AN ABSTRACT OF THE DISSERTATION OF

Alija Bajro Mujic for the degree of Doctor of Philosophy in Botany and Plant Pathology presented on March 13, 2015

Title: Symbiosis in the Pacific Ring of Fire: Evolutionary-Biology of *Rhizopogon* Subgenus *Villosuli* as Mutualists of *Pseudotsuga*.

Abstract approved: _____

Joseph W. Spatafora

Rhizopogon is a large genus of ectomycorrhizal (EM) fungi that grows in association with host trees of the family Pinaceae. We have conducted a series of studies investigating the ecology and evolutionary biology of the EM symbiosis shared between *R.* subgenus *Villosuli* and trees of the genus *Pseudotsuga*. Two members of *R.* subgenus *Villosuli,* the sister taxa *R. vinicolor* and *R. vesiculosus,* are the primary focus of the first two investigations presented in this dissertation. These two species share a sympatric range in the Pacific Northwest where they share a host association with *Pseudotsuga menziesii* (Douglas fir). In Chapter two we perform a series of container plant experiments, which manipulate substrate stratification and competitive conditions, in order to investigate ecological factors influencing vertical resource partitioning between these species. From the findings of this work we conclude that priority effects and substrate heterogeneity allow *R. vinicolor* to maintain EM root tips colonized during independent mating events in the face of competition from individuals of *R. vesiculosus,* which compete for root tips through vegetative expansion. The need for high rates of mating in *R. vinicolor* populations to successfully compete with *R. vesiculosus* may be

driving microevolutionary adaptation in these species and our second investigation seeks to elucidate the genetic mechanisms that underlie this adaptation.

In Chapter three we report the draft genomes of *R. vinicolor* and *R. vesiculosus*, which were sequenced in order to characterize the structure and gene content of the loci responsible for mating-type recognition in these fungi. The functional traits of mating-type loci can have dramatic effects upon the population structure and adaptability of fungi and we have sought to characterize these loci to better understand the role of mating in competitive interactions between these species. We have found that the differential life history strategies that shape the population structure of *R. vinicolor* and *R. vesiculosus* correlate with differential structure of the mating type B-locus between these species. The B-locus is involved in recognition of compatible mates through the production of lipopeptide pheromones and complementary pheromone receptors. *R. vinicolor* possesses a greater diversity of both pheromones and receptors than *R. vesiculosus* and this property is likely reinforced by *R. vinicolor's* need for frequent mating events to compete successfully with *R. vesiculosus.* The loss of diversity in this genomic region is a derived trait in *R. vesiculosus,* which competes successfully through vegetative expansion of a few individuals independent of the frequency of mating.

Rhizopogon species possess one of the highest rates of host specificity of any EM fungus and we hypothesize that *R.* subgenus *Villosuli* has undergone a process of comigration and cospeciation with their *Pseudotsuga* hosts. Chapter four seeks to test these hypotheses by conducting a genome-scale phylogeographical analysis of *Rhizopogon* species growing in association with *Pseudotsuga* species. We have performed a multinational and intercontinental sampling of *Rhizopogon* species from the

Pacific Rim distribution of all *Pseudotsuga* species and sequenced low coverage genomes of *Rhizopogon* species from across this range. Phylogenetic analysis of a 1.97 mbp sequence alignment dataset mined from these genomes demonstrates that members of *R.* subgenus *Villosuli* have undergone a process of comigration and cospeciation with their *Pseudotsuga* hosts. Our work presents macroevolutionary evidence that the EM symbiosis between these genera evolved only once between the common ancestor of *R.* subgenus *Villosuli* and the common ancestor of *Pseudotsuga* in Western North America.

Symbiosis in the Pacific Ring of Fire: Evolutionary-Biology of *Rhizopogon* Subgenus *Villosuli* as Mutualists of *Pseudotsuga.*

by
Alija Bajro Mujic

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented March 13, 2015
Commencement June 2015

Doctor of Philosophy dissertation of Alija Bajro Mujic presented on March 13, 2015

APPROVED:

Major Professor, representing Botany and Plant Pathology

Head of the Department of Botany and Plant Pathology

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.

