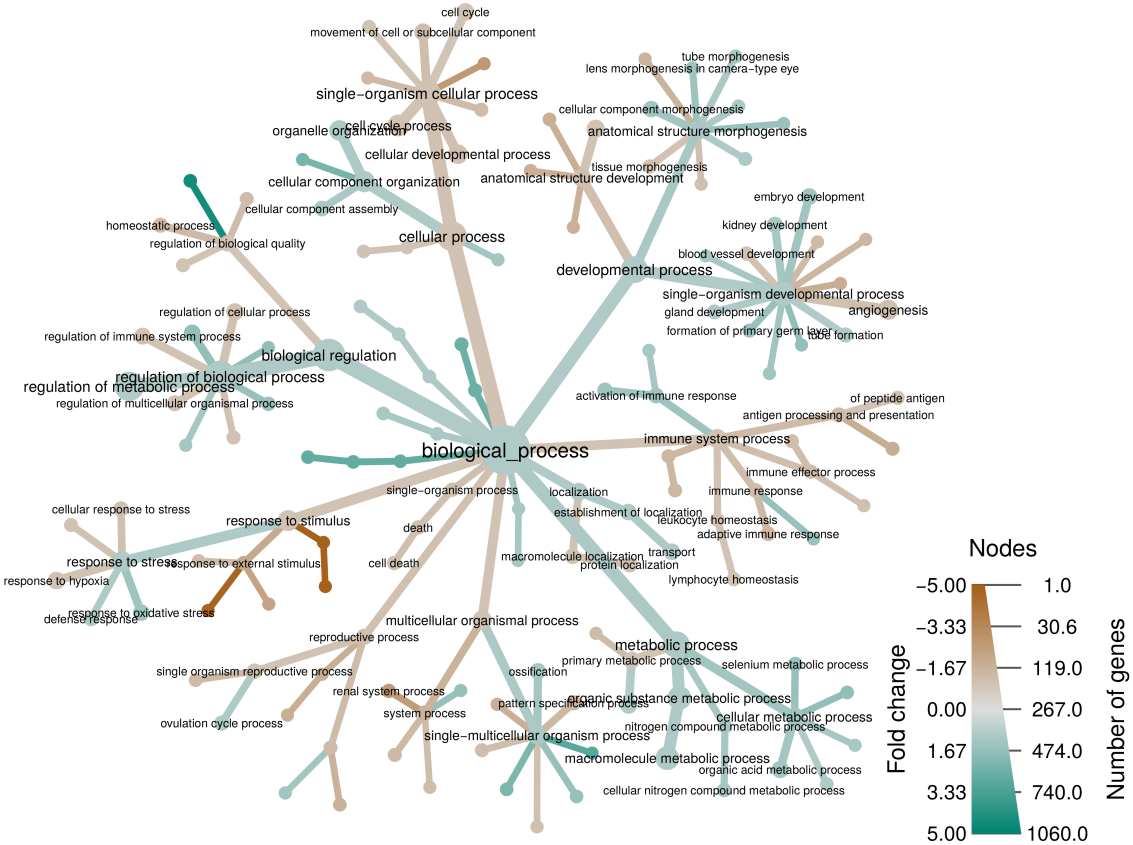

Figure 3.7: Another alternate use example: Visualizing gene expression data in a GO hierarchy. The gene ontology for all differentially expressed genes in a study on the effect of a glucocorticoid on airway smooth muscle tissue. Color indicates the sign and intensity of averaged changes in gene expression and the size indicates the number of genes classified by a given gene ontology term.

3.5 Availability and future directions

The R package `metacoder` is an open-source project under the MIT License. Stable releases of `metacoder` are available on CRAN while recent improvements can be

downloaded from github (https://github.com/grunwaldlab/metacoder). A manual with documentation and examples is provided. This manual also provides the code to reproduce all figures included in this manuscript.

We are currently continuing development of `metacoder`. We welcome contributions and feedback from the community. We want to make `metacoder` functions and classes compatible with those from other bioinformatic R packages such as `phyloseq`, `ape` (Paradis, Claude, & Strimmer, 2004), `seqinr` (Charif & Lobry, 2007), and `taxize` (Chamberlain & Szöcs, 2013). We might integrate more options for digital PCR and barcode gap analysis, perhaps using ecoPCR (Ficetola et al., 2010) or the R packages `PrimerMiner` (Elbrecht & Leese, 2017) and `Spider` (Brown et al., 2012). We are also considering adding additional visualization functions.

Chapter 4: The composition of the fungal and oomycete microbiome of *Rhododendron* roots under varying growth conditions, nurseries, and cultivars

Zachary S. L. Foster, Jerry E. Weiland, Carolyn F. Scagel, and Niklaus J. Grünwald

Published in:

Phytobiomes, 2020, 4:2

DOI: 10.1094/PBIOMES-09-19-0052-R

4.1 Abstract

The microbiome of agricultural crops influences processes such as nutrient absorption, drought stress, and susceptibility to pathogens. Interactions between a plant's genotype and its environment influence the composition of the microbiome, but these interactions are not well understood. We compared how the fungal and oomycete microbiomes of rhododendrons from Oregon nurseries differed among cultivars, growth conditions, and nurseries. Roots were sampled from randomly selected container and field-grown plants of three cultivars of *Rhododendron* at four nurseries. The internal transcribed spacer 1 (ITS1) barcode was sequenced with the Illumina MiSeq using two sets of primers specific to fungi and oomycetes, respectively. Sequences were used to infer community composition using VSEARCH and a custom reference database combining curated fungal and oomycete sequences. Comparisons of diversity and community composition were conducted in R using the `vegan` and `metacoder` packages. Organism lifestyle was inferred using the FUNGuild database. Few oomycetes were found and fungal communities were dominated by saprobes and mutualists. Nurseries that grew plants in containers and in-field had a significantly higher diversity of fungi than those that only grew plants in containers. Microbiome composition differed significantly among growth conditions and nurseries, but not among cultivars. This suggests that, among these cultivars of *Rhododendron*, environment is important in structuring the root microbiome, but cultivar is not.

4.2 Introduction

Given major advances in microbiome research, it might soon become possible to engineer the microbiome of plants to improve nutrient absorption and health. Plants influence the composition and function of their microbiome by selectively recruiting a subset of the microbes from the surrounding environment that tend to provide benefits to the host (Berendsen, Pieterse, & Bakker, 2012). Plant genotype can affect this process, influencing the composition of the microbiome (Bálint et al., 2013; Panke-Buisse, Poole, Goodrich, Ley, & Kao-Kniffin, 2015; Wagner et al., 2016). Different cultivars of maize are known to respond differently to inoculation with the nitrogen-fixing bacteria *Azospirillum*; in some cultivars, addition of *Azospirillum* is equivalent to 100 kg ha-1 of nitrogen, whereas other cultivars are unaffected (Salamone, Döbereiner, Urquiaga, & Boddey, 1996). Different cultivars of wheat and grape have been observed to harbor distinct microbiomes (Bokulich, Thorngate, Richardson, & Mills, 2014; Sapkota & Nicolaisen, 2015). For fermentation substrates like grapes, the microbiome has additional relevance due to its effects on the sensory qualities of the finished product (Swiegers, Bartowsky, Henschke, & Pretorius, 2005). Common garden experiments of wild plants also suggest that the fungal microbiome differs among genotypes of the same species (Bálint et al., 2013; Wagner et al., 2016). There is evidence that microbiome differences in natural stands of European beech correlate more with genotypic differences than with geographical distance (Cordier, Robin, Capdevielle, Desprez-Loustau, & Vacher, 2012). Agricultural management practices have a major influence on the plant environment, which determines which microbes

can be recruited and which pathogens will be present. For example, how a nursery recycles irrigation water could influence the community of oomycete pathogens (Redekar, Eberhart, & Parke, 2019). Given a mechanistic understanding of the factors required to assemble a microbiome and achieve certain functions, one can envision a future where direct management of the plant microbiome can improve crop production. However, before a microbiome can be managed it has to be well characterized including how different hosts, cultivars, and production conditions can affect the microbiome.

High-throughput sequencing allows for rapid, affordable, and comprehensive characterization of the diversity found in microbial communities compared with sequencing by cloning or culturing (Ji et al., 2013). For example, obligate symbionts and pathogens that have important implications for agriculture are difficult to culture and are therefore less likely to be detected by traditional culture-based microbial surveys (Yarza et al., 2014). New techniques, such as metabarcoding (i.e., amplicon metagenomics) and shotgun metagenomics, rely only on sequencing mixtures of PCR products or raw genomic DNA derived from environmental samples of complex communities, such as soil, and are therefore less biased by organismal lifestyle. Metabarcoding is a particularly powerful technique that involves extracting DNA from complex samples, amplifying a common barcode gene with PCR, and sequencing amplicons using high-throughput sequencing (Taberlet et al., 2012). Sequences can then be used to estimate community diversity and compared with reference databases to estimate community composition (Cole et al., 2009). However, metabarcoding also has its own set of biases, such as differential PCR efficiency and limited taxonomic

resolution (Nichols et al., 2018). Recently, there have been attempts to use metabarcoding to characterize the communities and the distribution of pathogenic organisms, such as *Phytophthora* spp. in agricultural settings (Prigigallo et al., 2016; Riddell et al., 2019).

Rhododendron is a major ornamental crop in the Pacific Northwest and is known to host both mycorrhizal symbionts and plant pathogens (Farr, Esteban, & Palm, 1996; Knaus, Fieland, Graham, & Grünwald, 2015; Parke & Grünwald, 2012; Parke, Knaus, Fieland, Lewis, & Grünwald, 2014). The nursery and greenhouse industry is a leading agricultural sector in the Pacific Northwest, with gross sales in Oregon of $948 million in 2017 and rhododendrons are one of the leading ornamental plants sold. Plants in the family Ericaceae, including *Rhododendron*, are known to form a distinct type of mutualism with fungi known as ericoid mycorrhizae. Ericoid mycorrhizae are known to aid their host in nutrient absorption and survival in poor and polluted soils (Cairney & Meharg, 2003). Rhododendrons in nurseries are also known to be vectors of the sudden death pathogen *Phytophthora ramorum* (Gruenwald, Goss, & Press, 2008; Werres et al., 2001) and other oomycetes such as *Phytophthora plurivora* (Weiland et al., 2018). The presence of well-known mutualists and multiple pathogens combined with its economic importance make *Rhododendron* a good system to study the effects of management and cultivar on phytobiome composition (Jones, Benson, & others, 2001).

Root pathogens and symbionts described on *rhododendron* include both fungi and oomycetes. Oomycete pathogens generally cause root rot or damping off. Commonly found oomycetes on *Rhododendron* include *Phytophthora plurivora*, *Phytophthora cin-*

namomi, and *Pythium cryptoirregulare* (Weiland et al., 2018). Fungal pathogens include *Cylindrocladium scoparium*, *Cylindrocladium theae*, *Rhizoctonia solani*, *Armillaria mellea*, and *Thielaviopsis basicola* (Dreistadt, 2001; Jones et al., 2001). Common ericoid mycorrhizal fungi isolated from ericaceous plants include ascomycetes such as *Rhizoscyphus ericae*, *Oidiodendron maius*, and dark septate endophytes, such as *Phialocephala fortinii*. Some basidiomycetes also occur, including *Clavaria* and members of the order Sebacinales (Dighton & Coleman, 1992; Vohnı'k & Albrechtová, 2011; Vohnı'k, Albrechtová, Vosátka, & others, 2005).

The goal of this study is to characterize the fungal and oomycete root microbiome of *Rhododendron* in Oregon nurseries and determine what factors might influence its composition. Therefore, we evaluated the relative importance of host genotype, environment, and management practices on structuring *Rhododendron* microbiomes in the rhizosphere using high-throughput sequencing. Specifically, we used metabarcoding to survey fungal and oomycete rhizosphere communities from three cultivars in four nurseries. We tested the hypothesis that microbiomes would differ among cultivars, nurseries, and production systems (container versus field-grown). We also expected to detect well-known *Rhododendron* pathogens, such as *Phytophthora cinnamomi* or *Phytophthora plurivora*, and well-known *Rhododendron* symbionts, such as *Rhizoscyphus ericae* or *Phialocephala fortinii* (Jones et al., 2001; Knaus et al., 2015; Parke et al., 2014; Weiland et al., 2018). Our work provides novel insights into the makeup of oomycete and fungal communities in *Rhododendron* roots and is one of the few studies to apply Illumina sequencing to oomycete metabarcoding.

4.3 Materials and methods

4.3.1 Sample collection

The rhizosphere of *Rhododendron* was sampled to determine the effect of nursery, production system (container versus field-grown plants), and cultivar. Plants that appeared healthy were sampled in four nurseries from three cultivars in the Willamette Valley, Oregon, United States during May 2014. The nurseries varied in size and management practices. Nurseries A and B grew plants in both field and container systems while nurseries C and D grew plants only in containers. Both field-grown and container-grown plants were sampled. *Rhododendron* cultivars Nova Zembla (RHS 58), Roseum Elegans (RHS 58), and PJM (ARS874) were sampled in all nurseries and in both growing conditions. Five plants were sampled at random from each combination of nursery, cultivar, and production system (i.e., field-grown versus container-grown). Two nurseries only grew the selected cultivars in containers, so field-grown plants were only sampled from the other two nurseries. Each sample consisted of four equally-sized subsamples, each roughly 50 cm^3, taken from opposite sides of the root ball. Root balls were sampled by hand using sterile gloves and transported to the lab on ice. In the case of field-grown plants, a hand trowel was used to expose the root ball on four sides and this trowel was sterilized between plants by rinsing in distilled water and 10% bleach.

4.3.2 Sample processing

Root ball samples were broken apart by hand using sterile gloves and excess dirt or potting media was shaken off to obtain rhizosphere samples. Gloves were changed between samples to avoid cross-contamination. Each frozen rhizosphere sample was ground by hand using a clean mortar and pestle with liquid nitrogen. Mortars and pestles were autoclaved for 1 h and soaked in 10% bleach for at least 4 h between uses (Prince & Andrus, 1992). Ground rhizosphere samples were stored at -80°C in prechilled Falcon tubes (Corning, MA) until DNA extraction of approximately 150 mg of the sample using the Fast DNA Spin Kit (MP Biomedicals, Santa Ana, CA). The samples were not allowed to thaw at any time during this process.

4.3.3 PCR

We characterized both fungal and oomycete diversity using two internal transcribed spacer 1 (ITS1) primer pairs specific to each group. The fungal PCR used primers ITS1F (5' **TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG** *NC* AAACTTGGTCATTTAGAGGAAGTAA 3') and ITS2 (5' **GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG** GCTGCGTTCTTCATCGATGC 3') (White, Bruns, Lee, Taylor, & others, 1990). The bold sequences are the Illumina adapters and the italic sequence is a spacer added to increase the annealing temperature to what is recommended in the Illumina 16S sample preparation guide (Illumina Inc., San Diego, CA). The spacer was designed to be complementary to fungal DNA based on alignments of the primer to a random selection of sequences

downloaded from GenBank. Each reaction consisted of 1× PCR buffer, 0.2 mM dNTP mixture, 1 µM of each primer, 0.15 µl of GeneScript Taq polymerase (GenScript, Piscataway, NJ), and 2 µl of template DNA extract in a total volume of 15 µl. The thermocycler profile was 3 min at 94°C, followed by 30 cycles of 30 s at 94°C, 45 s at 60°C, and 1 min at 72°C with a final elongation for 7 min at 72°C. The oomycete PCR was seminested and used ITS6 (GAAGGTGAAGTCGTAACAAGG) and ITS4 (TCCTCCGCTTATTGATATGC) without the MiSeq adapters followed by ITS6 (TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGKGAAGGTGAAGTCGTAACAAGG) and ITS7 (GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGCGTTCTTCATCGATGTGC) with the MiSeq adapters (Sapkota et al., 2015). The reactions for the first PCR consisted of 1× PCR buffer, 1.5 mM MgCl2, 0.2 mM dNTP mixture, 0.2 µM of each primer, 0.04 units Platinum Taq polymerase (ThermoFisher Scientific, Waltham, MA), and 7.5 µl of template DNA extract in a total volume of 15 µl. The thermocycler profile was 2 min at 94°C, followed by 25 cycles of 30 s at 94°C, 30 s at 60°C, and 1 min at 72°C with a final elongation for 2 min at 72°C. The second PCR had the same reaction composition and thermocycler profile, except that 1.5 µl of a 1:10 dilution of the first PCR was used as the template.

4.3.4 Sequencing

The pooled oomycete and fungal libraries were sequenced on the Illumina MiSeq (Illumina) at the Center for Genome Research and Biocomputing (CGRB) at Oregon State University. Fungal and oomycete PCR products were mixed in equal volumes

and used by the CGRB Core Lab to create sequencing libraries. In brief, the mixture of PCR products was cleaned with AMPure XP beads (Beckman Coulter, Brea, CA). Sample indexes were added using the Illumina Nextera XT kit and cleaned again with AMPure XP beads. The samples were then quantified using an Agilent Bioanalyzer (Agilent, Palo Alto, CA), diluted to the same concentration, and pooled. A total of 25% PhiX was added to the pooled samples to correct for low diversity sequence bias. This library was then run on the CGRB's Illumina MiSeq using 250 bp paired-end sequencing.

4.3.5 Data analysis

An abundance matrix of operational taxonomic units (OTUs) versus samples was generated in order to compare the diversity and composition of the different cultivars, nurseries, and management practices. Primers and low-quality sequences were removed with `cutadapt` (Martin, 2011). Any sequences with greater than 10% mismatch to the primers or more than two "N" ambiguity codes were filtered out. All sequences from the ends of each read that had a phred score of less than 20 were removed. VSEARCH (Rognes, Flouri, Nichols, Quince, & Mahé, 2016) was then used to merge forward and reverse reads. Pairs with less than 15 bp overlap or more than three mismatches were filtered out. Unique reads were found using the `derep_fulllength` command and singletons were filtered out. Predicted chimeric sequences were filtered out using the `uchime_denovo` command. The remaining sequences were then clustered into OTUs using the `cluster_size` command at 99 and

97% to approximate species-scale differences for oomycetes and fungi respectively. Finally, the `usearch_global` command was used to assign taxonomic classifications to OTUs by comparing them to a custom reference database composed of the UNITE database (Kõljalg et al., 2005) for fungi, and Phytophthora-ID (Grünwald et al., 2011), Phytophthora-DB (Park et al., 2008), and the sequences from Robideau et al. (2011) for oomycetes. OTUs with less than 10 reads were filtered out. The taxonomy assigned to OTUs was used to infer organism lifestyle using FUNGuild (Nguyen et al., 2016). The FUNGuild results presented are those for OTUs that had at least a 97% identity with the reference sequence supplying the taxonomy and that were considered "Probable" or "Highly probably" by FUNGuild.

The diversity of communities was compared with alpha (diversity within a sample) and beta (the compositional dissimilarity between two samples) diversity statistics and ordination techniques using the R packages `vegan` (Dixon, 2003), `taxa` (Foster et al., 2018), and `metacoder` (Foster et al., 2017). The inverse Simpson index was calculated for each sample as a measure of alpha diversity and the Bray-Curtis index was calculated for each pair of samples as a measure of beta diversity. Differences in alpha diversity among factors were determined using a Tukey's honest significant difference (HSD) test following analysis of variance (ANOVA). Differences in beta diversity were visualized using nonmetric multidimensional scaling (NMDS) using the `metaMDS` function from `vegan` (Kruskal, 1964). Permutational multivariate analysis of variance (PERMANOVA), as implemented by the adonis function in the `vegan` package, was used to test which factors (cultivars, nurseries, and management practices) might be important for explaining differences in beta diversity (Anderson,

2001). PERMANOVA is a nonparametric approach allowing partitioning of variance among factors, analogous to a factorial ANOVA. All factors and interactions were included in the model.

To test for taxa with differential abundance between factors, nonparametric Wilcoxon rank-sum tests (Mann & Whitney, 1947) with a false discovery rate (FDR) correction for multiple tests were performed on the median read proportions for each taxon, at all taxonomic ranks. The results were visualized with differential heat trees using metacoder. For experimental factors that had more than two types (i.e., cultivar and nursery), differential abundance tests were conducted for each pairwise combination of types.

4.3.6 Data availability

All supplementary materials including R scripts to reproduce the analysis and figures, FASTQ files, and the OTU abundance matrix were deposited at Open Science Framework. The raw MiSeq sequences are also deposited in NCBI's Sequence Read Archive (BioProject PRJNA561631).

4.4 Results

4.4.1 Sequencing

Raw reads were filtered and grouped into OTUs to simulate species-level differences. After processing the raw sequences with VSEARCH, a total of 3,120,565 reads were

assigned to OTUs. At the primer removal step, approximately 6.3% of the reads were derived from the oomycete primers and 93.7% of the reads were derived from the fungal primers. After quality filtering, 3.6% of the reads were assigned to oomycetes and 96.4% of the reads were assigned to fungi. The number of raw reads per sample ranged from 8,869 to 82,015. There were 1,915 OTUs in the raw data and 731 remaining after removing OTUs with fewer than 10 reads. The number of OTUs per sample ranged from 68 to 259 for raw counts and from 29 to 98 for filtered counts.