Alija Bajro Mujic, Author

ACKNOWLEDGEMENTS

There are many people who have contributed to the work presented in this dissertation and I humbly thank those of you past, present, and future who have had a hand in bringing this work to completion. Your influence has helped to transform this dissertation's content, as well as my own, into the state that you find it in today. Firstly, I wish to thank my advisor, Dr. Joseph Spatafora, for his patience, guidance, experience, wisdom, and humor. By the welcome extended through him and his family I have been made to feel truly at home and comfortable during my time at Oregon State University. Indeed, you all have been as a second family to me during my time here and I am most grateful. This feeling of family extends also to my fellow members of the department of Botany and Plant Pathology. I had several options in my choice of graduate program, but the choice was made immediately clear when I came to visit OSU for the first time. Whether it was John Fowler's humble and welcoming demeanor or Dianne Simpson's immediate willingness to loan a favor to a complete stranger I am uncertain, but you all made a strong impression upon me from the beginning. You are now, and will remain to be, my family in Oregon. And to my family in California I must also give thanks for your patience and understanding through my time in Oregon. To my mother and father, Judith and Alija, who waited patiently for me to return their phone calls and for my visits. To my sister, Nafija, and brother in-law, James, who kept me laughing and always knew when I needed to talk.

Thank you also to my fellow students. There have been many of us which shared our time and I cannot begin to name all of you. I have grown immensely through our

shared experience of learning, succeeding, failing, mending, and breaking. Thank you to my lab mates in the Spatafora lab. To Alisha Quandt for her indominable spirit and love for the music of Ween. To Ryan Kepler for his relaxed approach to life and for guiding me through my first experiences in the lab. To Cedar Hesse for teaching me that there is always time for fishing, making music, and finishing that last Illumina library prep. To Derek Johnson for reminding me that there is always a better way of doing something if we only take the time to look. To Dabao Sun Lu for sharing my passion for *Rhizopogon* and cooking with me. To Rheannon Arvidson for inspiration to make time at the beginning or end of the day for another mountain bike ride (or both if you please). To Ming Jun Chen for sharing many a mushroom hunt and her first camping experience with me. To Kathryn Bushley for her optimism and infectious laugh. Thank you to Stephanie Vandruff, Joseph Cagle, and Kazuya Tsukagoshi for your inspirational motivation to volunteer countless hours behind the flow hood and for doing the job well.

While working to complete this dissertation I had the great opportunity to work in many forests and laboratories both at home and in the far flung reaches of the globe. I have been shown the greatest kindness and hospitality by the people and forests I have met during these travels. Firstly, this work was begun upon the good graces of my PhD committee who instilled confidence in me by believing in me and encouraging my work. Drs. Kentaro Hosaka and Tsuyoshi Hosoya were both my teachers and my companions in Japan. By their guidance we discovered the first species of *Rhizopogon* subgenus *Villosuli* in Japan. Utashiro-san, Nishibori-san, Otsuka-san, Uno-san, Melanie Cousineau, Ross Hatton, and Jiro-san all ensured that my time time in Japan was both culturally and spiritually fulfilling and I am grateful for their friendship. Dr. Pi-Han Wang, Wan-Rou

Lin, Shih Feng Ming, Sun Hua Jun, and Ming Hsiu Kao made my work in Taiwan a reality and I cannot begin to thank them enough for their patience and contributions to my cultural education and success in Taiwan. Drs. Bo Huang, Ming Jun Chen, and Yu Long were my ever present guides and mentors through Chinese field work. Their guidance and assistance in China allowed me to complete the "final chapter" of my dissertation field work. They also ensured that my heart and my belly were always full of good things during my time in China. Drs. David Gernandt and Roberto Garibay Orijel helped me to navigate the highlands of central Mexico and made me feel as though I will always have a home in Mexico City. Thank you to Drs. Betsy Arnold and Greg Douhan for your hospitality during my field work in the United States. Thank you to Aaron Ung, Justin Schaffer, Kuni Kitajima, Cecelic, and Bob Chapman for your company and assistance in the field. Thank you to Dr. Richard Cronn for waiting patiently for the last *Pseudotsuga* specimens to be processed and for sharing your lab, time, and experience. Lastly, thank you to the all of the funding sources, large and small, which funded my travels and collections. Monetary support was provided for this dissertation by the NSF East Asia Pacific Institutes program, the North American Truffling Society, the Sonoma Mycological Association, the Cascade Mycological Society, the Oregon Mycological Society, the BPP Anita Summers Travel grant and Gottlieb Award, and the OSU Oregon Lottery and Graduate Internationalization grants.

I would like to thank Susan Theis for standing with me through the completion of this work and for loving and encouraging both the artist and the scientist that dwell inside of me.

Peace, Alija Bajro Mujic

CONTRIBUTIONS OF AUTHORS

Chapter 2: Peter Kennedy, Daniel Durall, and Joseph W. Spatafora helped to design, set-up, and harvest container plant experiments. Peter Kennedy developed the initial concept and experimental design of the study and provided growth chamber facilities and raw materials for the study. Financial support for this study was provided through an NSERC grant to Daniel Durall. Peter Kennedy, Daniel Durall, and Joseph W. Spatafora advised and assisted with statistical analyses and edited the manuscript.

Chapter 3: Joseph Spatafora assisted in the design, analysis, and edititing of the manuscript.

Chapter 4: Joseph Spatafora assisted in the design, analysis, and edititing of the manuscript.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

LIST OF FIGURES

Figure Page

LIST OF FIGURES (Continued)

Figure Page

LIST OF TABLES

Chapter 1. Introduction

The functional and ecological importance of ectomycorrhizal (EM) symbioses between plants and fungi are subjects that have long fascinated the scientific community. With the development of modern genetic methodologies we have seen an unprecedented expansion in our knowledge of EM ecology from previously unreachable reserves of information held within the so called "black box" of below ground interactions (Horton and Bruns 2001). With current advances have come many additional questions and there is little currently known concerning the role of EM symbiosis in the evolution of species that participate. Competition and resource heterogeneity are known to affect the spatial organization and community structure of EM fungi (Dickie et al. 2002, Baier et al. 2006, Anderson et al. 2014). Such spatial patterns are likely to affect fungal population structure and change at the microevolutionary scale but the specific ecological mechanisms that shape these patterns are uncertain. Patterns of host specificity ranging from broad to narrow are common in EM fungi (Bruns et al. 2002) and because both partners rely upon the symbiosis for establishment, host specificity likely affects species on micro and macroevolutionary scales.

This dissertation investigates the evolutionary biology and systematics of the EM relationship shared by the genera *Rhizopogon* Fries (Boletales: Rhizopogonaceae) and *Pseudotsuga* Carrière (Pinales: Pinaceae). The research presented here utilizes phylogenomic techniques to develop hypotheses concerning the comigration and cospeciation of trees in the genus *Pseudotsuga* and their *Rhizopogon* EM fungal symbionts. Comparative genomic analyses and ecological studies of the sister species *Rhizopogon vinicolor* and *R. vesiculosus* are performed to investigate ecological and genetic mechanisms underlying cryptic speciation and differential population structure

between these species. Taken together, this work expands the knowledge of EM evolutionary biology and addresses the ecological mechanisms that drive evolutionary change.

The systematics and evolutionary origins of ectomycorrhizal symbiosis

EM interactions begin when growing plant fine roots are ensheathed by fungal mycelium, which forms a mantle surrounding the root tip. The fungal mantle completely separates the root from the soil medium and absorptive root hairs are no longer produced once the mantle is formed. The intercellular space of root epidermal and cortical cells is then colonized by a net of fungal hyphae known as the Hartig net (Figure 1.1). The Hartig net is a site of direct resource exchange where plant roots provide photosynthates to the fungal mycelial network that in turn provides mineral nutrition and water to the plant. Both symbionts realize many benefits from this association. The fungus receives access to an abundant source of carbon and plant hosts have been demonstrated to provide up to 10% of all photosynthetic carbon produced to their EM fungal symbionts. In addition to carbon, fungi receive additional benefits such as physical refuge and cofactors necessary for spore germination. The fungal mycelium emanates from the EM mantle as extramatrical hyphae, often forming large rhizomorphic structures capable of transporting water and minerals over some distance through soil. These extramatrical hyphae and the finely branched mycelium that feed them represent a several-fold increase in the potential surface area for absorption than would be available to the plant host without mycorrhizae (Smith and Read 2008).