4.4.2 Alpha diversity

Differences in alpha diversity (i.e., the diversity within a sample) were determined using ANOVA followed by a Turkey's HSD test. Overall, differences in alpha diversity among factors were small or not significant (Figure 4.1). There were minor, but significant, differences in diversity among the three cultivars. Nursery B had significantly higher alpha diversity than the others. Although there was no difference in alpha diversity between container-grown and field-grown plants, the species diversity of container-grown plants in nurseries that also grew plants in-field was greater than from container-grown plants in nurseries that only grew rhododendrons in containers ($P < 0.05$). The same analyses were done at the OTU and genus levels with similar results.

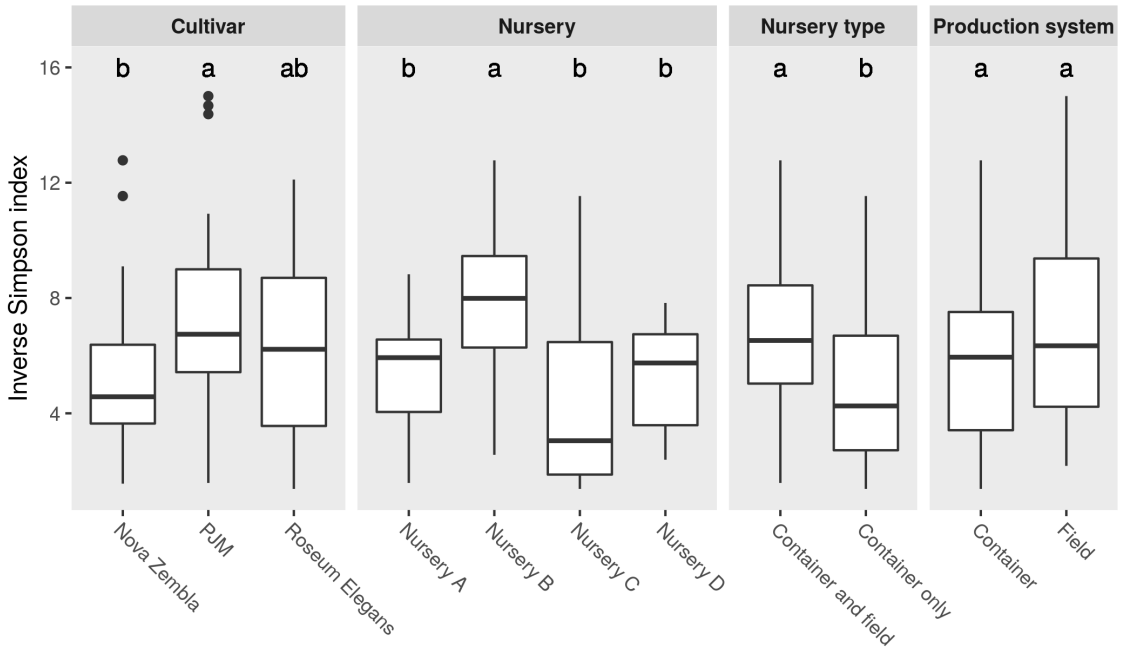

Figure 4.1: Alpha diversity of the combined fungal and oomycete species in the *Rhododendron* rhizosphere in different cultivars (Nova Zembla, PJM, and Roseum Elegans), nurseries (A, B, C, and D), nursery types (plant grown in containers and field soil versus nurseries growing only in container soil), and production systems (container versus field-grown). Letters represent significantly different distributions as determined by analysis of variance followed by a Tukey's honest significant difference test.

4.4.3 Beta diversity

Differences in rhizosphere community composition correlated with differences in nursery and production system, but not with differences in cultivar. NMDS revealed distinct communities associated with each combination of production system and

nursery when using two (Figure 4.2) or three dimensions. Production systems (e.g., container versus field-grown) separated communities based on axis NMDS1, whereas axis NMDS2 separated nurseries (Figure 4.2). There was little, if any, clustering associated with cultivar. PERMANOVA supported the NMDS results, showing highly significant associations between community similarity and production system ($R2 = 0.149$; $P < 0.001$) or nursery ($R2 = 0.234$; $P < 0.001$) (Table 4.1). Cultivar was also significantly correlated, but the effect size was small ($R2 = 0.025$; $P = 0.001$). The same analysis on other levels of the taxonomic hierarchy (e.g., OTU, species, genus, and family) showed similar results.

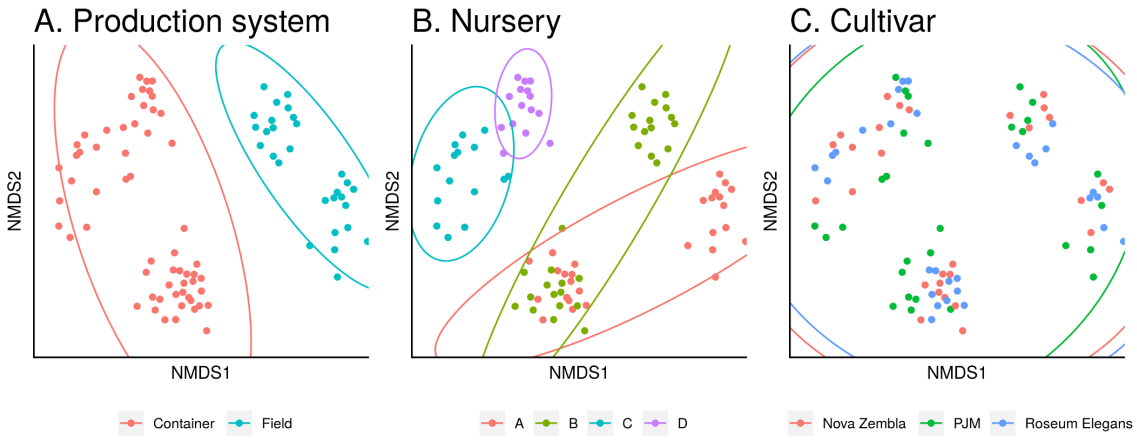

Figure 4.2: Two-dimensional nonmetric multidimensional scaling of Bray-Curtis distances between samples based on operational taxonomic unit relative abundance for both oomycetes and fungi. Ellipses represent 95% confidence intervals, assuming a multivariate t-distribution. Samples colored by A, container versus field-grown *Rhododendron*, B, the four nurseries sampled, and C, the three *Rhododendron* cultivars sampled.

Table 4.1: Results of permutational multivariate analysis of variance (PERMANOVA) of Bray-Curtis distances between samples.

Factor	R2	P
Cultivar	0.025	0.0012
Production system	0.149	0.0001
Nursery	0.234	0.0001
Cultivar × production system	0.025	0.0007
Cultivar × nursery	0.076	0.0001
Production system × nursery	0.045	0.0001
Cultivar × production system × nursery	0.028	0.0003
Residuals	0.418	-

Only the comparison between container and field-grown samples indicated significant differences in median read abundance for numerous taxa, as determined by Wilcoxon rank sum tests followed by a false discovery rate correction (Figure 4.3). Taxa with a greater proportion of reads assigned to them in field-grown samples than container-grown samples include: *Phialocephala fortinii*, *Cladophialophora chaetospira*, *Galerina atkinsoniana*, *Solicoccozyma terrea*, and *Trichoderma crassum*. Taxa with a higher read abundance in container-grown samples than in field-grown samples include: *Coniochaeta lignicola*, *Lecythophora fasciculata*, *Pleurostoma richardsiae*, *Sporothrix lignivora*, and *Exophiala hetermorpha*. No significant differences between experimental factors were observed for any oomycete taxa, probably due to the infrequency of detecting individual oomycete taxa relative to the number

of samples for each factor.

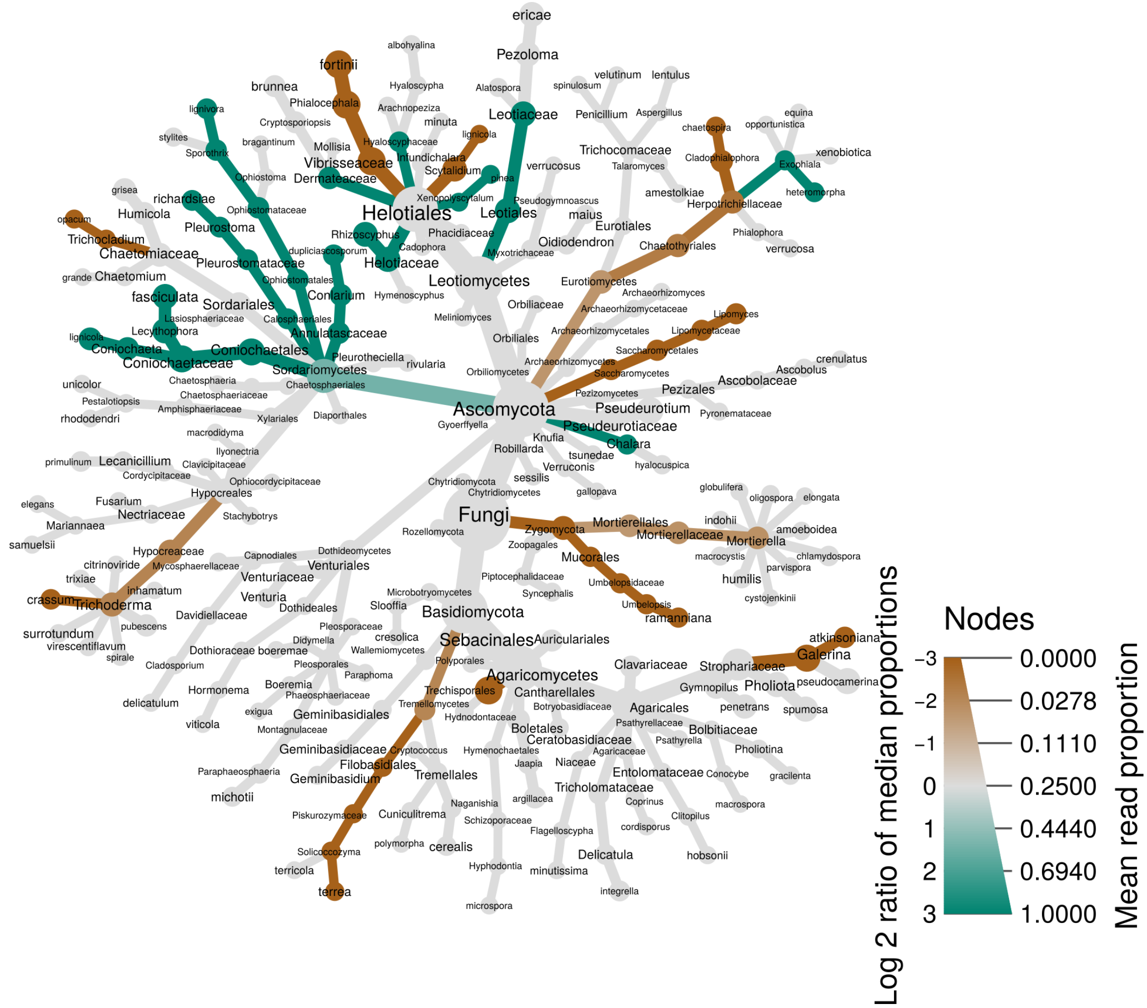

Figure 4.3: Differential heat tree (i.e., taxonomic tree) showing significant differences in median read proportion between fungal operational taxonomic units found in the rhizosphere from container-grown (green) versus field-grown (brown) *Rhododendron*. Significance was determined using Wilcox rank-sum tests with a false discovery rate correction for multiple comparisons. Only taxa that appear in at least five samples are shown.

4.4.4 Organismal diversity

Saprobes and beneficial fungi dominated the fungal and oomycete communities in *rhododendron* roots, according to FUNGuild results. Most sequence reads were assigned to saprobes and beneficial organisms. Pathogens were much less common than either saprobic or beneficial fungi. The three most common putative mutualists were OTUs matching reference sequences (>99.5% identity) for *Lecythophora fasciculata*, *Trichoderma pubescens*, and *Phialocephala fortinii*, appearing in 74, 68, and 53% of samples, respectively. The 10 most common saprobes occurred individually in 38 to 18% of samples and included OTUs that matched reference sequences (>99.5% identity) of *Xenopolyscytalum pinea*, *Humicola grisea*, *Cladophialophora chaetospira*, *Scytalidium lignicola*, and *Trichocladium opacum*. We also found taxa typically associated with ericaceous plants to be common, including taxa in the genera *Rhizoscyphus*, *Meliniomyces*, *Oidiodendron*, *Pezoloma*, *Hymenoscyphus*, *Phialocephala*, and the order Sebacinales (Figure 4.4). *Trichoderma* (some with purported biocontrol properties) and the saprophytic genera *Galerina* and *Mortierella* were also common and diverse.

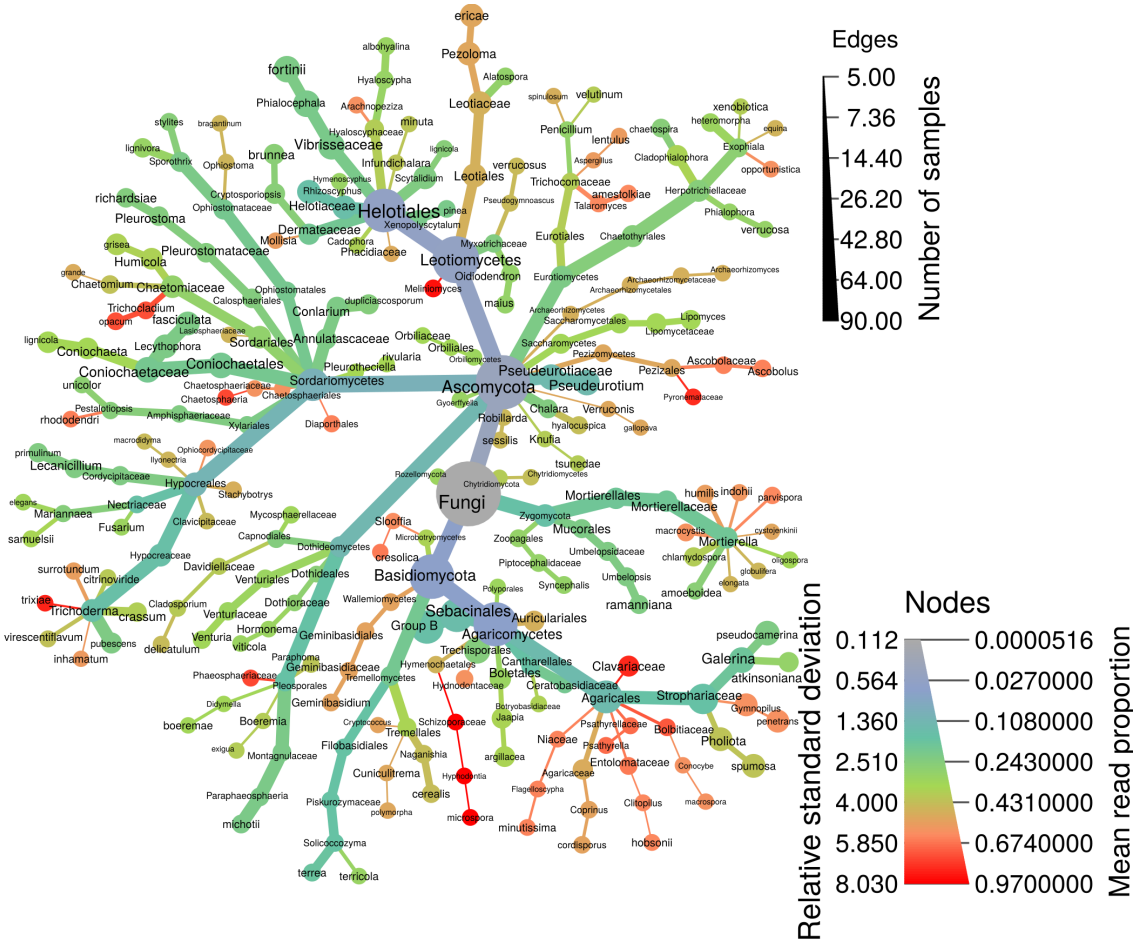

Figure 4.4: A heat tree (i.e., taxonomic tree) of fungal taxa with unambiguous classifications found in at least five samples. Edge width is proportional to the number of samples a given taxon was found in. The color and size of nodes is the relative standard deviation of operational taxonomic unit read proportions and the mean read proportion, respectively. Cooler colors indicate less variation between samples and hotter colors indicate more.

Putative plant pathogens were generally much less common than saprobes or mutualists. Some potential fungal pathogens found include *Pestalotiopsis unicolor*, *Pestalotiopsis rhododendri*, and *Ophiostoma bragantinum*, which were found in 23,

10, and 8% of samples, respectively. The genus *Microbotryum* was also found. There were much fewer OTUs assigned to oomycetes than fungi and these OTUs occurred in much fewer samples than the common fungal OTUs. Due to the nature of oomycete ITS1 sequences, OTUs that differ by more than 1% in sequence identity from their closest reference sequences are very likely different species than the one assigned and even exact matches do not always resolve individual species (Redekar et al., 2019). The reference sequences that were matched well (>99% identity) by the most common oomycete OTUs include *Pythium irregulare*, *Pythium sylvaticum*, *Phytophthora cactorum*, *Phytophthora citricola*, and *Phytophthora infestans*. These putative pathogens appeared in 11 to 2% of samples. Less frequently, sequences matching *Pythium macrosporum* and *Pythium dissotocum* were found. Other OTUs found had no close match to a reference sequence or occurred in only one sample. These likely represent either technical error or species not in the reference database.

4.5 Discussion

The environment of plants has a strong influence on the composition, and therefore function, of the microbiome (Bonito et al., 2014; Wagner et al., 2016). The plants sampled in this study differed by nursery and production system (container versus field-grown), both of which constitute different environments that are expected to change the composition of associated microbiomes. Of all the factors tested, production system and nursery were the two best predictors of differences in the root microbiome according to NMDS (Figure 4.2) and PERMANOVA (Table 4.1). The

effect of production system is likely due in part to the difference in growing media: container-grown plants were grown in nearly 100% Douglas fir bark chips whereas field-grown plants were grown in soil. Surprisingly, there was no significant difference in alpha diversity between the two types of production systems (Figure 4.1). We expected that field soil would support a more diverse community of fungi than a potting mix largely composed of bark and that would be reflected in the diversity of the rhizosphere community. However, it could be that rhododendrons select a subset of the bulk soil community, as happens in other plants (Uroz, Buée, Murat, Frey-Klett, & Martin, 2010), and that both environments have more diversity than is selected, even if the potting media is less diverse. Since we did not analyze bulk soil samples, we cannot confirm this. It could also be that the soil in the fields had relatively little organic matter and thus supported less saprotrophic diversity than the potting mix that was almost entirely organic matter. The composition of rhizosphere communities also varied among nurseries (Figure 4.2), even though production system was the dominant effect. For example, the container-grown plants from nurseries A and B clustered together in NMDS ordination, as to a lesser degree did the field-grown plants. This suggests that there could be microbiomes characteristic of container-grown and field-grown plants regardless of where they are grown, even though the location also has an effect. If this is true, it would mean that cultivation practices can have predictable effects on the microbiome of *Rhododendron* and that there is potential for optimizing these practices to maximize the benefits received from the microbiome. However, a study of a much greater number of nurseries, representative of a larger geographical area and greater diversity of cultivation practices, would be

needed to verify that this result applies to nursery-grown *Rhododendron* in general.

In many agricultural plant species, slight genotypic differences influencing interactions with the plant's microbiome can have a large impact on plant productivity (Salamone et al., 1996). Surprisingly, we observed little, if any, differences between the microbiome compositions of the three cultivars of *Rhododendron* sampled (Figure 4.2). This suggests that either the cultivars we selected are not very distinct genotypically or that their differences are not relevant for structuring the microbiome under the conditions sampled. If the microbiomes of other cultivars of *Rhododendron* are similarly indistinct, attempts at breeding for beneficial microbial communities will likely be ineffective. A study of the perennial herb *Boechera stricta* revealed an association between genotype and leaf microbiome but no association between genotype and root microbiome, although a genotype by environment interaction was significant (Wagner et al., 2016). In our results, interactions between environmental factors (production system and nursery) and cultivar were also more significant than cultivar alone according to PERMANOVA, but the effect size was small relative to the environmental factors alone (Table 4.1). In addition, the relative effects of environment and genotype can be different for bacteria and fungi (Bonito et al., 2014), so it is possible that bacterial communities, which we did not characterize, have a stronger association with genotype.