The most ancient and geographically widespread mycorrhizal associations known are the arbuscular endomycorrhizal (AM) symbioses involving fungi of the phylum Glomeromycota. AM symbioses are estimated to occur with approximately 90% of all vascular plant lineages and occur in nearly all terrestrial ecosystems where plants occur (Smith and Read 2008). It is hypothesized that the colonization of land by ancient plants was facilitated by the AM symbiosis (Pirozynski and Malloch, 1975) and early Devonian (~400 MYA) plant fossils contain AM structures (Taylor et al. 1995). Mycorrhizal synthesis experiments demonstrate AM type symbiosis in the earliest diverging land plants, thalloid liverworts (Humphreys et al. 2010), as well as spore producing vascular plants (Bidartondo et al. 2011). In contrast to AM symbioses, EM symbioses are a much more recent evolutionary development that have arisen independently across multiple lineages of the fungal phyla Ascomycota, Basidiomycota and Zygomycota (Smith and Read 2008, Tedersoo et al. 2009). At least 162 fungal genera are demonstrated to form EM symbioses with an additional 85 genera supported as EM through phylogenetic or habitat association with known EM taxa (Tedersoo et al. 2009). While AM fungi associate with most lineages of vascular plants, EM fungi are predominantly associated with woody trees and shrubs belonging to the Angiosperm Rosid clade (e.g. Fagales, Rosaceae, Ericaceae, Salicaceae, Dipterocarpaceae) and the Gymnosperm family Pinaceae (Wang and Qiu 2006). The earliest evidence of EM structures from the fossil record are from the middle Eocene period (at least 50 MYA) and are associated with the roots of fossil *Pinus* species (Lepage et al. 1997). Both the fossil record and molecular clock analyses indicate that the earliest EM symbioses most likely involved trees in the family Pinaceae (Hibbett and Matheny 2009), which began to diversify approximately

130 MYA in the early Cretaceous (Wang and Qiu 2006, Miller and Charles 1976). This time period also corresponds with the diversification of flowering plants in the early Cretaceous period (Ligard and Crane 1988, Crane et al. 2000) and it is likely that EM symbiosis rose to prevalence as potential host plant species were undergoing diversification.

EM symbiosis exists along a broad morphological and ecological spectrum and assignment of taxa to specific EM types can be difficult. EM symbioses are broadly distributed across a diverse range of taxonomically distinct fungi. Within Agaricomycetes alone a minimum of six to eight independent origins of the EM habit are predicted (Hibbett and Matheny 2009). Multiple independent origins of the EM symbiosis across a diversity of plant and fungal taxa suggest that convergence upon common symbiotic strategies imparts strong evolutionary advantages upon both partners. Root morphology of floral lineages is strongly correlated with mycorrhizal status and it has been suggested that mycorrhizal plant and fungal lineages have undergone a coevolutionary process that placed selective advantage upon the development of roots and colonization strategies suitable for maintenance of the symbiosis (Brundrett et al. 2002). The plasticity of EM mycorrhizal forms is apparent even within individuals of a single fungal species. Many EM fungal taxa are capable of forming arbutoid (Kennedy et al. 2012, Molina et al. 1997, Molina and Trappe 1994), monotropoid (Cullings et al. 1996, Bidartondo et al. 2002), and orchid (Bidartondo et al. 2004) mycorrhizae with secondary hosts while in concurrent EM association with host trees. Generalist host tree association by EM fungal species is a common phenomenon (Bruns et al. 2002) and host generalist fungi are the prevalent form of EM association in some EM systems (Horton and Bruns 1998, Horton

et al. 1999). Such host generalism is thought to function in promoting ecosystem stability and resilience by facilitating the establishment of seedling recruits through association with shared common mycorrhizal networks with mature trees (Molina and Trappe 1982, Horton et al. 1999, Simard and Durall 2004, Teste et al. 2009, Kennedy et al. 2012)

In contrast to the host generalist strategy some EM fungi function as host specialists that associate with only a single species or genus of host tree. This pattern is common and well documented in the Suilloid clade of EM Agaricomycetes (*Suillus, Rhizopogon, Truncocolumnella*, *Gomphidius,* and *Chroogomphus*), which are almost entirely restricted to host trees in the Pinaceae (Bruns et al. 2002). The biological filter that determines host specificity is usually present in the EM fungal symbiont (Bruns et al. 2002) and host specialist fungi in the Suilloid clade possess adaptations that are consistent with this pattern. Species of both *Suillus* and *Rhizopogon* possess spores, which are stimulated to germinate through exposure to exudates of their preferred host roots (Fries et al. 1987, Miller et al. 1993). The spores of many *Rhizopogon* species exhibit strong dormancy in the absence of suitable host roots and this ability is coupled with extended viability of spores in soil (Bruns et al. 2009). This combination of traits is a likely mechanism for offsetting the costs of host specialization in EM fungi (Bruns et al. 2002). The benefit gained by host specialization is uncertain but it has been hypothesized that the EM structures formed by specialist EM fungi and host plants are more efficient in the net transfer of nutrients (Bruns et al. 2002). Early colonization of previously unforested habitats by *Pinus* species is often observed in association with *Rhizopogon* EM symbionts (Horton et al. 1998, Ashkannejhad and Horton 2006, Peay et al. 2007) and the long term viability and host specificity of *Rhizopogon* spores is a likely

contributing factor to this host benefit (Bruns et al. 2009). When the abundance of a fungal specialist's preferred host tree is high, conspecific seedling recruits are likely to receive an additional advantage by gaining exclusive access to shared mycorrhizal networks with mature host trees.