Saprobes and symbionts are the most ubiquitous organisms observed in the *Rhododendron* rhizosphere. Some of the most commonly observed organisms, *Humicola grisea* and *Cladophialophora chaetospira*, have been isolated from *Rhododendron* in other studies and *Trichoderma opacum* has been isolated from other ericaceous

plants (Bruzone, Fehrer, Fontenla, & Vohnı'k, 2017; Kowalik, Kierpiec-Baran, & Duda-Franiak, 2015; Vano, Sakamoto, Inubushi, & others, 2011). *Cladophialophora chaetospira* has been observed to form intracellular structures in root cells of *Rhododendron*, similar to ericoid mycorrhizae (Vano et al., 2011). *Scytalidium lignicola* is an anamorphic ascomycete generally found in wood or compost and is related to pathogens of Citrus and Manihot (Büttner, Gebauer, Hofrichter, Liers, & Kellner, 2018). The common ericoid mycorrhizal fungus *Pezoloma ericae* (aka *Hymenoscyphus ericae*) was found often (Vrålstad, Schumacher, & Taylor, 2002). Ericoid mycorrhizal fungi, such as *Pezoloma ericae*, have been previously reported on *Rhododendron* and as root endosymbionts shown to aid in the breakdown of organic debris, releasing mineral forms of nutrients that are available to the plant host (Smith & Read, 2010). The dark septate endophyte *Phialocephala fortinii* was also common and is known to aid in the breakdown of organic compounds, potentially leading to increased growth (Narisawa & others, 2017). The commonly found genus *Trichoderma* is known to contain many plant rhizosphere symbionts with biocontrol applications (Harman, 2006). The frequent occurrence of individual beneficial symbionts and saprobes combined with the negative results of differential taxon abundance analysis among nurseries and cultivar suggest that there might be a subset of the fungal community that is typical of *Rhododendron*.

Plant pathogens were much less common than saprobes and mutualists. The most common pathogen found was *Pestalotiopsis unicolor* (99.6% identity), which appeared in 23% of samples. The genus *Pestalotiopsis* contains a diverse group of anamorphic plant pathogens that causes a variety of diseases such as leaf spots, blights,

and cankers in many agricultural hosts, including blueberry (Maharachchikumbura, Hyde, Groenewald, Xu, & Crous, 2014). An OTU similar to *Pestalotiopsis rhododendri* (99.1% identity) was also observed in 9% of samples. *Pestalotiopsis rhododendri* was originally described on *Rhododendron sinogrande* in 2013 in China (Zhang, Maharachchikumbura, Tian, Hyde, & others, 2013). The only oomycete pathogen that was commonly predicted and had a high sequence similarity to its reference was an OTU matching *Pythium irregulare*, which occurred in 11% of samples. The next most common *Pythium* observed was *Pythium sylvaticum*, occurring in 3% of samples. Oddly, an OTU matching *Phytophthora infestans* was also found in 3% of samples. However, it should be noted that we only sequenced ITS1 and even with the entire ITS sequence, it is not always possible to differentiate oomycetes species, so even a 100% sequence match to a reference sequence does not necessarily imply that an OTU is from that species. In particular, *Pythium irregulare* has the same sequence as *Pythium cryptoirregulare* and *Phytophthora infestans* has the same sequence as *Phytophthora andina*, *Phytophthora mirabilis*, and *Phytophthora ipomoeae* in this region of the ITS (Redekar et al., 2019). The OTU matching *Phytophthora infestans* could also be from an undescribed species, considering that the known species sharing a sequence with *Phytophthora infestans* are not known to occur on *Rhododendron*. The rare OTU classified as *Phytophthora citricola* could also be *Phytophthora plurivora*, which was found to be common in diseased rhododendrons in Oregon (Carleson, Fieland, Scagel, Weiland, & Grünwald, 2018). Notably, we did not detect some pathogens commonly found when sampling symptomatic tissue, such as *Phytophthora syringae* and *Phytophthora citrophthora* (Parke et al., 2014). We found a rare OTU in two samples

that most closely matched *Phytophthora cinnamomi* with 98.9% sequence identity, but for the previously mentioned reasons, this level of similarity does not indicate that *Phytophthora cinnamomi* was actually present. We also found fewer *Phytophthora* species compared with some other oomycete metabarcoding studies, which find around 20 species total, albeit in diseased plants for a greater diversity of sites (Prigigallo et al., 2016; Riddell et al., 2019). The relative infrequency of pathogens and the different pathogens found compared with some previous research is probably due to sampling apparently healthy plants, whereas many studies target symptomatic plants (Weiland et al., 2018).

This study used high-throughput amplicon sequencing to characterize how the oomycete and fungal rhizobiome of rhododendrons in Oregon nurseries is structured by plant cultivar and environmental factors. The overall diversity of samples varied little between production system, nursery, and cultivar, although plants in nurseries that had both container and in-field production systems were significantly more diverse than those in nurseries that only grew plants in containers. The differences between the compositions of microbiome communities were correlated with production system and nursery, but not cultivar. Communities were dominated by saprobes and symbionts. This study provides novel insights into potential factors influencing the rhizosphere microbiome of rhododendrons in nurseries, and more generally, how plant genotype and environment impact the makeup of the rhizobiome of woody plants.

Chapter 5: Rps10: a new barcode for high throughput amplicon sequencing of oomycete communities

5.1 Abstract

Oomycetes are a group of eukaryotes related to brown algae and diatoms, many of which cause disease in plants and animals. Improved methods are needed for rapid and accurate characterization of oomycete communities. With the unique order of tRNA coding regions flanking the mitochondrial *rps10* gene it was possible to design oomycete-specific primers that may be useful for oomycete metabarcoding. We evaluated the utility of this locus with a mock community and environmental samples using MiSeq Illumina sequencing. The amplification primers described herein are predicted to amplify all oomycetes tested, but analysis of sequence data from a mock community revealed that some biases are present. Simulated PCR and sequencing of environmental samples indicates the proposed *rps10*-based technique results in less amplification of non-target organisms than the ITS1-based method. We also provide a new website with a *rps10* reference database and all protocols needed for oomycete metabarcoding. Our results indicate that the *rps10* locus has greater taxonomic resolution for the oomycetes tested than the ITS1 locus and the primers proposed result in less non-target amplification.

5.2 Introduction

Oomycetes are microscopic eukaryotes related to brown algae and diatoms that often cause disease in plants and animals (Baldauf, Roger, Wenk-Siefert, & Doolittle, 2000; Yoon, Hackett, Pinto, & Bhattacharya, 2002). They include highly destructive pathogens with major impacts on agriculture (Fry, 2008), aquaculture (Phillips et al., 2008), and natural ecosystems (Cahill, Rookes, Wilson, Gibson, & McDougall, 2008; Grünwald et al., 2019). Oomycetes are primarily known for causing agricultural diseases, such as potato late blight caused by *Phytophthora infestans*, implicated in the Irish Potato Famine (Fry, 2008). Other well-known oomycete genera include *Aphanomyces euteiches*, responsible for damping off and root rot of legumes (Gaulin et al., 2007), and *Pythium* species that cause damping off and root rot on a large variety of agricultural and horticultural crops (Martin & Loper, 1999). In addition to agriculture, invasive oomycete pathogens have detrimental effects on forests, managed landscapes, and aquatic ecosystems. Forests in North America and Europe have suffered significant tree mortality due to sudden oak and larch death respectively, caused by *Phytophthora ramorum* (Brasier & Webber, 2010; Gruenwald et al., 2008). Eucalyptus forests in Australia are experiencing massive dieback caused by the invasive *Phytophthora cinnamomi* (Burgess et al., 2017) and seedlings in natural ecosystems regularly suffer damping off, which is frequently associated with *Pythium* species (Augspurger & Wilkinson, 2007). Some oomycetes are also major pathogens of fish, such as *Saprolegnia parasitica*, and are of great concern in aquaculture (Van West, 2006).

Improving methods to quickly, accurately, and economically characterize oomycete communities would be useful for controlling and understanding this important, but relatively understudied, group of organisms. While many specific oomycete pathogens are well known for the extensive damage they cause to natural ecosystems (Wills, 1993) and agriculture, oomycete diversity as a whole is much less well characterized than other microbes, such as fungi and bacteria. This is due, in part, to the challenge of collecting samples during periods when oomycetes are active, isolating them from diverse host and substrate materials (e.g., water, soil, plant and animal tissues), culturing on an assortment of media, and identifying species using morphological characteristics. Methods relying on DNA sequencing are increasingly being used to complement or replace these traditional techniques. Standardized regions of DNA known as "barcodes" are sequenced and compared to ever-growing collections of reference sequences from known isolates (Choi et al., 2015). These regions should ideally vary enough between closely related taxa to provide species-level identifications. When used on individual isolates, this process is known as "barcoding" and when used on collections of unknown organisms in an environmental sample, such as soil or plant tissue, it is known as "metabarcoding" (Taberlet et al., 2012).

Culture-independent DNA-based methods like metabarcoding have the potential to overcome many of challenges associated with characterizing oomycete communities (Tedersoo et al., 2019) as has been demonstrated with fungi (Nilsson et al., 2018; Schoch et al., 2012) and bacteria (Bukin et al., 2019; Tringe & Hugenholtz, 2008). Reliable metabarcoding methods for fungi, bacteria, and archaea have enabled the discovery of major undescribed groups of microorganisms (Fuhrman, McCallum, &

Davis, 1992; Orchard et al., 2017) and revealed a previously unexpected diversity of microbes associated with almost every habitat and multicellular organism on earth. In order for metabarcoding of oomycetes to be effective, a region of DNA not only must identify species-level differences, but also needs to be small enough for high throughput sequencing and be flanked by regions conserved in oomycetes but diverged in other organisms, so primers can be designed to amplify all oomycetes, and, ideally, nothing else (Cristescu, 2014). In addition, a curated database of high quality reference sequences of the barcode region must be publicly available so that environmental sequences produced by metabarcoding can be assigned a taxonomic classification.

There have been several attempts to create DNA barcodes for the identification of oomycetes (Choi et al., 2015; Robideau et al., 2011; Yuan, Feng, Zhang, & Zhang, 2017), but the most popular, the first internal transcribed spacer of the ribosomal DNA (ITS1), has insufficient taxonomic resolution to identify many oomycetes to the species level (Redekar et al., 2019) and currently available primers amplify other organisms (Coince et al., 2013) as well or only amplify some oomycetes (Legeay et al., 2019). Among the first primers used for barcoding individual oomycete isolates were ITS6 and ITS4 (Cooke, Drenth, Duncan, Wagels, & Brasier, 2000). These primers were designed to amplify the entire ITS region from pure cultures and were effective for phylogenetic research. Since ITS6 and ITS4 produces an amplicon too long for the most commonly used high throughput sequencing methods, they were used in a semi nested PCR with ITS6 and ITS7 to amplify only the ITS1 locus in early metabarcoding attempts. However, this technique resulted in as little as 5.3% of the OTUs (57% of the reads) recovered being assigned to oomycetes (Coince et al.,

2013). Sapkota & Nicolaisen (2015) proposed increasing the annealing temperature to increase specificity to oomycetes and reported 60% of OTUs (95% of the reads) being assigned to oomycetes using the higher annealing temperature. However, there is concern that these more restrictive PCR conditions could exclude some taxa and the requirement that labs fine-tune their PCR conditions is not ideal (Riit et al., 2016). Also, nested PCR protocols are problematic for use with metabarcoding since they increase the chance for contamination, increase the chance of sequence errors and chimera formation due to additional PCR cycles, and add to the cost of lab reagents. Riit et al. (2016) developed primers targeting the ITS1 and ITS2 regions without the need for a semi-nested approach and could assign 22% and 29% of OTUs (25% and 30% of reads) to oomycetes respectively. Additional mismatches to plants and fungi likely make these primers more specific to oomycetes with less strict PCR conditions, but there is still substantial room for improvement.

The *rps10* gene exhibits interspecific variability in the genus *Phytophthora* and has been useful for delineating species and estimating phylogenetic relationships in the genus (Martin, Blair, & Coffey, 2014). More recent analysis with the gene and flanking sequences extracted from assembled oomycete mitochondrial genomes revealed a similar level of interspecific sequence diversity at the taxonomic level of class, making it useful for understanding the evolutionary relationships among oomycetes at large (F. Martin, unpublished). In the process of working with the assembled mitochondrial genomes, a conserved order of tRNAs flanking the *rps10* gene that was unique to oomycetes was observed (tRNA-Phe, *rps10*, tRNA-Arg, tRNA-Gln, tRNA-Ile). Designing amplification primers from conserved regions of tRNA-Arg and tRNA-Ile am-

plified sequencing template from a range of oomycetes evaluated. While this locus was useful for phylogenetic analysis and as a barcode for species identification, at approximately 600bp (varies depending on the taxon) it was not suitable as a metabarcode locus for Illumina sequencers. Recently, Yuan et al. (2017) also noted the sequence divergence of the mitochondrial *rps10* locus (among others) for 14 oomycete taxa and suggested the locus as a candidate barcode for oomycete barcoding.

In this study, we propose primers for *rps10* suitable for oomycete metabarcoding on the Illumina MiSeq and compare their effectiveness to a semi-nested method to amplify ITS1 similar to the one proposed by Sapkota & Nicolaisen (2015). By compiling a curated reference database of *rps10* sequences, simulating PCR, and metabarcoding environmental samples and a mock community, we compared the specificity and taxonomic resolution of each method. We also developed a companion website to host the *rps10* reference database and all protocols needed for researchers to immediately apply this method to oomycete metabarcoding.

5.3 Materials and methods

5.3.1 Primer design

Sequence alignments of amplicons generated using *rps10* amplification primers annealing in flanking tRNA-Arg and tRNA-Ile (Martin et al., 2014) or extracted from assembled mitochondrial genomes (F. Martin, unpublished) representing 16 oomycete genera and 92 taxa were searched for highly conserved regions flanking the *rps10* gene

that would generate an amplification template suitable for MiSeq analysis. In particular, we searched for pairs of potential primer binding sites that would (1) result in an amplicon length of less than 500bp and would therefore be appropriate for short-read sequencing using platforms like the Illumina MiSeq, (2) be conserved in all known oomycete sequences, and (3) not match the sequences of other organisms, particularly fungi and plants (Cristescu, 2014). Sequences were aligned with Clustal Omega (Sievers et al., 2011) and inspected in Geneious v8.1.9 (Biomatters, Auckland, New Zealand). Forward and reverse primers were designed for the highly conserved flanking regions in the *trnF* and *trnR* genes (Figure 5.1). Potential primers were evaluated with OligoAnalyzer (Owczarzy et al., 2008) to assess their melting temperatures, CG content, and problematic secondary structures, such as dimers and hairpins.

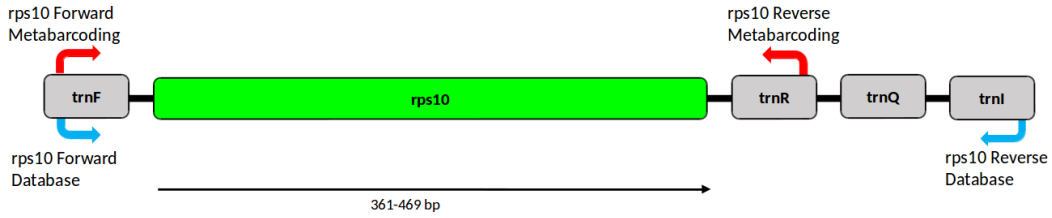

Figure 5.1: Details on the location of the 40S ribosomal protein S10 (*rps10*) locus in the circular, mitochondrial genome of oomycetes. The *rps10* locus, using the example of *Phytophthora sojae*, is flanked by several tRNAs. Two sets of primers are shown in blue for the primers for amplification of the sequence found in the reference database, and red for the primers for the *rps10* metabarcode locus suitable for high throughput sequencing of the 400-550bp amplicon. More details on the primers are provided in Table 5.1.

5.3.2 Simulated PCR

To test the newly developed primers for taxonomic specificity, coverage, and species resolution, we used mitochondrial genomes of 121 oomycete species, 38 non-oomycete stramenopiles, 19 fungi from different families, and four *Rickettsia* species. *Rickettsia* spp. were included because segments of their genomes resemble mitochondrial genomes (Andersson et al., 1998). We conducted simulated PCR using all oomycete and non-oomycete sequences to predict the sensitivity and specificity of the primers using Geneious v8.1.9 (Biomatters, Auckland, New Zealand) with the following parameters: no mismatches allowed, SantaLucia (1998) formula and salt correction, 50nM concentration of oligos, and 0.6mM dNTPs. The results were analyzed and visualized in a taxonomic context using the R packages `taxa` (Foster et al., 2018) and `metacoder` (Foster et al., 2017).

5.3.3 Isolate and environmental DNA collection

In order to test for amplification and species resolution, we collected 12 oomycete DNA samples from different laboratories in the USA (see acknowledgments). In addition, we cultured 12 oomycete isolates on media amended with PARP or appropriate host tissue (Jeffers, Martin, & others, 1986) and extracted DNA using various protocols depending on isolate origin. We then created a synthetic community by pooling DNA from each of the 24 oomycetes for a per species final concentration of 2ng/µL, with the exception of *Saprolegnia diclina* and *Phytophthora pluvialis*. The concentration of DNA measured represented all the DNA in the sample, so DNA extracts from

infected plant tissue (for obligate pathogens) also include an unknown proportion of plant DNA.

To test for non-target amplification, we obtained DNA extracted from Panama soils (Schappe et al., 2017) in addition to DNA extracted from soils, canopy drip water, and needles collected from the Wind River research forest plot in southern Washington, USA. DNA from soil was extracted following a previously reported protocol (Schappe et al., 2017). DNA from needles and canopy water, collected from under the canopy drip line of old growth *Pseudotsuga menziesii* trees, was extracted using the DNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA), following the manufacturer's instructions. Water samples were collected by filtering up to 1L of water dripping through tree canopies on 0.45µm filters and extracting DNA from the filters.

5.3.4 DNA amplification and high-throughput sequencing

DNA from the synthetic community and from environmental samples was amplified with both the newly developed *rps10* method and the ITS1 method to generate amplicons for high-throughput sequencing. Negative PCR controls were included in both assays. The *rps10* assay is a multiplex PCR reaction comprising two forward *rps10* primers (rps10_F_Conserved and rps10_F2_Conserved) and seven reverse *rps10* primers (rps10_R1 through rps10_R7). All amplifications of the *rps10* locus were carried out using the QIAGEN Type-it Mutation Detect PCR Kit (QIAGEN, 206343, Valencia, CA). Multiplex PCR reactions were performed in 35.0µL with 14ng template DNA and 1X final buffer concentration. Final primer concentrations were

0.2µM except for rps10_F_Conserved which has a single degenerate base (Y = C, T) representing two primers and was therefore added at a final concentration of 0.4µM. Amplifications were run in a Veriti thermal cycler (Life Technologies, Grand Island, NY) with an initial denaturation at 95°C for 5 minutes, followed by 35 cycles of 95°C for 30 seconds, 58°C for 3 minutes, and 72°C for 30 seconds, and a final extension at 60°C for 30 minutes.

ITS1 was amplified following a semi-nested protocol similar to a previously published one using the ITS6/ITS4 primer set and the ITS6/ITS7 primer set (Sapkota & Nicolaisen, 2015). The initial PCR reaction was performed in 1X PCR buffer (with 1.5mM MgCl2), 1.5mM MgCl2, 0.2mM dNTP mix, 0.2µM of each primer (ITS6 and ITS4), and 2 U/rxn of Platinum Taq (#10966018, ThermoFisher Scientific). The non-proofreading Platinum Taq was used since previous efforts to use various proofreading Taqs resulted in unacceptably-strong amplification of plant DNA using these primers (unpublished data), possibly due to the 3' to 5' exonuclease activity of proofreading taqs removing bases at the 3' end of the ITS7 primer that are divergent in plants (Sapkota & Nicolaisen, 2015). DNA (2.6ng/µl, except for *P. pluvialis* at 1.27ng/µl and S. diclina at 0.7ng/µl) was added to each reaction for a total volume of 15µL. The second PCR reaction was identical to the first with the following exceptions: Template was 1.0µL of the initial PCR reaction, primers were ITS6 and ITS7, and the total reaction volume was 25µL. Both ITS1 PCR amplifications were conducted in a Bio-Rad T100 thermocycler (Bio-Rad, Hercules, CA, USA) under the following thermal cycling conditions: 2 minutes at 94°C, 25 cycles of 30 seconds at 94°C, 30 seconds at 60°C, 1 minutes at 72°C, and a final extension of 2 minutes at 72°C. The

protocol differs from that of Sapkota & Nicolaisen (2015) in that the first PCR had 25 cycles instead of 15 and that the annealing temperatures of the PCR were raised from 55°C and 59°C to 60°C for both. These changes were based on an optimization of the PCR for a previous experiment (Foster et al., 2020) that was necessary to avoid non-target amplification.

Rps10 amplicons as well as amplicons from the second ITS1 PCR reaction were cleaned, ligated to Illumina Nextera XT indices and adapters, and then purified and pooled following Illumina's protocols (Illumina, 2013). Sequencing was done on the Illumina MiSeq platform with 300bp paired-end reads at the Center of Genome Research and Biocomputing at Oregon State University.

5.3.5 The *rps10* database and associated website

A curated reference database was developed for assigning taxonomic classifications to sequences generated from the metabarcoding primers. The database is composed of manually curated sequences from online databases, sequences contributed from other labs, sequences extracted from whole mitochondrial genomes, and sequences produced from known isolates for this study. Species classification was confirmed by *cox1* or ITS sequence analysis and compared to vouchered specimens in GenBank. To generate sequences for the reference database, we modified previously reported amplification primers (Martin et al., 2014) by the addition of degenerate bases. This protocol results in an amplicon that includes one of the metabarcoding primer binding sites, so that the effectiveness of that primer can continue to be evaluated as more

sequences are added. We could not find a suitable site that would allow the other primer to be included. Amplification of the *rps10* database amplicon was conducted in 25.0µL with 0.025 U/µL, GenScript Taq (GenScript, Cat. No. E00007), 1X Taq Buffer, 0.2µM dNTPs, 1.5mM MgCl2, 2.0ng DNA, 0.5µM of each primer (Prv9r-M and Prv9f-M). Thermal cycling was performed using a Veriti thermal cycler (Life Technologies, Grand Island, NY) with an initial denaturation at 94°C for 3 minutes, followed by 35 cycles of 94°C for 30 seconds, 55°C for 45 seconds, and 72°C for 45 seconds, and a final extension at 72°C for 7 minutes. This is the suggested protocol for future researchers to use to contribute sequences to the reference database.

To host the *rps10* database and laboratory protocols, the website www.oomycetedb.org was created. The website is a combination of static HTML produced from Rmarkdown documents (Xie, Allaire, & Grolemund, 2018) and interactive R Shiny applications (Chang, Cheng, Allaire, Xie, & Mcpherson, n.d.). Lab protocols can be viewed on the website or downloaded as printer-friendly PDFs. Users can download all or a specific subset of the database based on a search term. Users can also BLAST their own sequences against the database, view the results online, and download the results in any format BLAST can output. Updates to the database are released on this website with a unique version number and old versions will continue to be available. All tools on the website can be used with any version of the database so that researchers can reproduce analyses. All source code and documents for the website are available on Github at https://github.com/grunwaldlab/OomyceteDB.

5.3.6 Abundance matrix preparation

An ASV abundance matrix with associated taxonomic annotations was created from MiSeq reads using `cutadapt` (Martin, 2011) and the R package `dada2` (Callahan et al., 2016). Primer sequences were trimmed from reads using `cutadapt`. Reads were then filtered out using the `filterAndTrim` command of `dada2` if they were expected to contain 2 or more errors, based on their quality scores. Reads were also truncated at the first instance of a quality score less than 4 and any reads that were then shorter than 50bp were removed. Error rates were estimated and used to infer ASVs using the `learnErrors` and `dada` commands. This is intended to distinguish sequencing and PCR errors from true biological sequences, but does not cluster sequences into species-level differences like OTU-based approaches do (Callahan et al., 2016). ASV read pairs were merged using the `mergePairs` function and predicted chimeras were removed using `removeBimeraDenovo`. Any sequences less than 50bp long were also removed. A taxonomic classification was assigned to each ASV using the RDP Naive Bayesian Classifier algorithm implemented in the `assignTaxonomy` command (Wang et al., 2007). The algorithm assigns a bootstrap value to each taxonomic rank for each classification, providing a kind of confidence measure for which taxon in the reference database is most similar. Each ASV was also optimally aligned to the best-matching reference sequence to calculate a percent identity using the `pairwiseAlignment` function from the `biostrings` R package (Pages, Aboyoun, Gentleman, & DebRoy, 2009).

ASVs were clustered into OTUs using VSEARCH (Rognes et al., 2016) to create

an OTU abundance matrix, in addition to the ASV abundance matrix. To inform the choice of clustering thresholds for each locus, ASVs present in the mock community were clustered at thresholds ranging from 90% to 100% in 0.1% increments and the number of resulting OTUs recorded. Thresholds were chosen that best reproduced the number of species used in the mock community and conformed with previous experience using ITS1 as an oomycete barcode. *Rps10* samples were clustered at the 97% threshold and ITS1 samples were clustered at the 99% clustering thresholds. Taxonomy was assigned to OTUs using the same methods as taxonomy was assigned to ASVs.

5.3.7 Mock community

The inferred composition of the mock community based on sequencing results was compared with the composition of species put into the mock community to evaluate the performance of the *rsp10* and ITS1-based methods. For this analysis, only ASVs/OTUs representing at least 30 reads were used. ASVs/OTUs that were found in the mock community samples were classified as "expected", "near expected", or "non-target". ASVs/OTUs were considered "expected" if their taxonomic assignment matches a member of the mock community. ASVs/OTUs assigned to other taxa but had at least 99% sequence identity to a mock community sequence were classified as "near expected". Otherwise they were classified as "non-target". The numbers of reads and species found in each of these categories were then calculated for each locus. A single representative ASV for each unique taxonomic annotation was used

to construct bootstrapped neighbor-joining trees for each locus using the `ape` package (Paradis et al., 2004). Reference database sequences for ASVs that did not exactly match their closest reference sequence were included in the tree as well. Reference database sequences for species that were put into the mock community, but not detected were also included in the tree. This makes it apparent which taxa were not amplified and which amplified but had incorrect taxonomic classifications.

5.3.8 Non-target amplification

ASVs/OTUs associated with environmental samples were used to assess non-target amplification. Since our reference databases do not contain many non-target sequences, ASVs/OTUs in this analysis were assigned an alternative taxonomic classification using BLAST against the NCBI nucleotide database (Altschul et al., 1990). Although NCBI taxonomic annotations can be unreliable (Nilsson et al., 2006), we only considered kingdom-level portions of the taxonomy, which we considered more likely to be correct. The best BLAST hit was chosen for each ASV/OTU based on the E-value and percent identity of the matching region. BLAST hits with an E-value less than 0.001 were not considered. Using the taxonomy associated with the best BLAST hit, ASVs were grouped into “Oomycetes”, “Fungi”, and “Other” categories. They were considered “Unknown” when no acceptable BLAST hit was found. The proportions of reads, ASVs, and OTUs in each category for each locus were then compared to evaluate which locus had more non-target amplification.

5.3.9 Taxonomic resolution

The ability of each locus to distinguish different taxa was evaluated by comparing the distribution of bootstrap scores of ASV taxonomic assignments and pairwise alignments of the reference database sequences. The bootstrap values are those assigned by the RDP Naive Bayesian Classifier algorithm implemented by the `assignTaxonomy` command of `dada2` and measure how consistent the taxonomic assignment is for a given reference database when parts of the sequences are subsampled. The name of the specific reference database sequence was included as part of the taxonomy of each reference sequence, as a "rank" below species, so bootstrap values were also generated for which reference sequence matched, which is useful for when there are multiple reference sequences for the same species. The distribution of bootstrap values for mock community samples for the genus, species, and reference database sequence ranks were compared for ITS1 and *rps10*. The portion of the reference database sequences that was predicted to be amplified by each primer was aligned using MAFFT (Katoh, Kuma, Toh, & Miyata, 2005) and all pairwise differences in sequence identity were calculated using the `ape` package. Reference sequences with the complete amplicon were used, as determined by the presence of primer binding sites, using a modified version of `matchProbePair` function from the `Biostrings` package that allows for ambiguity codes. Reference sequences were also included if they aligned to at least 90% of one of the predicted amplicon, in which case the portion aligned to the amplicon was used. For each locus, the distributions of the percent identity of each species to the closest sequence from a different species was compared to assess taxonomic

resolution and lists of indistinguishable species were generated.

5.4 Results

5.4.1 Development and validation of primers for the *rps10* region

New primers were developed for specific amplification of oomycetes using the *rps10* region (Table 5.1; Figure 5.1). Simulated PCR with the oomycete-specific forward "rps10-F" and reverse "rps10-R" primer sets amplified the *rps10* mitochondrial gene with no mismatches on any of 121 oomycetes sequences analyzed (Figure 5.2). The forward primer rps10-F has 40% GC content, a mean melting temperature of 59.8°C, and a predicted maximum homodimer delta G of -63.9 kcal/mole, while the reverse primer rps10-R has 31% CG content, a mean melting temperature of 58.2°C, and a predicted maximum homodimer delta G of -9.83 kcal/mole. The difference between the primer melting temperature is predicted to be 1.6°C. The predicted maximum heterodimer delta G between rps10-F and rps10-R is -5.41 kcal/mole. The *rps10* locus-specific primers were predicted to amplify a sequence of median length of 481bp (including primers), with *Albugo laibachii* producing the shortest with 448bp, and *Peronospora tabacina* producing the longest with 513bp. Compared with predicted ITS1 amplicons, *rps10* amplicons have less variation in length. For other Stramenopiles, fungi, and *Rickettsia*, the *rps10* primers were predicted not to bind anywhere in the whole mitochondrial genome under the amplification conditions selected (Figure 5.2).

Table 5.1: Primer sequences used in this study.

Primer	Sequence (5'-3')
rps10__F__Conserved	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GTTGGTTAGAGYAAAAGACT
rps10__F2__Conserved	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GTTGGTTAGAGTAGAAGACT
rps10__R1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG ATGCTTAGAAAGATTTGAACT
rps10__R2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG ATACTTAGAAAGATTTGAACT
rps10__R3	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG ATGCTTAGAAAGACTTGAACT
rps10__R4	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG ATGCTTAGAAAGACTCGAACT
rps10__R5	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG ATGCCTAGAAAGACTCGAACT
rps10__R6	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG ATGTTTAGAAAGATTCGAACT
rps10__R7	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG ATGCTTAGAAAGATTCGAACT
rps10__DB__F	GTTGGTTAGAGYARAAGACT
rps10__DB__R	RTAYACTCTAACCAACTGAGT
ITS6	GAAGGTGAAGTCGTAACAAGG
ITS4	TCCTCCGCTTATTGATATGC
ITS7	AGCGTTCTTCATCGATGTGC

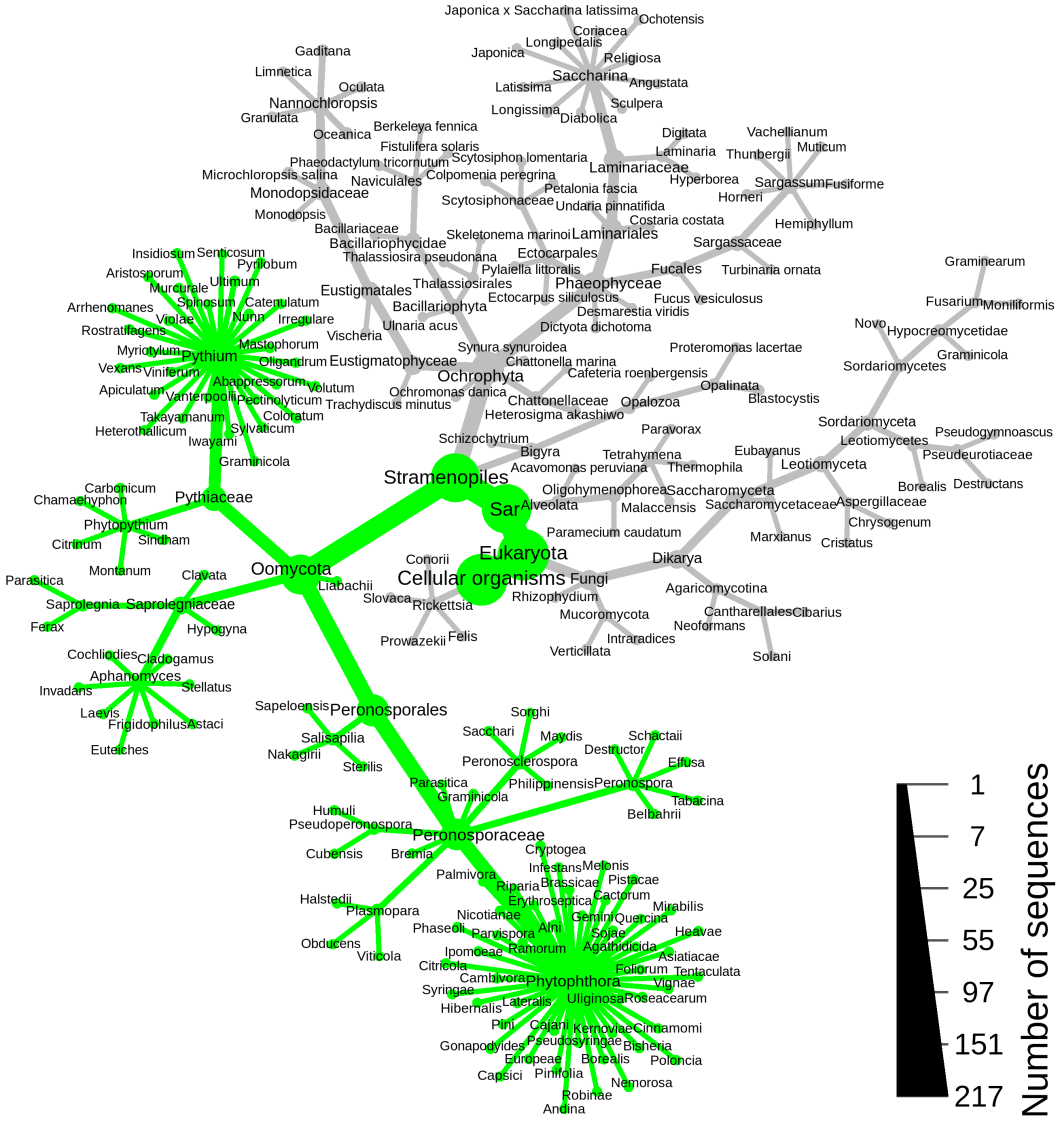

Figure 5.2: Heat tree showing specific amplification of oomycetes for the *rps10* barcode using primers rps10-F and rps10-R predicted by simulated PCR. In green are all the taxa predicted to be successfully amplified with the new *rps10* locus-specific primers. The analysis includes unrelated organisms to demonstrate specificity.

5.4.2 Metabarcoding of the mock community

Sequencing of a mock community composed of 24 oomycete species suggests that the *rps10* provides a more accurate reconstruction of community composition (Figure 5.3). Both methods resulted in more ASVs than were assembled in the mock community, but the *rps10* results were closer to the true count with no non-target ASVs (Figure 5.3A). The ASVs detected by *rps10* were all assigned to species present in the mock community or to closely related taxa with nearly identical sequences, whereas many ASVs generated from ITS1 were assigned to taxa not present in the mock community. Many of these unexpected ASVs generated by ITS1 were within 99% sequence identity to a species in the mock community (i.e., "near expected"), but some were more different (i.e., "non-target"). The "near expected" and "non-target" ASVs accounted for a large proportion of the reads generated by the ITS1-based method (Figure 5.3B). Although more ASVs were found than there were species in the mock community, the taxa these ASVs were assigned to did not account for all the species included in the mock community, since multiple ASVs were assigned to some taxa and some taxa were undetected (Figure 5.3C). Although both methods failed to detect some taxa, *rps10* detected more of the expected species than ITS1.

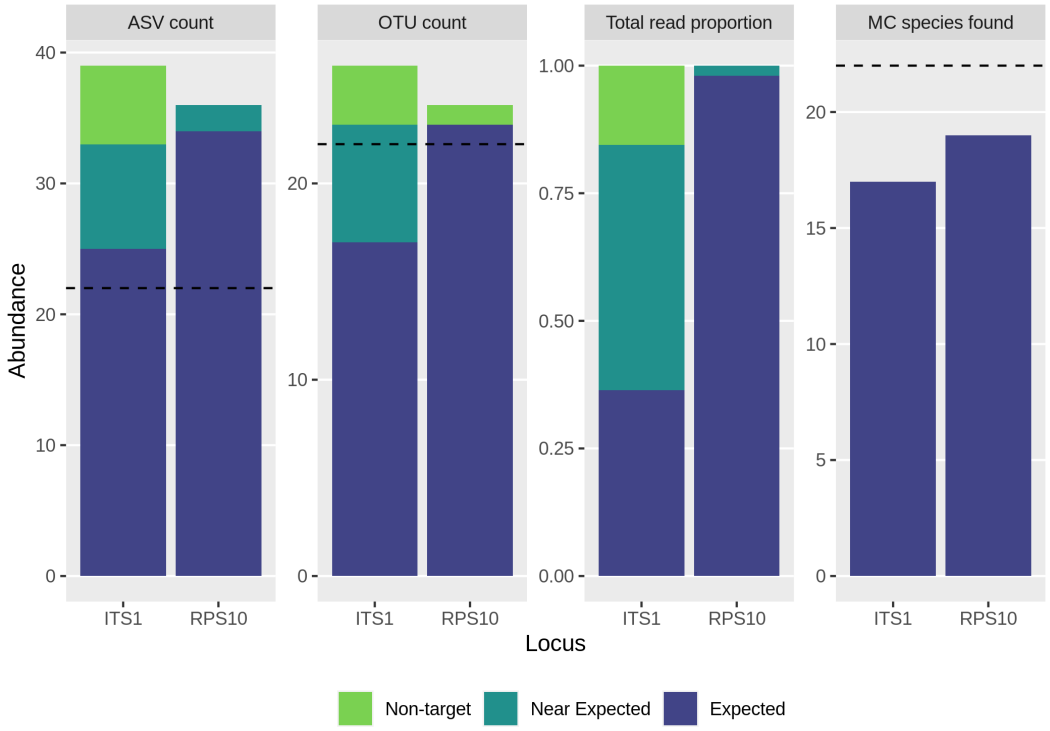

Figure 5.3: The abundance of ASVs, reads, and successfully detected species in the sequenced mock community, using the ITS1 and *rps10* loci. Sequences were considered "expected" when they were assigned to a taxon included in the mock community, "near-expected" when they were assigned to another taxon but had at least a 99% identity match to a species in the mock community, and "non-target" otherwise. The dotted line indicates the number of mock community species expected.

A neighbor-joining tree of ASVs in the mock community representing each unique taxonomic assignment, along with selected reference sequences, suggested that *rps10* is somewhat better at reconstructing the mock community than ITS1, and that both methods might miss or misclassify some taxa (Figure 5.4). For example, with the ITS1 analysis there were three taxa misidentified as a closely related species and four

species listed that were not present. This compares with only two species misidentified as a closely related taxa in the *rps10* analysis. After taking into account obvious taxonomic misclassifications, we could not detect sequences similar to *Aphanomyces euteiches* or *Plasmopara halstedii* using *rps10*. Rerunning the analysis with minimal read quality filtering (data not shown) revealed that reads for *Plasmopara halstedii* were present but were filtered out by our rather stringent quality filtering used for this analysis. Using ITS1, we could not detect sequences similar to *Pythium oligandrum*. The ITS1-based method also resulted in a few unexpected ASVs with low-confidence taxonomic classifications, whereas the *rps10*-based method did not.

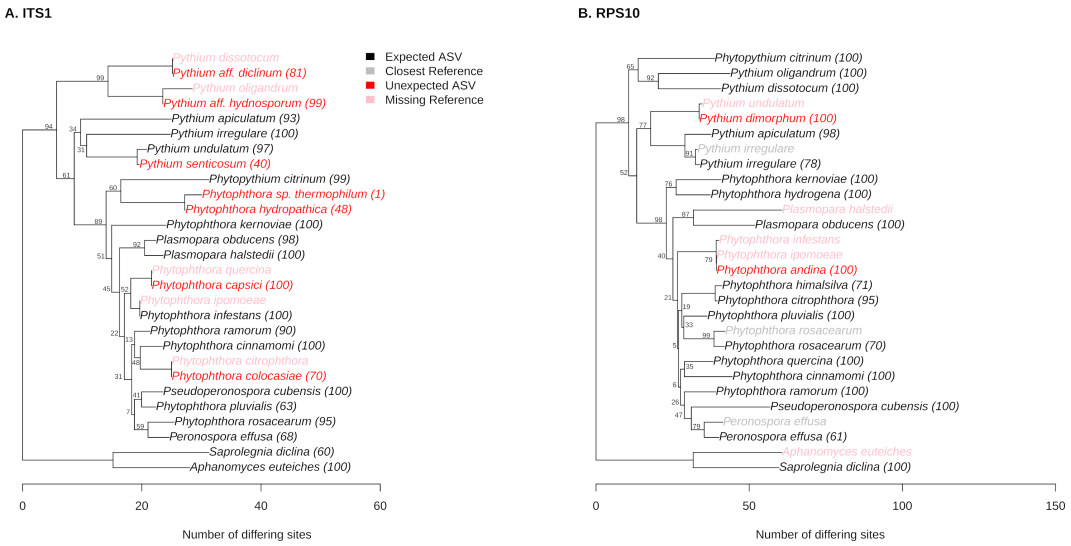

Figure 5.4: Bootstrapped neighbor-joining tree of the sequenced mock community and some selected reference sequences for the ITS1 (A) and *rps10* (B) loci. Only one ASV per unique species is shown. Species names in black and red are ASV sequences and those in grey and pink are selected reference database sequences. Grey species are the closest reference sequences for each ASV that was assigned to a member of the mock community but did not have an exact sequence match. Pink species are reference sequences for members of the mock community not found. Red species are ASVs assigned to species not included in the mock community. Numbers in parentheses at the end of the species names are the bootstrap values for the taxonomic assignment.

5.4.3 Non-target amplification of environmental samples

Sequencing of environmental samples from water, soil, and plant material suggested that the *rps10* barcode resulted in less non-target amplification, in terms ASV, OTU, and read counts (Figure 5.5). Most of the ASVs generated from *rps10* were assigned to oomycetes and these ASVs accounted for nearly all of the reads (Figure 5.5). Many

of the *rps10* ASVs could not be assigned to a taxon based on BLAST searches against the NCBI nucleotide database, particularly for the plant-derived samples, although they accounted for a low proportion of the total reads. Since the *rps10* locus is not fully represented in taxonomic databases, these could represent unknown oomycetes. When these unknown ASVs were clustered into OTUs, they accounted for a larger fraction of the total OTUs, indicating that many of these unknown ASVs are at least somewhat diverse. In contrast, relatively few of the ITS1 ASVs were assigned to oomycetes and these accounted for a little over half of the reads and a small fraction of the OTUs and ASVs. Most of the non-target ITS1 ASVs, OTUs, and reads were assigned to fungi.

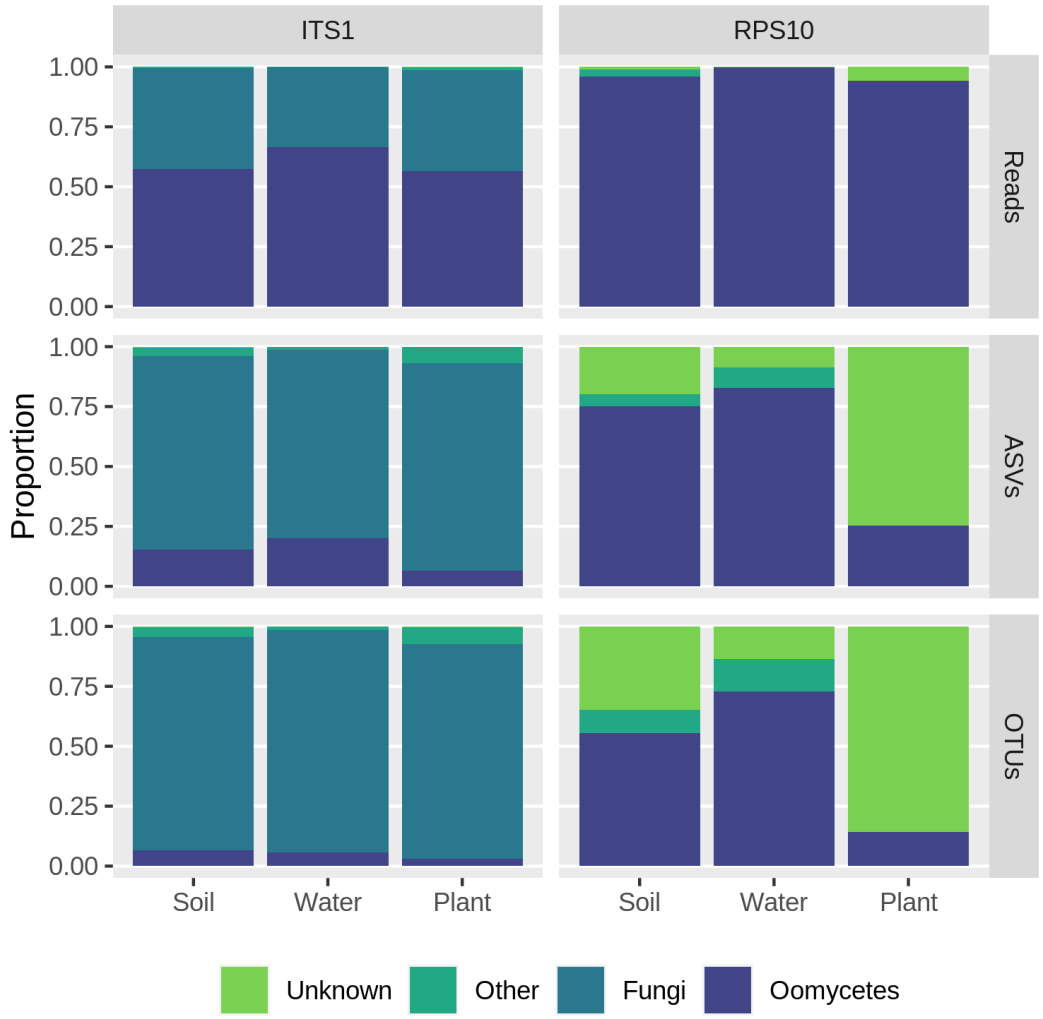

Figure 5.5: Target vs non-target amplification using oomycete-specific primers for the ITS1 and *rps10* loci displayed as counts of ASVs, OTUs, and reads from a variety of environmental samples, grouped into soil, water, and plant tissue samples. ASV sequences were given a coarse taxonomic assignment based on BLAST searches against the NCBI nucleotide sequence database. Those assigned "Unknown" did not have a match with an E-value of at least 0.001. Sequences in "Other" include plant, animal, bacterial, and protist sequences.

5.4.4 Taxonomic resolution

Bootstrap scores for taxonomic classifications and neighbor-joining trees of ASVs in mock community samples were higher for *rps10* amplicons than ITS1 amplicons (Figure 5.6). The taxonomic classification bootstrap scores from the RDP Naive Bayesian Classifier implemented in `dada2` at the reference sequence and species level were higher for *rps10* amplicons than ITS1 amplicons. Scores at the genus rank and the coarser taxonomic ranks were nearly always 100 (the highest score) for both methods. The bootstrap scores for the neighbor-joining trees of the mock community sequences were also higher in *rps10* than in ITS1. The branch lengths of the tree were also higher for *rps10* than ITS1 (Figure 5.4), suggesting that *rps10* sequences from different taxa are generally more diverged from each other than sequences from different taxa for ITS1.

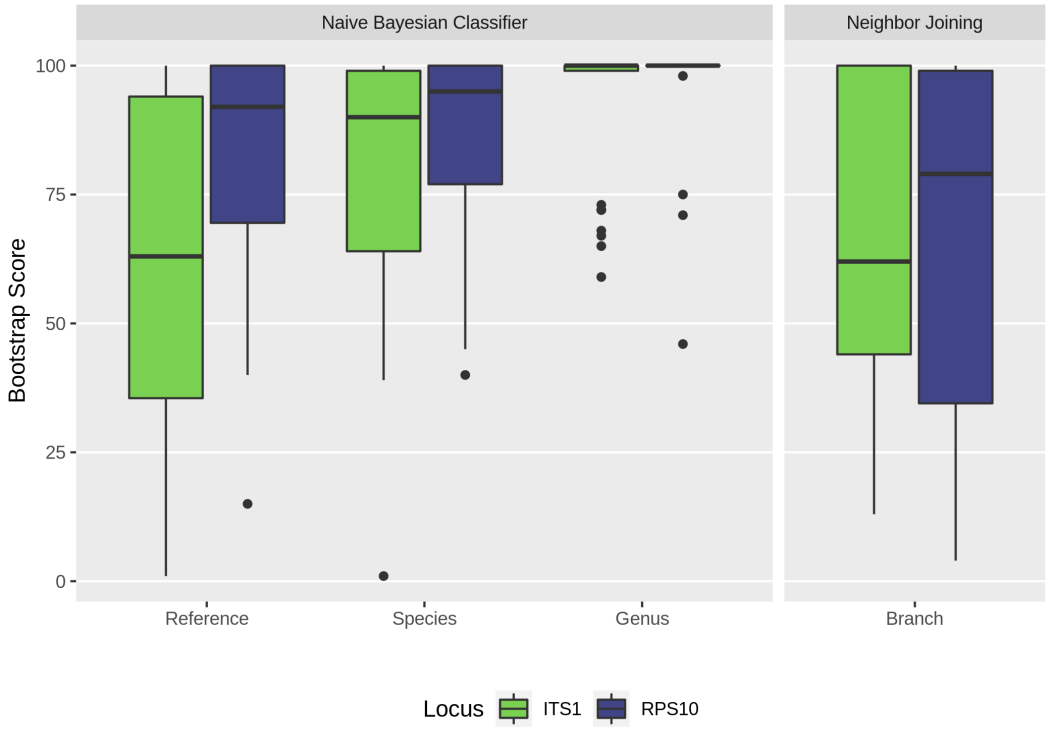

Figure 5.6: The distribution of bootstrap scores for the taxonomic assignment of ASVs in the mock community for the ITS1 and *rps10* loci. The RDP Naive Bayesian Classifier "Reference", "Species", and "Genus" scores refer to the ability to consistently assign ASVs to a particular reference sequence, species, or genus respectively when the data is resampled. The neighbor-joining tree scores quantify how consistent the branching pattern of the resulting tree is when the data is resampled.

The number of base pairs differentiating the amplified region of the most similar species in the reference databases also indicate that the *rps10* amplicon differentiates closely related taxa more effectively than ITS1. Over 50% of the predicted amplicons derived from unique species in the ITS1 references database shared an identical amplicon with a different species whereas only about 12% for species in the *rps10*

database did (Figure 5.7). In addition, over 50% of the species in the *rps10* database were distinguished from their most closely related species by 5 or more base pairs, whereas less than 5% of the ITS1 species were. When ASVs from the mock community were clustered into OTUs at a range of clustering thresholds, it was found a clustering threshold of 97% for *rps10* ASVs and 98.5% for ITS1 ASVs resulted in the correct number of sequences.

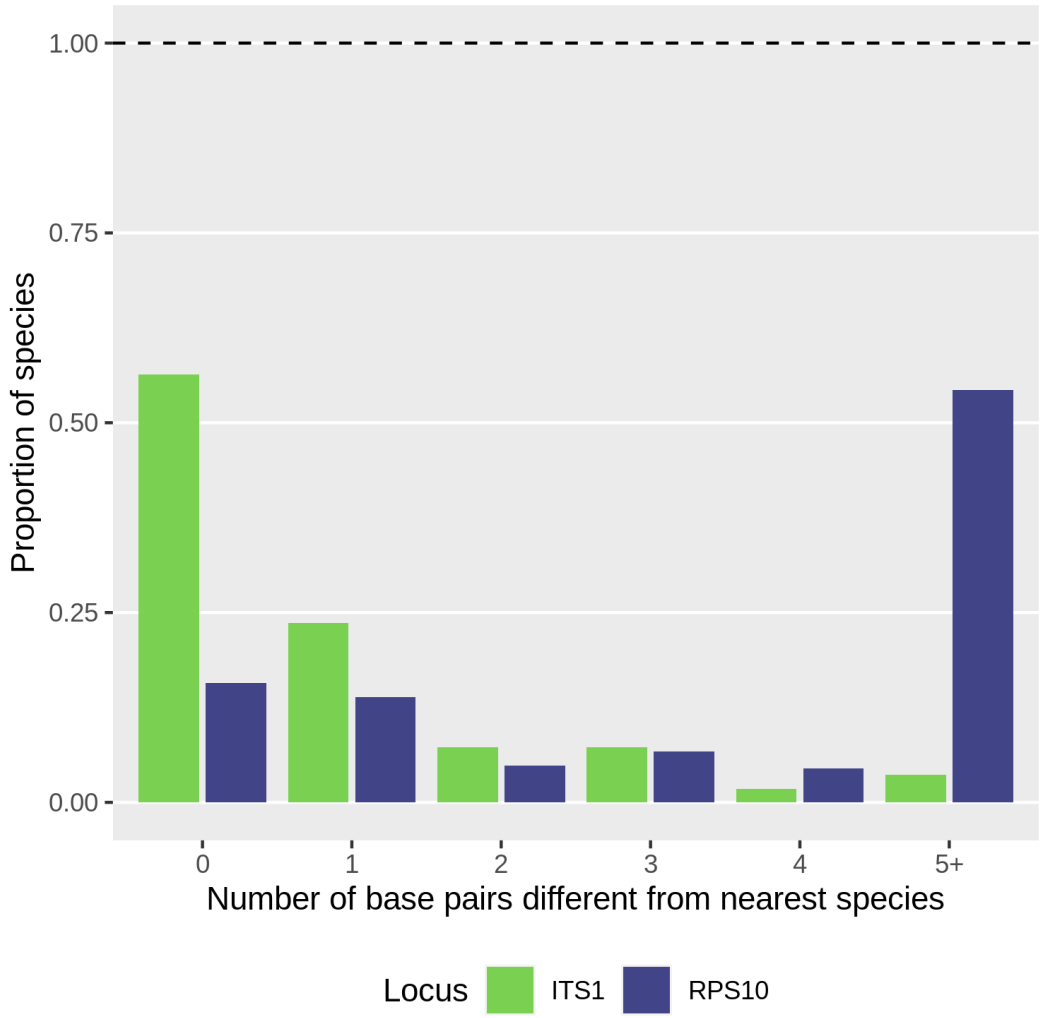

Figure 5.7: The distribution of the smallest inter-species distances for predicted amplicons for each species in the *rps10* and ITS1 databases. Only sequences with unambiguous, species-level taxonomic classifications that contain the entire amplicon are included. Zero differences for a species mean that at least one other different species is predicted to have an identical amplicon.

5.5 Discussion

Effective methods for studying microbial communities without the need for culturing and manual identification have greatly increased understanding of bacterial and fungal communities in recent decades, but methods for studying oomycetes communities are still being developed. Specific oomycete pathogens have important, and occasionally catastrophic, effects on agriculture and natural ecosystems, but relatively little is known about this group of organisms as a whole and their distribution and function in nature. Metabarcoding has the potential to accelerate our understanding of oomycete biodiversity and distribution. A few methods have been proposed for metabarcoding of oomycetes, but have not yet been widely implemented, partly due to high levels non-target amplification, limited taxonomic resolution of the loci used, and incomplete reference databases. Here we propose a new method for oomycete metabarcoding with an associated reference database and compare its effectiveness to another ITS1-based method.

Our results suggest that the *rps10* locus has much greater taxonomic resolution than ITS1, but cannot uniquely distinguish all species even so. Pairwise alignments of the predicted amplicons from reference database sequences indicate that over 50% of species tested in the ITS1 database share their exact sequence with at least one other different species (Figure 5.7). Assuming the species tested are representative of oomycete diversity, this suggests that many of the species-level taxonomic assignments of oomycete microbiome experiments using ITS1 could be wrong, unless a method is used that provides confidence measures for each rank in the classification,

like the RDP Naive Bayesian Classifier. However, such methods are usually not used in fungal and oomycete microbiome research, since they rarely yield confident results at the species level (Tedersoo et al., 2019). Since all sequencing methods have some degree of error, a single base pair difference could be considered insufficient to distinguish species. If 2bp differences are required to reliably distinguish species, the proportion of species not able to be uniquely identified by ITS1 increases to around 75%. In contrast, only 12% of *rps10* species shared their exact sequence with at least one other species and this increases to 20% of species if a 2bp difference is required. The *rps10* sequences of 70% of species are separated from the most similar sequence from a different species by 5 or more base pairs, indicating that most species tested can be confidently assigned taxonomy with *rps10* even in the presence of significant sequencing error. Comparisons of bootstrap scores of the taxonomic assignments of the mock community species using the RDP Naive Bayesian Classifier and bootstrapped neighbor-joining trees suggest that the *rps10* method typically results in more confident taxonomic assignments (Figure 5.6). The confidence of the taxonomic assignment of the RDP Naive Bayesian Classifier is influenced by the reference database and since the two loci have different reference database characteristics, these comparisons must be interpreted with caution. However, only bootstrap values for members of mock community were included in this analysis and all these species are in both databases, which should minimize this bias somewhat. These results are corroborated by the lower optimal clustering threshold of 97% for *rps10* versus 98.5% for ITS1, suggesting the average difference between sequences from different species is greater in *rps10*. Although *rps10* cannot be used to identify all species, it is much

better than the currently used ITS1.

The *rps10* method proposed here is predicted to amplify almost all oomycetes, but sequencing of a mock community revealed biases still exist. Simulated PCR of 121 *rps10* reference database sequences with primer binding sites indicated that the proposed primers should be able to amplify all species tested (Figure 5.2). However, when a mock community of 22 isolates was sequenced, one or two isolates were not detected, depending on the quality filtering settings used (Figure 5.4). We could find no reads matching *Aphanomyces euteiches* and the few reads that matched *Plasmopara halstedii* were removed during quality filtering. The sample for *Plasmopara halstedii* was a mixture of plant and pathogen DNA, with pathogen DNA most likely much less than plant, and this could be responsible for the lower read depth relative to other oomycetes included in the analysis. These two species might be biased against somewhat by having unusually long amplicons (Nichols et al., 2018) of 489bp and 491bp respectively, but *Plasmopara obducens* was detected even though it had a longer amplicon length of 492bp. When the DNA extracts of *Aphanomyces euteiches* and *Plasmopara halstedii* used in the mock community were amplified individually, they produced bright bands on a gel, suggesting that their absence in the metabarcoding results was due to biases in sequencing depth in the context of an oomycete community rather than an inability of the method to detect them individually. The ITS1 method yielded sequences at least somewhat similar to all of the species in the mock community although they were misclassified at a much higher rate and numerous unexpected sequences were also present. Therefore, when judging by the number of correct classifications alone, without manual interpretation of the results,

the *rps10* method outperformed the ITS1 method (Figure 5.3). This measure is perhaps more relevant to metabarcoding studies, since manual curation of taxonomic classifications is not practical for most studies. Further research will be needed to verify these findings and determine if the proposed *rps10* primers are effective at detecting *Aphanomyces* species when in a mixture of other organisms.

Simulated PCR and sequencing of environmental samples suggests that the proposed *rps10*-based method results in far less amplification of non-target organisms than the ITS1-based method. Simulated PCR using the *rps10* reference database indicates that all oomycete sequences tested should be amplified by the proposed primers and that no non-target sequences, even other stramenopiles, should be amplified (Figure 5.2). However, the results of this analysis only apply to the taxa tested, and as new species are sequenced and discovered, new biases will likely become apparent. We could not compare the *rps10* to the ITS1 method using simulated PCR because we could not find publicly-available ITS1 sequences with primer binding sites for many of the species tested in the *rps10* analysis. This is probably because most ITS1 reference sequences were produced with at least one of these primers or similar ones with overlapping binding sites (Bellemain et al., 2010). However, other analyses suggest that the ITS1 method produces more non-target amplification. When the same environmental samples were sequenced with both methods, 88% of ITS1 ASVs and 40% of ITS1 reads were assigned to non-target sequences, compared to 40% and 3% respectively for *rps10* (Figure 5.5). Other studies using this method have reported similar results, such as Coince et al. (2013), where only 5% of OTUs found were assigned to oomycetes. Although Sapkota & Nicolaisen (2015) describes increasing the

specificity of the primers to oomycetes by raising the annealing temperature, we still observed much non-target amplification using the ITS1 method, even though we used an annealing temperature 1°C higher than was recommended in Sapkota & Nicolaisen (2015).

The proposed method is appropriate for use with the Illumina MiSeq and should reduce biases from multiple sources when used for metabarcoding, compared to the ITS1-based method we tested. A single PCR step should reduce the chance of contamination and lower the cost of reagents compared with the nested PCR of the ITS1 method. The *rps10* amplicon also has much less variation in length, which should reduce the amount of read count bias (Nichols et al., 2018). Although the sequenced *rps10* amplicon of some oomycetes is near the upper limit for MiSeq 300bp paired-end sequencing, we were able to amplify and merge the paired-end reads of *Plasmopara obducens*, the longest amplicon in the mock community and the 5th longest in the reference database. Less amplification of non-target organisms should reduce the need to optimize PCR conditions as well as improve amplification efficiency of target organisms due to reduced competition for primers. This is an improvement compared to the ITS1 method, which is very sensitive to the chosen annealing temperature of the PCR. All resources needed to use apply this method to metabarcoding oomycete communities using the Illumina MiSeq, including lab protocols and a reference database for taxonomic classification of results, are provided at www.oomycetedb.org.

The global diversity of oomycetes is still largely unknown, with little knowledge of where invading species come from or their habitat ranges, both natural and invaded. This is underlined by results from our environmental samples, where most of the

OTUs found had less than 90% similarity to a reference sequence, meaning OTUs could only be classified at the genus or family level. This is partially because the *rps10* reference database is incomplete and because many, if not most, oomycetes that occur in natural ecosystems have not yet been described. There are thousands of oomycete specimens in herbariums around the world. A collaborative effort to sequence *rps10* barcodes from a wider range of oomycetes is in progress and will improve the accuracy and usefulness of this barcode. To be able to confirm the species classification it is important to also provide the *cox1* or ITS sequence of the isolates under study. We encourage the oomycete community to assist with this by uploading sequences of their oomycete isolates to www.oomycetedb.org (Table 5.2).

Table 5.2: Overview of the number of species currently available in the *rps10* reference database.

Genus	Number of species
Achlya	1
Albugo	1
Aphanomyces	13
Bremia	3
Halophytophthora	4
Hyaloperonospora	1
Perofascia	1
Peronosclerospora	4
Peronospora	72
Phytophthora	509
Phytopythium	11
Plasmopara	4
Pseudoperonospora	3
Pythium	103
Salisapilia	4
Saprolegnia	3
Thraustotheca	1
Total	739

To conclude, the proposed *rps10* method provides a superior alternative to the ITS1 method we tested. Oomycetes are an extremely important but relatively understudied group of organisms. Understanding their diversity and distribution will be helpful in understanding future outbreaks of destructive pathogens like *Phytophthora ramorum* and *Phytophthora infestans* as well as characterizing the role of the less destructive majority of oomycetes present in natural ecosystems. Whatever method is used, it is important that an effective method be developed to efficiently characterize oomycete communities. Currently, metabarcoding using Illumina sequencing is the most cost-effective technique (Tedersoo et al., 2019). We hope the method presented here will facilitate new insights into oomycete diversity and biology, just as robust methods for metabarcoding of fungi and bacteria have revolutionized our understanding of these organisms in recent decades.

5.6 Data availability

A curated *rps10* database is available at www.oomycetedb.org for downloading with open access protocols. We encourage submission of new species to the database to improve this resource for the community (see instructions online). The version-controlled code for the analysis presented here is available at https://github.com/grunwaldlab/rps10_barcode.

Chapter 6: Conclusion

6.1 The impact of molecular methods in community ecology

Microbial community ecology has advanced at an unprecedented rate in recent decades, largely due to culture-independent molecular methods such as metabarcoding and shotgun metagenomics. It is now clear that the vast majority of microbial diversity has been overlooked until recently, including entire phylum-level groups of organisms living in poorly-characterized habitats (Brown et al., 2015; Hug et al., 2016). Since most of this diversity does not appear to be culturable, many of these microbes might exist in complex interconnected communities that will remain difficult to study experimentally. The discovery that nearly all macroscopic animals and plants are actually "super organisms" (Berg & Raaijmakers, 2018; Boursnell, 1950; Mendes et al., 2011; Vandenkoornhuyse, Quaiser, Duhamel, Le Van, & Dufresne, 2015) with essential communities of microbes forces us to reconsider organisms that were thought to be well-characterized, including ourselves (Arnold et al., 2003; Hosokawa, Kikuchi, Nikoh, Shimada, & Fukatsu, 2006; Rodriguez et al., 2008; Rosshart et al., 2017). In order to fully understand the health and reproduction of plants and animals, we must also understand the complex community of microbes that inhabit them and have co-evolved with them from their beginning. Health for plants and animals is no longer as simple as killing all "germs". Now we must

consider which microbes are dangerous and which are beneficial and under what conditions. Nutrition is no longer as simple as standardized daily amounts of a variety of chemicals. Now microbe-mediated bioavailability and nutrient-microbe interactions must be considered. All these are complex interactions that will likely take centuries to elucidate if not longer.

Most of the recent advances in microbial community ecology have been associated with the study of bacteria and fungi (Walters et al., 2016). Bacterial ecology in particular has seen the greatest advances and molecular tools for studying bacteria are the most developed. Massive collaborative projects such as the earth microbiome project (Gilbert et al., 2014) have revealed complex communities of bacteria in environments ranging from the surface of Antarctic snow pack (Michaud et al., 2014) to bare basalt at the bottom of the ocean (Santelli et al., 2008). The human microbiome project has revealed that most of the gene diversity associated with humans is actually encoded by symbiotic microbes (Peterson et al., 2009). Studies of the microbiome of plants, particularly crop plants, has now become commonplace. Many researchers hope advances in the understanding of microbiomes of agricultural plants will allow engineering microbial communities to suppress disease and promote nutrition in order to find sustainable replacements for artificial pesticides and fertilizers.

For this reason, elucidating the interacting effects of environment and host genotype on microbiome composition has been a priority (Lundberg et al., 2012; Vandenkoornhuyse et al., 2015). In general, current research suggests that the environment is the main driver of fungal and bacterial community composition in plants, but plant genotype also has an effect, especially as an interaction with environmental

differences. In our study on how the fungal and oomycete microbiome of *Rhododendron* varies with genotype and environment, we found that environment was the main driver and genotype had very little effect. For those systems where genotype is shown to have an effect, there is hope that plants could eventually be bred to encourage beneficial microbiomes. Even in systems where genotype does not have a distinct effect, there is potential to engineer microbes that could be applied to plants as biocontrols or as symbionts useful for making nutrients available to plants. If environmental conditions have a predictable effect on microbiome structure, then the environment could also be manipulated to result in a favorable microbiome composition.

6.2 The state of oomycete metabarcoding

While bacterial and fungal metabarcoding methods are mostly well developed at this point, methods to metabarcode oomycete communities are still quite experimental and no standard method exists. There have been various primers proposed based on modifications to fungal ribosomal primers, but these have either targeted only a small fraction of oomycete diversity (Legeay et al., 2019), have targeted regions with insufficient variation to distinguish species-level differences (Redekar et al., 2019), have produced an amplicon too long for Illuminia sequencers, or have had other technical difficulties. We recently developed metabarcoding primers for the *rps10* gene that should allow for the identification of the vast majority of oomycetes using an amplicon short enough to be sequenced on the Illumina MiSeq platform. We evaluated the relative performance of our *rps10*-based method with a current ITS1-

based method by sequencing both a mock community of known organisms and diverse environmental samples. Our results indicate that the *rps10* method produces fewer non-target sequences, fewer erroneous sequences, and has better ability to distinguish closely related species. Along with the sequencing primers, we have also compiled a manually curated database of reference sequences that can be used to assign taxonomy to sequences produced by the *rps10* method. To host this database, a website was created where any version of the database can be downloaded or searched. We hope this resource will be a major improvement in the methods used for characterizing oomycete communities.

6.3 The state of oomycete ecology

Although oomycete are present in nearly every ecosystem on earth and include some of the most devastating pathogens, they have been studied much less than bacteria and fungi (Davis, 2016). Major groups of oomycetes are barely known, although there is evidence that they are diverse and abundant in some ecosystems. Most well known are the terrestrial pathogens of plants, including *Phytophthora*, *Pythium*, and the downy mildews (Akino et al., 2014; Gessler et al., 2011). However, there are also diverse oomycete pathogens of other organisms that are less well known, such as those infecting algae (Sekimoto et al., 2009), fish (Van West, 2006), crustaceans (Tharp & Bland, 1977), diatoms (Klochkova et al., 2016), rotifiers (Molloy et al., 2014), nematodes (Glockling & Dick, 1997), and mammals (Spies et al., 2016). Even less studied are the various saprophytic oomycetes in aquatic and moist terrestrial

ecosystems (Blackwell et al., 2015). Some of these, such as pathogens of algae and diatoms affect the main primary producers in many ecosystems, making their influence on the ecosystem as a whole potentially quite large. Similarly, oomycetes are the primary saprobes in some ecosystems, such as *Halophytopthora* in mangrove swamps and *Pythium* in some moist soils (Bennett & Thines, 2019). In the past, the relatively few researchers studying oomycetes and the difficulty of morphological identification has slowed research into oomycete communities and restricted the focus to a few economically important pathogens (Hatai, 2012; Muraosa et al., 2009; Telle et al., 2011). Now however, molecular tools have lowered the barrier to entry for microbial ecologists and phylogeneticists to include oomycetes in their studies.

Molecular methods for oomycete community ecology are still in their infancy, but there is progress being made. As more attention is given to phylogenetic studies of oomycetes, much of the oomycete taxonomy will likely be revised in the near future. There are many studies of the phylogeny of economically important groups of plant pathogens, but less work has been done on relationships between major oomycete groups as a whole. A robust and stable taxonomy will be invaluable for classifying results of community ecology studies and a taxonomy that mirrors phylogeny will be the most useful for molecular methods. In addition, curated reference databases of sequences from known organisms are crucial to make use of methods like metabarcoding and metagenomics. It is likely that many new species and entirely new groups of oomycetes will be discovered as molecular community ecology techniques like metabarcoding are increasingly applied to natural habitats.

6.4 The need for modular open source tools

The study of biology has benefited greatly from the use of computers and programming languages, but as the complexity and scale of analyses increase, it is increasingly important that programs are modular, open source, and include extensive documentation. The ideal form of these modular tools are functions in programming languages or executable scripts with defined inputs and outputs. Since these modular tools are usually designed to be used in an automated fashion in scripts, their use vastly improves reproducibility of published research when the scripts are published with the results. Examples include packages for R and command line functions used in Linux/Unix systems. This kind of modularity allows tools to be chained together to make more complex tools and analyses customized for the problem being addressed. Such analyses and tools usually take the form of plain text files that are easily examined and can be run with a single command once the proper software is installed. This allows for a level of reproducibility not possible with "point and click" software, since the order of specific commands cannot typically be recorded and redone in an automated fashion. Modular tools reduce redundant effort for programmers since many complex tools require components that might not be the primary focus of the software. Instead of each program implementing separate solutions to these common problems, a single well-developed tool for that specific purpose can be used by multiple softwares. For example, the R package `readr` focuses on reading different types of file formats into R and is used by hundreds of other packages that would otherwise have to implement this functionality individually. This also makes these tools more

intuitive for users, since this tends to make different tools behave the same way when they do the same task. Such tools are most useful as open source code, since this allows programmers to build complex tools from more simple tools, which can then be used by even more complex tools, without the need to negotiate with the owners of the software.

All computational tools used for research should ideally be open source and version controlled to ensure maximum reproducibility and potential for reuse. If the researcher has a question about the details of how a specific function works or suspects a bug in the code, only open source software can be examined manually to address the issue. For proprietary software, the specific implementation of a function might be considered important intellectual property and will not likely be made available by the owners of the software. In addition, the implementation of features might change between versions without the knowledge of the researcher, leading to unexplained differences in results. It is common practice in open source software to be version controlled, so researchers can rerun analyses with the original version used if needed. The owners of proprietary software might stop maintaining widely used software, which could lead to bugs that can make the software unusable over time. In contrast, open source software can always be adopted by new maintainers if it is abandoned by the original creators. Finally open source software is typically free to use, so its use does not discriminate against poorer countries and allows anyone to attempt to reproduce results of a study regardless of wealth. This is particularly important for research that is used to inform public policy since it allows researchers and informed citizens to verify that an analysis was done correctly without paying

for software they would not otherwise need.

The tools `metacoder` and `taxa` presented here are examples of this kind of software. They were designed to be modular flexible tools rather than ready-made pipelines. `Taxa` in particular is meant to be a building block for other packages to use rather than a stand-alone tool set, although it can also function as such. `Metacoder` is structured as a set of discrete modular operations commonly needed to analyze metabarcoding data, rather than implementing ready-made analyses. This is meant to make it more flexible and useful by other packages. Both packages are open source, version controlled, and come with extensive included and online documentation.

6.5 The need for collaborative projects

The data produced by microbial community research are particularly challenging to analyze statistically and draw consistent conclusions from, but these difficulties might be addressed by massive collaborative studies with many samples to better capture complex ecological phenomena. Microbial communities often have hundreds to thousands of species in a single sample and associated with each species in each sample is a read count. Each species found in a study can be thought of as a variable and a single study rarely has more than a few hundred samples. Trying to draw inferences based on thousands of variables with hundreds of samples is not possible using most statistical techniques, so summary statistics are often analyzed instead, such as the alpha and beta diversity of samples. While summary statistics might reveal interesting patterns, they are often not what is most relevant to the biology

of the system and the priorities of the researcher. However, massive collaborative projects with dozens of labs might be able to sequence tens of thousands of samples and reveal patterns associated with complex ecological interactions among groups of microbes. Such studies could reveal things like the influence of host genotype on microbiome composition, which microbes are beneficial or damaging to which host under what conditions, or which assemblages of microbes are best suited for particular ecological functions, like biodegradation of pollutants. Knowledge of this kind would be useful for agriculture, biotechnology, and human health, among other areas. For subjects as complicated as microbial ecology, progress will likely be slow and inefficient as long as studies are restricted to single labs and a time span of only a few years.

Bibliography

Akino, S., Takemoto, D., & Hosaka, K. (2014). Phytophthora infestans: A review of past and current studies on potato late blight. *Journal of General Plant Pathology*, *80*(1), 24–37.

Akrofi, A. (2015). Phytophthora megakarya: A review on its status as a pathogen on cacao in west africa. *African Crop Science Journal*, *23*(1), 67–87.

Allen, B., Kon, M., & Bar-Yam, Y. (2009). A new phylogenetic diversity measure generalizing the shannon index and its application to phyllostomid bats. *The American Naturalist*, *174*(2), 236–243.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, *26*(1), 32–46.

Andersson, S. G., Zomorodipour, A., Andersson, J. O., Sicheritz-Pontén, T., Alsmark, U. C. M., Podowski, R. M., ... Kurland, C. G. (1998). The genome sequence of rickettsia prowazekii and the origin of mitochondria. *Nature*, *396*(6707), 133–140.

Andrivon, D. (1996). The origin of phytophthora infestans populations present in europe in the 1840s: A critical review of historical and scientific evidence. *Plant Pathology*, *45*(6), 1027–1035.

Arnold, A. E., Mejı'a, L. C., Kyllo, D., Rojas, E. I., Maynard, Z., Robbins, N., & Herre, E. A. (2003). Fungal endophytes limit pathogen damage in a tropical tree. *Proceedings of the National Academy of Sciences*, *100*(26), 15649–15654.

Atkins, D. (1954). A marine fungus plectospira dubia n. Sp.[Saprolegniaceae], infecting crustacean eggs and small crustacea. *Journal of the Marine Biological Association of the United Kingdom*, *33*(3), 721–732.

Augspurger, C. K., & Wilkinson, H. T. (2007). Host specificity of pathogenic pythium species: Implications for tree species diversity. *Biotropica*, *39*(6), 702–708.

Baldauf, S. L., Roger, A., Wenk-Siefert, I., & Doolittle, W. F. (2000). A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science*, *290*(5493), 972–977.

Bálint, M., Tiffin, P., Hallström, B., O'Hara, R. B., Olson, M. S., Fankhauser, J. D., ... Schmitt, I. (2013). Host genotype shapes the foliar fungal microbiome of balsam poplar (populus balsamifera). *PLoS One*, *8*(1), e53987.

Baxter, L., Tripathy, S., Ishaque, N., Boot, N., Cabral, A., Kemen, E., ... others. (2010). Signatures of adaptation to obligate biotrophy in the hyaloperonospora arabidopsidis genome. *Science*, *330*(6010), 1549–1551.

Beakes, G. W., Glockling, S. L., & James, T. Y. (2014). A new oomycete species parasitic in nematodes, chlamydomyzium dictyuchoides sp. Nov.: developmental biology and phylogenetic studies. *Fungal Biology*, *118*(7), 527–543.

Bellemain, E., Carlsen, T., Brochmann, C., Coissac, E., Taberlet, P., & Kauserud, H. (2010). ITS as an environmental dna barcode for fungi: An in silico approach reveals potential pcr biases. *BMC Microbiology*, *10*(1), 189.

Bennett, R., & Thines, M. (2019). Revisiting salisapiliaceae. *Fungal Systematics and Evolution*, *3*(1), 353–366.

Benson, D., Cavanaugh, M., Clark, K., Karsch, I. M., & DJ, L. (2013). J. Ostell, ew sayers, genbank. *Nucleic Acids Res*, *41*, D36–42.

Berendsen, R. L., Pieterse, C. M., & Bakker, P. A. (2012). The rhizosphere microbiome and plant health. *Trends in Plant Science*, *17*(8), 478–486.

Berg, G., & Raaijmakers, J. M. (2018). Saving seed microbiomes. *The ISME Journal*, *12*(5), 1167.

Blackwell, W. H., Letcher, P. M., & Powell, M. J. (2015). A review and update of the genus sapromyces (straminipila: Oomycota). *Phytologia*, *97*(2), 82–93.

Blazer, V. S., Lilley, J., Schill, W., Kiryu, Y., Densmore, C. L., Panyawachira, V., & Chinabut, S. (2002). Aphanomyces invadans in atlantic menhaden along the east coast of the united states. *Journal of Aquatic Animal Health*, *14*(1), 1–10.

Bock, C., Jeger, M., Fitt, B. D., & Sherington, J. (1997). Effect of wind on the dispersal of oospores of peronosclerospora sorghi from sorghum. *Plant Pathology*, *46*(3), 439–449.

Bokulich, N. A., Thorngate, J. H., Richardson, P. M., & Mills, D. A. (2014). Microbial biogeography of wine grapes is conditioned by cultivar, vintage, and climate. *Proceedings of the National Academy of Sciences*, *111*(1), E139–E148.

Bonito, G., Reynolds, H., Robeson, M. S., Nelson, J., Hodkinson, B. P., Tuskan, G., ... Vilgalys, R. (2014). Plant host and soil origin influence fungal and bacterial assemblages in the roots of woody plants. *Molecular Ecology*, *23*(13), 3356–3370.

Boursnell, J. G. (1950). The symbiotic seed-borne fungus in the cistaceae: I. Distribution and function of the fungus in the seeding and in the tissues of the mature plant. *Annals of Botany*, *14*(54), 217–243.

Brasier, C., & Webber, J. (2010). Sudden larch death. *Nature*, *466*(7308), 824–825.

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., ... Banfield, J. F. (2015). Unusual biology across a group comprising more than 15% of domain bacteria. *Nature*, *523*(7559), 208–211.

Brown, S. D., Collins, R. A., Boyer, S., LEFORT, M.-C., MALUMBRES-OLARTE, J., Vink, C. J., & Cruickshank, R. H. (2012). Spider: An r package for the analysis of species identity and evolution, with particular reference to dna barcoding. *Molecular Ecology Resources*, *12*(3), 562–565.

Bruzone, M. C., Fehrer, J., Fontenla, S. B., & Vohnı'k, M. (2017). First record of rhizoscyphus ericae in southern hemisphere's ericaceae. *Mycorrhiza*, *27*(2), 147–163.

Bukin, Y. S., Galachyants, Y. P., Morozov, I., Bukin, S., Zakharenko, A., & Zemskaya, T. (2019). The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Scientific Data*, *6*, 190007.

Burgess, T. I., Scott, J. K., Mcdougall, K. L., Stukely, M. J., Crane, C., Dunstan, W. A., ... others. (2017). Current and projected global distribution of phytophthora cinnamomi, one of the world's worst plant pathogens. *Global Change Biology*, *23*(4), 1661–1674.

Büttner, E., Gebauer, A. M., Hofrichter, M., Liers, C., & Kellner, H. (2018). Draft genome sequence of scytalidium lignicola dsm 105466, a ubiquitous saprotrophic fungus. *Microbiol Resour Announc*, *7*(14), e01208–18.

Cahill, D. M., Rookes, J. E., Wilson, B. A., Gibson, L., & McDougall, K. L. (2008). Phytophthora cinnamomi and australia's biodiversity: Impacts, predictions and

progress towards control. *Australian Journal of Botany*, *56*(4), 279–310.

Cairney, J. W., & Meharg, A. A. (2003). Ericoid mycorrhiza: A partnership that exploits harsh edaphic conditions. *European Journal of Soil Science*, *54*(4), 735–740.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from illumina amplicon data. *Nature Methods*, *13*(7), 581.

Carleson, N. C., Fieland, V., Scagel, C. R., Weiland, J. E., & Grünwald, N. J. (2018). Population structure of phytophthora plurivora on rhododendron in oregon nurseries. *Plant Disease*, (ja).

Chamberlain, S. A., & Boettiger, C. (2017). *R python, and ruby clients for gbif species occurrence data*. PeerJ Preprints.

Chamberlain, S. A., & Szöcs, E. (2013). Taxize: Taxonomic search and retrieval in r. *F1000Research*, *2*.

Chamberlain, S., Ram, K., Barve, V., & Mcglinn, D. (2017). Rgbif: Interface to the global 'biodiversity'Information facility "api". R package version 0.9. 8.

Chang, W., Cheng, J., Allaire, J., Xie, Y., & Mcpherson, J. (n.d.). Shiny: Web application framework for r. R package version 1.4. 0. 2019.

Chao, A., Chiu, C.-H., & Jost, L. (2010). Phylogenetic diversity measures based on hill numbers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1558), 3599–3609.

Charif, D., & Lobry, J. R. (2007). SeqinR 1.0-2: A contributed package to the r project for statistical computing devoted to biological sequences retrieval and analysis. In *Structural approaches to sequence evolution* (pp. 207–232). Springer.

Choi, Y.-J., Beakes, G., Glockling, S., Kruse, J., Nam, B., Nigrelli, L., ... others. (2015). Towards a universal barcode of oomycetes–a comparison of the cox1 and cox2 loci. *Molecular Ecology Resources*, *15*(6), 1275–1288.

Choi, Y.-J., Shin, H.-D., & Thines, M. (2009). The host range of albugo candida extends from brassicaceae through cleomaceae to capparaceae. *Mycological Progress*, *8*(4), 329.

Choi, Y.-J., Thines, M., Tek, M. P., & Shin, H.-D. (2012). Morphological evidence

supports the existence of multiple species in pustula (albuginaceae, oomycota). *Nova Hedwigia*, *94*(1), 181–192.

Clara, M. I. E. da, Almeida Ribeiro, N. M. C. de, & others. (2013). Decline of mediterranean oak trees and its association with phytophthora cinnamomi: A review. *European Journal of Forest Research*, *132*(3), 411–432.

Coince, A., Caël, O., Bach, C., Lengellé, J., Cruaud, C., Gavory, F., … Buée, M. (2013). Below-ground fine-scale distribution and soil versus fine root detection of fungal and soil oomycete communities in a french beech forest. *Fungal Ecology*, *6*(3), 223–235.

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., … others. (2009). The ribosomal database project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, *37*(suppl_1), D141–D145.

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., … Tiedje, J. M. (2014). Ribosomal database project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, *42*(D1), D633–D642.

Coleman-Derr, D., Desgarennes, D., Fonseca-Garcia, C., Gross, S., Clingenpeel, S., Woyke, T., … Tringe, S. G. (2016). Plant compartment and biogeography affect microbiome composition in cultivated and native agave species. *New Phytologist*, *209*(2), 798–811.

Cooke, D., Drenth, A., Duncan, J., Wagels, G., & Brasier, C. (2000). A molecular phylogeny of phytophthora and related oomycetes. *Fungal Genetics and Biology*, *30*(1), 17–32.

Cordier, T., Robin, C., Capdevielle, X., Desprez-Loustau, M.-L., & Vacher, C. (2012). Spatial variability of phyllosphere fungal assemblages: Genetic distance predominates over geographic distance in a european beech stand (fagus sylvatica). *Fungal Ecology*, *5*(5), 509–520.

Cristescu, M. E. (2014). From barcoding single individuals to metabarcoding biological communities: Towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution*, *29*(10), 566–571.

Cuenca, M. F. P. (2016). *Temporal and host related variation of pythium and globisporangium species in floricultural crops* (PhD thesis). Oklahoma State University.

Czeczuga, B., Mazalska, B., Godlewska, A., & Muszyńska, E. (2005). Aquatic fungi

growing on dead fragments of submerged plants. *Limnologica*, *35*(4), 283–297.

Davis, K. S. (2016). *Biodiversity of aquatic oomycetes in the falkland islands* (PhD thesis). University of Aberdeen.

Davison, E. (2015). How phytophthora cinnamomi became associated with the death of eucalyptus marginata–the early investigations into jarrah dieback. *Australasian Plant Pathology*, *44*(3), 263–271.

Derevnina, L., Petre, B., Kellner, R., Dagdas, Y. F., Sarowar, M. N., Giannakopoulou, A., ... others. (2016). Emerging oomycete threats to plants and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1709), 20150459.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., ... Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with arb. *Appl. Environ. Microbiol.*, *72*(7), 5069–5072.

De Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., ... others. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, *348*(6237), 1261605.

Dick, M. (1969). MORPHOLOGY and taxonomy of the oomycetes, with special reference to saprolegniaceae, leptomitaceae and pythiaceae: I. SEXUAL reproduction. *New Phytologist*, *68*(3), 751–775.

Dick, M. W. (2001). The lagenidiaceous fungi and similar organisms. In *Straminipilous fungi* (pp. 171–265). Springer.

Dighton, J., & Coleman, D. C. (1992). Phosphorus relations of roots and mycorrhizas of rhododendron maximum l. In the southern appalachians, north carolina. *Mycorrhiza*, *1*(4), 175–184.

Dixon, P. (2003). VEGAN, a package of r functions for community ecology. *Journal of Vegetation Science*, *14*(6), 927–930.

Douglas, W. Y., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., & Ding, Z. (2012). Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, *3*(4), 613–623.

Dreistadt, S. H. (2001). *Integrated pest management for floriculture and nurseries* (Vol. 3402). University of California Agriculture; Natural Resources.

Edgar, R. C. (2017). Accuracy of microbial community diversity estimated by closed- and open-reference otus. *PeerJ*, *5*, e3889.

Elbrecht, V., & Leese, F. (2017). PrimerMiner: An r package for development and in silico validation of dna metabarcoding primers. *Methods in Ecology and Evolution*, *8*(5), 622–626.

Farr, D. F., Esteban, H. B., & Palm, M. E. (1996). *Fungi on rhododendron: A world reference*. Parkway Publishers, Inc.

Fernández-Benéitez, M. J., Ortiz-Santaliestra, M. E., Lizana, M., & Diéguez-Uribeondo, J. (2008). Saprolegnia diclina: Another species responsible for the emergent disease "saprolegnia infections" in amphibians. *FEMS Microbiology Letters*, *279*(1), 23–29.

Ficetola, G. F., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessière, J., ... Pompanon, F. (2010). An in silico approach for the evaluation of dna barcodes. *BMC Genomics*, *11*(1), 434.

Fisher, W., Nilson, E., & Shleser, R. (1975). Effect of the fungus haliphthoros milfordensis on the juvenile stages of the american lobster homarus americanus. *Journal of Invertebrate Pathology*, *26*(1), 41–45.

Foster, Z. S., Chamberlain, S., & Grünwald, N. J. (2018). Taxa: An r package implementing data standards and methods for taxonomic data. *F1000Research*, *7*.

Foster, Z. S., Sharpton, T. J., & Grünwald, N. J. (2017). Metacoder: An r package for visualization and manipulation of community taxonomic diversity data. *PLoS Computational Biology*, *13*(2), e1005404.

Foster, Z., Weiland, J., Scagel, C., & Grünwald, N. (2020). The composition of the fungal and oomycete microbiome of rhododendron roots under varying growth conditions, nurseries, and cultivars. *Phytobiomes Journal*, PBIOMES–09.

Fry, W. (2008). Phytophthora infestans: The plant (and r gene) destroyer. *Molecular Plant Pathology*, *9*(3), 385–402.

Fry, W., Birch, P., Judelson, H., Grünwald, N., Danies, G., Everts, K., ... others. (2015). Five reasons to consider phytophthora infestans a reemerging pathogen. *Phytopathology*, *105*(7), 966–981.

Fuhrman, J. A., McCallum, K., & Davis, A. A. (1992). Novel major archaebacterial

group from marine plankton. *Nature*, *356*(6365), 148–149.

Gachon, C. M., Strittmatter, M., Badis, Y., Fletcher, K. I., West, P. V., & Müller, D. G. (2017). Pathogens of brown algae: Culture studies of anisolpidium ectocarpii and a. Rosenvingei reveal that the anisolpidiales are uniflagellated oomycetes. *European Journal of Phycology*, *52*(2), 133–148.

Gachon, C. M., Strittmatter, M., Müller, D. G., Kleinteich, J., & Küpper, F. C. (2009). Detection of differential host susceptibility to the marine oomycete pathogen eurychasma dicksonii by real-time pcr: Not all algae are equal. *Appl. Environ. Microbiol.*, *75*(2), 322–328.

Garvetto, A., Perrineau, M.-M., Dressler-Allame, M., Bresnan, E., & Gachon, C. M. (2020). "Ectrogella" parasitoids of the diatom licmophora sp. Are polyphyletic. *Journal of Eukaryotic Microbiology*, *67*(1), 18–27.

Gaulin, E., Jacquet, C., Bottin, A., & Dumas, B. (2007). Root rot disease of legumes caused by aphanomyces euteiches. *Molecular Plant Pathology*, *8*(5), 539–548.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., ... others. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, *5*(10), R80.

Gessler, C., Pertot, I., & Perazzolli, M. (2011). Plasmopara viticola: A review of knowledge on downy mildew of grapevine and effective disease management. *Phytopathologia Mediterranea*, *50*(1), 3–44.

Gilbert, J. A., Jansson, J. K., & Knight, R. (2014). The earth microbiome project: Successes and aspirations. *BMC Biology*, *12*(1), 69.

Glockling, S., & Dick, M. (1997). New species of chlamydomyzium from japan and pure culture of myzocytiopsis species. *Mycological Research*, *101*(7), 883–896.

Glockling, S. L., & Beakes, G. W. (2000). A review of the taxonomy, biology and infection strategies of "biflagellate holocarpic" parasites of nematodes. *Fungal Diversity*, *4*, 1–20.

Glockling, S. L., & Beakes, G. W. (2006). An ultrastructural study of development and reproduction in the nematode parasite myzocytiopsis vermicola. *Mycologia*, *98*(1), 1–15.

Gruenwald, N. J., Goss, E. M., & Press, C. M. (2008). Phytophthora ramorum: A

pathogen with a remarkably wide host range causing sudden oak death on oaks and ramorum blight on woody ornamentals. *Molecular Plant Pathology*, *9*(6), 729–740.

Grünwald, N. J., Garbelotto, M., Goss, E. M., Heungens, K., & Prospero, S. (2012). Emergence of the sudden oak death pathogen phytophthora ramorum. *Trends in Microbiology*, *20*(3), 131–138.

Grünwald, N. J., LeBoldus, J. M., & Hamelin, R. C. (2019). Ecology and evolution of the sudden oak death pathogen phytophthora ramorum. *Annual Review of Phytopathology*, *57*, 301–321.

Grünwald, N. J., Martin, F. N., Larsen, M. M., Sullivan, C. M., Press, C. M., Coffey, M. D., ... Parke, J. L. (2011). Phytophthora-id. Org: A sequence-based phytophthora identification tool. *Plant Disease*, *95*(3), 337–342.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., ... others. (2012). The protist ribosomal reference database (pr2): A catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, *41*(D1), D597–D604.

Haas, B. J., Kamoun, S., Zody, M. C., Jiang, R. H., Handsaker, R. E., Cano, L. M., ... others. (2009). Genome sequence and analysis of the irish potato famine pathogen phytophthora infestans. *Nature*, *461*(7262), 393–398.

Hansen, E. M., Reeser, P. W., & Sutton, W. (2012). Phytophthora beyond agriculture. *Annual Review of Phytopathology*, *50*, 359–378.

Harman, G. E. (2006). Overview of mechanisms and uses of trichoderma spp. *Phytopathology*, *96*(2), 190–194.

Hatai, K. (2012). Diseases of fish and shellfish caused by marine fungi. In *Biology of marine fungi* (pp. 15–52). Springer.

Himes, B. E., Jiang, X., Wagner, P., Hu, R., Wang, Q., Klanderman, B., ... others. (2014). RNA-seq transcriptome profiling identifies crispld2 as a glucocorticoid responsive gene that modulates cytokine function in airway smooth muscle cells. *PloS One*, *9*(6).

Holt, C., Foster, R., Daniels, C. L., Giezen, M. van der, Feist, S. W., Stentiford, G. D., & Bass, D. (2018). Halioticida noduliformans infection in eggs of lobster (homarus gammarus) reveals its generalist parasitic strategy in marine invertebrates. *Journal of Invertebrate Pathology*, *154*, 109–116.

Hosokawa, T., Kikuchi, Y., Nikoh, N., Shimada, M., & Fukatsu, T. (2006). Strict host-symbiont cospeciation and reductive genome evolution in insect gut bacteria. *PLoS Biology*, *4*(10), e337.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., ... others. (2016). A new view of the tree of life. *Nature Microbiology*, *1*(5), 16048.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., ... others. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207.

Jee, H., Ho, H., & Cho, W. (2000). Pythiogeton zeae sp. Nov. Causing root and basal stalk rot of corn in korea. *Mycologia*, *92*(3), 522–527.

Jeffers, S., Martin, S., & others. (1986). Comparison of two media selective for phytophthora and pythium species. *Plant Disease*, *70*(11), 1038–1043.

Jeronimo, G. H., Jesus, A. L., Rocha, S. C., Goncalves, D. R., & Pires-Zottarelli, C. L. (2017). New insights into plectospira genus (oomycetes, straminipila): Morphological and molecular analyses. *Phytotaxa*, *307*(3), 191–198.

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., ... others. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, *16*(10), 1245–1257.

Jones, R. K., Benson, D. M., & others. (2001). *Diseases of woody ornamentals and trees in nurseries.* American Phytopathological Society (APS Press).

Jung, T., Durßn, A., Sanfuentes von Stowasser, E., Schena, L., Mosca, S., Fajardo, S., ... others. (2018). Diversity of phytophthora species in valdivian rainforests and association with severe dieback symptoms. *Forest Pathology*, *48*(5), e12443.

Jung, T., Scanu, B., Bakonyi, J., Seress, D., Kovács, G., Durán, A., ... others. (2017). Nothophytophthora gen. Nov., a new sister genus of phytophthora from natural and semi-natural ecosystems. *Persoonia: Molecular Phylogeny and Evolution of Fungi*, *39*, 143.

Kamoun, S., Furzer, O., Jones, J. D., Judelson, H. S., Ali, G. S., Dalio, R. J., ... others. (2015). The top 10 oomycete pathogens in molecular plant pathology. *Molecular Plant Pathology*, *16*(4), 413–434.

Katoh, K., Kuma, K.-i., Toh, H., & Miyata, T. (2005). MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, *33*(2), 511–518.

Kerwin, J. L. (2007). Oomycetes: Lagenidium giganteum. *Journal of the American Mosquito Control Association*, *23*(sp2), 50–57.

Klochkova, T. A., Shim, J. B., Hwang, M. S., & Kim, G. H. (2012). Host–parasite interactions and host species susceptibility of the marine oomycete parasite, olpidiopsis sp., from korea that infects red algae. *Journal of Applied Phycology*, *24*(1), 135–144.

Klochkova, T. A., Shin, Y. J., Moon, K.-H., Motomura, T., & Kim, G. H. (2016). New species of unicellular obligate parasite, olpidiopsis pyropiae sp. Nov., that plagues pyropia sea farms in korea. *Journal of Applied Phycology*, *28*(1), 73–83.

Knaus, B., Fieland, V., Graham, K., & Grünwald, N. (2015). Diversity of foliar phytophthora species on rhododendron in oregon nurseries. *Plant Disease*, *99*(10), 1326–1332.

Kõljalg, U., Larsson, K.-H., Abarenkov, K., Nilsson, R. H., Alexander, I. J., Eberhardt, U., ... others. (2005). UNITE: A database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytologist*, *166*(3), 1063–1068.

Kõljalg, U., Nilsson, R. H., Abarenkov, K., Tedersoo, L., Taylor, A. F., Bahram, M., ... others. (2013). Towards a unified paradigm for sequence-based identification of fungi. *Molecular Ecology*, *22*(21), 5271–5277.

Kowalik, M., Kierpiec-Baran, B., & Duda-Franiak, K. (2015). Micromycetes colonizing and damaging leaves of evergreen rhododendron (rhododendron l.) in nursery. *Acta Agrobotanica*, *68*(2), 179–185.

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*(2), 115–129.

Lava, S. S., Heller, A., & Spring, O. (2013). Oospores of pustula helianthicola in sunflower seeds and their role in the epidemiology of white blister rust. *IMA Fungus*, *4*(2), 251.

Leafio, E., Jones, E., & Vrijmoed, L. (2000). Why are halophytophthora species well adapted to mangrove habitats. *Aquatic Mycology Across the Millenium*, 125–145.

Lebeda, A., & Cohen, Y. (2011). Cucurbit downy mildew (pseudoperonospora cubensis)—biology, ecology, epidemiology, host-pathogen interaction and control. *European Journal of Plant Pathology*, *129*(2), 157–192.

Legeay, J., Husson, C., Cordier, T., Vacher, C., Marcais, B., & Buée, M. (2019). Comparison and validation of oomycetes metabarcoding primers for phytophthora high throughput sequencing. *Journal of Plant Pathology*, *101*(3), 743–748.

Lundberg, D. S., Lebeis, S. L., Paredes, S. H., Yourstone, S., Gehring, J., Malfatti, S., ... others. (2012). Defining the core arabidopsis thaliana root microbiome. *Nature*, *488*(7409), 86–90.

Maharachchikumbura, S. S., Hyde, K. D., Groenewald, J. Z., Xu, J., & Crous, P. W. (2014). Pestalotiopsis revisited. *Studies in Mycology*, *79*, 121–186.

Maidak, B. L., Olsen, G. J., Larsen, N., Overbeek, R., McCaughey, M. J., & Woese, C. R. (1996). The ribosomal database project (rdp). *Nucleic Acids Research*, *24*(1), 82–85.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 50–60.

Martin, F. N., Blair, J. E., & Coffey, M. D. (2014). A combined mitochondrial and nuclear multilocus phylogeny of the genus phytophthora. *Fungal Genetics and Biology*, *66*, 19–32.

Martin, F. N., & Loper, J. E. (1999). Soilborne plant diseases caused by pythium spp.: ecology, epidemiology, and prospects for biological control. *Critical Reviews in Plant Sciences*, *18*(2), 111–181.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, *17*(1), 10–12.

Matari, N. H., & Blair, J. E. (2014). A multilocus timescale for oomycete evolution estimated under three distinct molecular clock models. *BMC Evolutionary Biology*, *14*(1), 101.

McCarthy, C. G., & Fitzpatrick, D. A. (2017). Phylogenomic reconstruction of the oomycete phylogeny derived from 37 genomes. *Msphere*, *2*(2).

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst,

A., ... Hugenholtz, P. (2012). An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, *6*(3), 610.

McMullan, M., Gardiner, A., Bailey, K., Kemen, E., Ward, B. J., Cevik, V., ... others. (2015). Evidence for suppression of immunity as a driver for genomic introgressions and host range expansion in races of albugo candida, a generalist parasite. *Elife*, *4*, e04550.

McMurdie, P. J., & Holmes, S. (2013). Phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. *PloS One*, *8*(4), e61217.

Mendes, R., Kruijt, M., De Bruijn, I., Dekkers, E., Voort, M. van der, Schneider, J. H., ... others. (2011). Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science*, *332*(6033), 1097–1100.

Mendoza, L., Hernandez, F., & Ajello, L. (1993). Life cycle of the human and animal oomycete pathogen pythium insidiosum. *Journal of Clinical Microbiology*, *31*(11), 2967–2973.

Michaud, L., Giudice, A. L., Mysara, M., Monsieurs, P., Raffa, C., Leys, N., ... Van Houdt, R. (2014). Snow surface microbiome on the high antarctic plateau (dome c). *PloS One*, *9*(8).

Misra, R. S., Sharma, K., & Mishra, A. K. (2008). Phytophthora leaf blight of taro (colocasia esculenta)—a review. *Asian Australas J Plant Sci Biotechnol*, *2*, 55–63.

Molloy, D. P., Glockling, S. L., Siegfried, C. A., Beakes, G. W., James, T. Y., Mastitsky, S. E., ... Nemeth, M. J. (2014). Aquastella gen. Nov.: a new genus of saprolegniaceous oomycete rotifer parasites related to aphanomyces, with unique sporangial outgrowths. *Fungal Biology*, *118*(7), 544–558.

Muraosa, Y., Morimoto, K., Sano, A., Nishimura, K., & Hatai, K. (2009). A new peronosporomycete, halioticida noduliformans gen. Et sp. Nov., isolated from white nodules in the abalone haliotis spp. From japan. *Mycoscience*, *50*(2), 106–115.

Nakamura, K., & Hatai, K. (1994). Atkinsiella parasitica sp. Nov. Isolated from a rotifer, brachionus plicatilis. *Mycoscience*, *35*(4), 383–389.

Nakamura, K., & Hatai, K. (1995). Atkinsiella dubia and its related species. *Myco-*

science, *36*(4), 431–438.

Nam, B., & Choi, Y.-J. (2019). Phytopythium and pythium species (oomycota) isolated from freshwater environments of korea. *Mycobiology*, *47*(3), 261–272.

Narisawa, K., & others. (2017). The dark septate endophytic fungus phialocephala fortinii is a potential decomposer of soil organic compounds and a promoter of asparagus officinalis growth. *Fungal Ecology*, *28*, 1–10.

Newell, S., & Fell, J. (1992). Distribution and experimental responses to substrate of marine oomycetes (halophytophthora spp.) in mangrove ecosystems. *Mycological Research*, *96*(10), 851–856.

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

Nichols, R. V., Vollmers, C., Newsom, L. A., Wang, Y., Heintzman, P. D., Leighton, M., ... Shapiro, B. (2018). Minimizing polymerase biases in metabarcoding. *Molecular Ecology Resources*.

Nilsson, R. H., Anslan, S., Bahram, M., Wurzbacher, C., Baldrian, P., & Tedersoo, L. (2018). Mycobiome diversity: High-throughput sequencing and identification of fungi. *Nature Reviews Microbiology*, 1.

Nilsson, R. H., Ryberg, M., Kristiansson, E., Abarenkov, K., Larsson, K.-H., & Kõljalg, U. (2006). Taxonomic reliability of dna sequences in public sequence databases: A fungal perspective. *PloS One*, *1*(1).

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'hara, R., ... others. (2013). Package "vegan". *Community Ecology Package, Version*, *2*(9), 1–295.

Orchard, S., Hilton, S., Bending, G. D., Dickie, I. A., Standish, R. J., Gleeson, D. B., ... others. (2017). Fine endophytes (glomus tenue) are related to mucoromycotina, not glomeromycota. *New Phytologist*, *213*(2), 481–486.

Owczarzy, R., Tataurov, A. V., Wu, Y., Manthey, J. A., McQuisten, K. A., Almabrazi, H. G., ... others. (2008). IDT scitools: A suite for analysis and design of nucleic acid oligomers. *Nucleic Acids Research*, *36*(suppl_2), W163–W169.

Pages, H., Aboyoun, P., Gentleman, R., & DebRoy, S. (2009). String objects repre-

senting biological sequences, and matching algorithms. *R Package Version*, *2*(2).

Panke-Buisse, K., Poole, A. C., Goodrich, J. K., Ley, R. E., & Kao-Kniffin, J. (2015). Selection on soil microbiomes reveals reproducible impacts on plant function. *The ISME Journal*, *9*(4), 980.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in r language. *Bioinformatics*, *20*(2), 289–290.

Park, J., Park, B., Veeraraghavan, N., Jung, K., Lee, Y.-H., Blair, J. E., ... others. (2008). Phytophthora database: A forensic database supporting the identification and monitoring of phytophthora. *Plant Disease*, *92*(6), 966–972.

Parke, J. L., & Grünwald, N. J. (2012). A systems approach for management of pests and pathogens of nursery crops. *Plant Disease*, *96*(9), 1236–1244.

Parke, J. L., Knaus, B. J., Fieland, V. J., Lewis, C., & Grünwald, N. J. (2014). Phytophthora community structure analyses in oregon nurseries inform systems approaches to disease management. *Phytopathology*, *104*(10), 1052–1062.

Pelizza, S. A., LASTRA, C. C. L., Becnel, J. J., Bisaro, V., & Garcia, J. J. (2007). Biotic and abiotic factors affecting leptolegnia chapmanii infection in aedes aegypti. *Journal of the American Mosquito Control Association*, *23*(2), 177–181.

Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A., ... others. (2009). The nih human microbiome project. *Genome Research*, *19*(12), 2317–2323.

Phillips, A. J., Anderson, V. L., Robertson, E. J., Secombes, C. J., & Van West, P. (2008). New insights into animal pathogenic oomycetes. *Trends in Microbiology*, *16*(1), 13–19.

Ploch, S., Telle, S., Choi, Y.-J., Cunnington, J. H., Priest, M., Rost, C., ... Thines, M. (2011). The molecular phylogeny of the white blister rust genus pustula reveals a case of underestimated biodiversity with several undescribed species on ornamentals and crop plants. *Fungal Biology*, *115*(3), 214–219.

Ponce de León, I. (2011). The moss physcomitrella patens as a model system to study interactions between plants and phytopathogenic fungi and oomycetes. *Journal of Pathogens*, *2011*.

Prigigallo, M. I., Abdelfattah, A., Cacciola, S. O., Faedda, R., Sanzani, S. M., Cooke,

D. E., & Schena, L. (2016). Metabarcoding analysis of phytophthora diversity using genus-specific primers and 454 pyrosequencing. *Phytopathology*, *106*(3), 305–313.

Prince, A. M., & Andrus, L. (1992). PCR: How to kill unwanted dna. *Biotechniques*, *12*(3), 358–360.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., ... Glöckner, F. O. (2012). The silva ribosomal rna gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), D590–D596.

Radmer, L., Anderson, G., Malvick, D., Kurle, J., Rendahl, A., & Mallik, A. (2017). Pythium, phytophthora, and phytopythium spp. Associated with soybean in minnesota, their relative aggressiveness on soybean and corn, and their sensitivity to seed treatment fungicides. *Plant Disease*, *101*(1), 62–72.

Rasconi, S., Jobard, M., & Sime-Ngando, T. (2011). Parasitic fungi of phytoplankton: Ecological roles and implications for microbial food webs. *Aquatic Microbial Ecology*, *62*(2), 123–137.

Rattan, S., Muhsin, T. M., & Ismail, A. (1978). Aquatic fungi of iraq: Species of dictyuchus and calyptralegnia. *Sydowia*, *31*, 112–121.

Redekar, N. R., Eberhart, J. L., & Parke, J. L. (2019). Diversity of phytophthora, pythium, and phytopythium species in recycled irrigation water in a container nursery. *Phytobiomes*, (ja).

Riddell, C. E., Frederickson-Matika, D., Armstrong, A. C., Elliot, M., Forster, J., Hedley, P. E., ... others. (2019). Metabarcoding reveals a high diversity of woody host-associated phytophthora spp. In soils at public gardens and amenity woodlands in britain. *PeerJ*, *7*, e6931.

Riit, T., Tedersoo, L., Drenkhan, R., Runno-Paurson, E., Kokko, H., & Anslan, S. (2016). Oomycete-specific its primers for identification and metabarcoding. *MycoKeys*, *14*, 17.

Rizzo, D. M., Garbelotto, M., & Hansen, E. M. (2005). Phytophthora ramorum: Integrative research and management of an emerging pathogen in california and oregon forests. *Annu. Rev. Phytopathol.*, *43*, 309–335.

Robideau, G. P., Cock, A. W. de, Coffey, M. D., Voglmayr, H., Brouwer, H., Bala, K., ... others. (2011). DNA barcoding of oomycetes with cytochrome c oxidase

subunit i and internal transcribed spacer. *Molecular Ecology Resources*, *11*(6), 1002–1011.

Rocha, J., Sousa, N., Santos, L., Pereira, A., Negreiros, N., Sales, P., & Trindade Júnior, O. (2014). The genus pythiogeton (pythiogetonaceae) in brazil. *Mycosphere*, *5*(5), 623–634.

Rodriguez, R. J., Henson, J., Van Volkenburgh, E., Hoy, M., Wright, L., Beckwith, F., ... Redman, R. S. (2008). Stress tolerance in plants via habitat-adapted symbiosis. *The ISME Journal*, *2*(4), 404–416.

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, *4*, e2584.

Rosshart, S. P., Vassallo, B. G., Angeletti, D., Hutchinson, D. S., Morgan, A. P., Takeda, K., ... others. (2017). Wild mouse gut microbiota promotes host fitness and improves disease resistance. *Cell*.

Rost, C., & Thines, M. (2012). A new species of pustula (oomycetes, albuginales) is the causal agent of sunflower white rust. *Mycological Progress*, *11*(2), 351–359.

Ruthig, G. R. (2009). Water molds of the genera saprolegnia and leptolegnia are pathogenic to the north american frogs rana catesbeiana and pseudacris crucifer, respectively. *Diseases of Aquatic Organisms*, *84*(3), 173–178.

Sakr, N., & others. (2014). Evolution of new plasmopara halstedii races under the selection pressure with resistant sunflower plants: A review. *Hellenic Plant Protect. J*, *7*(1), 1–13.

Salamone, I. G. de, Döbereiner, J., Urquiaga, S., & Boddey, R. (1996). Biological nitrogen fixation in azospirillum strain-maize genotype associations as evaluated by the 15N isotope dilution technique. *Biology and Fertility of Soils*, *23*(3), 249–256.

SantaLucia, J. (1998). A unified view of polymer, dumbbell, and oligonucleotide dna nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences*, *95*(4), 1460–1465.

Santelli, C. M., Orcutt, B. N., Banning, E., Bach, W., Moyer, C. L., Sogin, M. L., ... Edwards, K. J. (2008). Abundance and diversity of microbial life in ocean crust. *Nature*, *453*(7195), 653–656.

Sapkota, R., Knorr, K., Jørgensen, L. N., O'Hanlon, K. A., & Nicolaisen, M. (2015).

Host genotype is an important determinant of the cereal phyllosphere mycobiome. *New Phytologist*, *207*(4), 1134–1144.

Sapkota, R., & Nicolaisen, M. (2015). An improved high throughput sequencing method for studying oomycete communities. *Journal of Microbiological Methods*, *110*, 33–39.

Savory, E. A., Granke, L. L., QUESADA-OCAMPO, L. M., Varbanova, M., Hausbeck, M. K., & Day, B. (2011). The cucurbit downy mildew pathogen pseudoperonospora cubensis. *Molecular Plant Pathology*, *12*(3), 217–226.

Schappe, T., Albornoz, F. E., Turner, B. L., Neat, A., Condit, R., & Jones, F. A. (2017). The role of soil chemistry and plant neighbourhoods in structuring fungal communities in three panamanian rainforests. *Journal of Ecology*, *105*(3), 569–579.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., ... others. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, *75*(23), 7537–7541.

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., ... others. (2012). Nuclear ribosomal internal transcribed spacer (its) region as a universal dna barcode marker for fungi. *Proceedings of the National Academy of Sciences*, *109*(16), 6241–6246.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., & Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, *9*(8), 811.

Sekimoto, S., Klochkova, T. A., West, J. A., Beakes, G. W., & Honda, D. (2009). Olpidiopsis bostrychiae sp. Nov.: an endoparasitic oomycete that infects bostrychia and other red algae (rhodophyta). *Phycologia*, *48*(6), 460–472.

Seymour, R. L. (1984). Leptolegnia chapmanii, an oomycete pathogen of mosquito larvae. *Mycologia*, *76*(4), 670–674.

Shearer, B. L., & Tippett, J. T. (1989). *Jarrah dieback: The dynamics and management of phytophthora cinnamomi in the jarrah (eucalyptus marginata) forest of south-western australia* (Vol. 3). Department of Conservation; Land Management Perth.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., ... others. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular Systems Biology*, *7*(1).

Smith, S. E., & Read, D. J. (2010). *Mycorrhizal symbiosis*. Academic press.

Sökücü, A., & Thines, M. (2014). A molecular phylogeny of basidiophora reveals several apparently host-specific lineages on astereae. *Mycological Progress*, *13*(4), 1002.

Sparrow, F. (1976). The present status of classification in biflagellate fungi. *Recent Advances in Aquatic Mycology*, 213–222.

Spencer, M. A., & others. (2002). Revision of aplanopsis, pythiopsis, and 'subcentric'Achlya species (saprolegniaceae) using 18S rDNA and morphological data. *Mycological Research*, *106*(5), 549–560.

Spies, C. F., Grooters, A. M., Lévesque, C. A., Rintoul, T. L., Redhead, S. A., Glockling, S. L., ... De Cock, A. W. (2016). Molecular phylogeny and taxonomy of lagenidium-like oomycetes pathogenic to mammals. *Fungal Biology*, *120*(8), 931–947.

Spring, O., Gomez-Zeledon, J., Hadziabdic, D., Trigiano, R. N., Thines, M., & Lebeda, A. (2018). Biological characteristics and assessment of virulence diversity in pathosystems of economically important biotrophic oomycetes. *Critical Reviews in Plant Sciences*, *37*(6), 439–495.

Steciow, M. M., Lara, E., Pillonel, A., Pelizza, S. A., Lestani, E. A., Rossi, G. C., & Belbahri, L. (2013). Incipient loss of flagella in the genus geolegnia: The emergence of a new clade within leptolegnia? *IMA Fungus*, *4*(2), 169–175.

Studholme, D. J., Panda, P., Sanfuentes Von Stowasser, E., González, M., Hill, R., Sambles, C., ... McDougal, R. L. (2019). Genome sequencing of oomycete isolates from chile supports the new zealand origin of phytophthora kernoviae and makes available the first nothophytophthora sp. Genome. *Molecular Plant Pathology*, *20*(3), 423–431.

Swiegers, J., Bartowsky, E., Henschke, P., & Pretorius, I. (2005). Yeast and bacterial modulation of wine aroma and flavour. *Australian Journal of Grape and Wine Research*, *11*(2), 139–173.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). To-

wards next-generation biodiversity assessment using dna metabarcoding. *Molecular Ecology*, *21*(8), 2045–2050.

Tedersoo, L., Drenkhan, R., Anslan, S., Morales-Rodriguez, C., & Cleary, M. (2019). High-throughput identification and diagnostics of pathogens and pests: Overview and practical recommendations. *Molecular Ecology Resources*, *19*(1), 47–76.

Telle, S., Shivas, R. G., Ryley, M. J., & Thines, M. (2011). Molecular phylogenetic analysis of peronosclerospora (oomycetes) reveals cryptic species and genetically distinct species parasitic to maize. *European Journal of Plant Pathology*, *130*(4), 521–528.

Tharp, T., & Bland, C. (1977). Biology and host range of haliphthoros milfordensis. *Canadian Journal of Botany*, *55*(23), 2936–2944.

Thines, M. (2014). Phylogeny and evolution of plant pathogenic oomycetes—a global overview. *European Journal of Plant Pathology*, *138*(3), 431–447.

Thines, M., & Choi, Y.-J. (2016). Evolution, diversity, and taxonomy of the peronosporaceae, with focus on the genus peronospora. *Phytopathology*, *106*(1), 6–18.

Thines, M., Telle, S., Choi, Y.-J., Tan, Y. P., & Shivas, R. G. (2015). Baobabopsis, a new genus of graminicolous downy mildews from tropical australia, with an updated key to the genera of downy mildews. *IMA Fungus*, *6*(2), 483–491.

Tippmann, S. (2015). Programming tools: Adventures with r. *Nature News*, *517*(7532), 109.

Tringe, S. G., & Hugenholtz, P. (2008). A renaissance for the pioneering 16S rRNA gene. *Current Opinion in Microbiology*, *11*(5), 442–446.

Tsirigoti, A., Beakes, G. W., Hervé, C., Gachon, C. M., & Katsaros, C. (2015). Attachment, penetration and early host defense mechanisms during the infection of filamentous brown algae by eurychasma dicksonii. *Protoplasma*, *252*(3), 845–856.

Tyler, B. M. (2001). Genetics and genomics of the oomycete–host interface. *TRENDS in Genetics*, *17*(11), 611–614.

Tyler, B. M. (2007). Phytophthora sojae: Root rot pathogen of soybean and model oomycete. *Molecular Plant Pathology*, *8*(1), 1–8.

Uroz, S., Buée, M., Murat, C., Frey-Klett, P., & Martin, F. (2010). Pyrosequencing

reveals a contrasted bacterial diversity between oak rhizosphere and surrounding soil. *Environmental Microbiology Reports*, *2*(2), 281–288.

Uzuhashi, S., Hata, K., Matsuura, S., & Tojo, M. (2017). Globisporangium oryzicola sp. Nov., causing poor seedling establishment of directly seeded rice. *Antonie van Leeuwenhoek*, *110*(4), 543–552.

Vandenkoornhuyse, P., Quaiser, A., Duhamel, M., Le Van, A., & Dufresne, A. (2015). The importance of the microbiome of the plant holobiont. *New Phytologist*, *206*(4), 1196–1206.

Vano, I., Sakamoto, K., Inubushi, K., & others. (2011). Phylogenetic relationship among non-pathogenic isolates of dark septate endophytes from ericaceae plants. *Hort Research*, *65*, 41–47.

Van West, P. (2006). Saprolegnia parasitica, an oomycete pathogen with a fishy appetite: New challenges for an old problem. *Mycologist*, *20*(3), 99–104.

Vilela, R., Humber, R. A., Taylor, J. W., & Mendoza, L. (2019). Phylogenetic and physiological traits of oomycetes originally identified as lagenidium giganteum from fly and mosquito larvae. *Mycologia*, *111*(3), 408–422.

Vohnı'k, M., & Albrechtová, J. (2011). The co-occurrence and morphological continuum between ericoid mycorrhiza and dark septate endophytes in roots of six european rhododendron species. *Folia Geobotanica*, *46*(4), 373–386.

Vohnı'k, M., Albrechtová, J., Vosátka, M., & others. (2005). The inoculation with oidiodendron maius and phialocephala fortinii alters phosphorus and nitrogen uptake, foliar c: N ratio and root biomass distribution in rhododendron cv. Azurro. *Symbiosis*, *40*(2), 87–96.

Vrålstad, T., Schumacher, T., & Taylor, A. F. (2002). Mycorrhizal synthesis between fungal strains of the hymenoscyphus ericae aggregate and potential ectomycorrhizal and ericoid hosts. *New Phytologist*, *153*(1), 143–152.

Wagner, M. R., Lundberg, D. S., Tijana, G., Tringe, S. G., Dangl, J. L., & Mitchell-Olds, T. (2016). Host genotype and age shape the leaf and root microbiomes of a wild perennial plant. *Nature Communications*, *7*, 12151.

Wallace, E. C., Salgado-Salazar, C., Gregory, N. F., & Crouch, J. A. (2018). Basidiophora delawarensis, a new downy mildew species infecting cultivated goldenrod (solidago sphacelata) in the usa. *Mycological Progress*, *17*(12), 1283–1291.

Walters, W., Hyde, E. R., Berg-Lyons, D., Ackermann, G., Humphrey, G., Parada, A., ... others. (2016). Improved bacterial 16S rRNA gene (v4 and v4-5) and fungal internal transcribed spacer marker gene primers for microbial community surveys. *Msystems*, *1*(1), e00009–15.

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, *73*(16), 5261–5267.

Weiland, J. E., Scagel, C. F., Grünwald, N. J., Davis, E. A., Beck, B. R., & Fieland, V. J. (2018). Variation in disease severity caused by phytophthora cinnamomi, p. Plurivora, and pythium cryptoirregulare on two rhododendron cultivars. *Plant Disease*, *102*(12), 2560–2570.

Werres, S., Marwitz, R., In't Veld, W. A. M., De Cock, A. W., Bonants, P. J., De Weerdt, M., ... Baayen, R. P. (2001). Phytophthora ramorum sp. Nov., a new pathogen on rhododendron and viburnum. *Mycological Research*, *105*(10), 1155–1165.

West, J. A., Klochkova, T. A., Kim, G. H., & Loiseaux-de Goër, S. (2006). Olpidiopsis sp., an oomycete from madagascar that infects bostrychia and other red algae: Host species susceptibility. *Phycological Research*, *54*(1), 72–85.

White, T. J., Bruns, T., Lee, S., Taylor, J., & others. (1990). Amplification and direct sequencing of fungal ribosomal rna genes for phylogenetics. *PCR Protocols: A Guide to Methods and Applications*, *18*(1), 315–322.

Whittaker, R. H. (1960). Vegetation of the siskiyou mountains, oregon and california. *Ecological Monographs*, *30*(3), 279–338.

Wickham, H., Francois, R., Henry, L., Müller, K., & others. (2015). Dplyr: A grammar of data manipulation. *R Package Version 0.4*, *3*.

Wills, R. (1993). The ecological impact of phytophthora cinnamomi in the stirling range national park, western australia. *Australian Journal of Ecology*, *18*(2), 145–159.

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R markdown: The definitive guide*. CRC Press.

Yang, X., & Hong, C. (2014). Halophytophthora fluviatilis sp. Nov. From freshwater in virginia. *FEMS Microbiology Letters*, *352*(2), 230–237.

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K.-H., ... Rosselló-Móra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology*, *12*(9), 635–645.

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., ... Glöckner, F. O. (2014). The silva and "all-species living tree project (ltp)" taxonomic frameworks. *Nucleic Acids Research*, *42*(D1), D643–D648.

Yoon, H. S., Hackett, J. D., Pinto, G., & Bhattacharya, D. (2002). The single, ancient origin of chromist plastids. *Journal of Phycology*, *38*, 40–40.

Yoshida, K., Schuenemann, V. J., Cano, L. M., Pais, M., Mishra, B., Sharma, R., ... others. (2013). The rise and fall of the phytophthora infestans lineage that triggered the irish potato famine. *Elife*, *2*, e00731.

Yuan, X., Feng, C., Zhang, Z., & Zhang, C. (2017). Complete mitochondrial genome of phytophthora nicotianae and identification of molecular markers for the oomycetes. *Frontiers in Microbiology*, *8*, 1484.

Zhang, Y., Maharachchikumbura, S. S., Tian, Q., Hyde, K. D., & others. (2013). Pestalotiopsis species on ornamental plants in yunnan province, china. *Sydowia*, *65*(1), 113–128.