The genus *Rhizopogon* displays a high degree of host specificity within several monophyletic lineages. The core subjects of this dissertation research are the ecology, evolutionary biology, and systematics of *Rhizopogon* subgenus *Villosuli* and trees of the genus *Pseudotsuga*. The following sections of this introduction will explore the general biology, systematic hypotheses and EM characteristics of the genera *Rhizopogon* and *Pseudotsuga*. Special attention is given to two EM associates of *P. menziesii, R. vinicolor* and *R. vesiculosus,* which are primary study organisms addressed in chapters 2 and 3 of this dissertation. Additional sections of the introduction will explore the mating genetics and phylogeography of EM fungi as these topics are the foundations for the work described in chapters 3 and 4 of this dissertation.

Genus *Rhizopogon:* Taxonomy and ecology

General description:

Rhizopogon is a large genus of EM Basidiomycota that produce hypogeous, truffle-like basidiomata (sporocarps) (Figure 1.2) in association with host trees of Pinaceae. The basidiomata of *Rhizopogon* possess a simple morphology with a loculate interior of spore bearing tissue (gleba) surrounded by a single or bilayered peridium. The interior surface of the gleba is coated with hymenial tissue, which contains the basidia.

Basidia are typically clavate and lack the ability to forcibly discharge spores. Upon reaching maturity, basidia degrade rapidly and passively release basidiospores into the locular chambers. Because of their ephemeral nature, the observation of basidia in dried or very mature specimens of *Rhizopogon* is difficult. A notable feature of *Rhizopogon* basidiomata is the presence of conspicuous rhizomorphic mycelium which often ensheath the outer peridium and emanate from the fruitbody into the surrounding substrate. These rhizomorphs aggregate into a basal tuft in many species of *Rhizopogon* and the regular occurrence of such a basal tuft upon the basidiomata of *R. luteous* (the type species of *Rhizopogon*) is the etymological derivation of the genus name (Gr. *rhizo-* root, *pogon-* beard).

Taxonomy:

Rhizopogon was first described in 1817 by Elias Magnus Fries (Fries 1817) and the foundation of modern species concepts for the genus was established in the North American monograph of Smith and Zeller (1966). A detailed monograph has also been produced for European *Rhizopogon* species (Martín 1996). Though Pinaceae is broadly distributed throughout the northern hemisphere (Farjon 2010) very few data on Asian species of *Rhizopogon* have been incorporated into phylogenetic and taxonomic studies and only a limited account of Asian *Rhizopogon* species has been published for EM associates of *Pinus* (Hosford and Trappe 1988). In total, greater than 150 species of *Rhizopogon* have been described. Historical species classification in *Rhizopogon* was based upon morphological characteristics. Early descriptive work of the genus was generated primarily from dried material (Zeller and Dodge 1918) and this work produced

species concepts that were often too broad because many morphological characters of *Rhizopogon* species are present only in fresh material. The monograph of Smith and Zeller (1966) refined and expanded species concepts for the genus by characterizing the staining reactions of the fresh peridium upon exposure and chemical staining reactions of the glebal, basidiospores, and peridium to chemical reagents. Based on phylogenetic analyses of the internal transcribed spacer (ITS) of the nuclear rDNA, Grubisha et al. (2002) reexamined the infrageneric relationships of Smith and Zeller (1966). Those analyses, as well as morphological and ecological evidence corroborated by their findings, resulted in a new infrageneric classification composed of five subgenera: *Rhizopogon*, *Villosuli*, *Amylopogon*, *Roseoli* and *Versicolores*.

Ecology:

The symbiosis between *Rhizopogon* and Pinaceae occurs globally throughout the natural and anthropogenic range of the family and plays an important ecological role in the establishment and maintenance of forest ecosystems. The hypogeous fruiting habit of *Rhizopogon* species and the inability of *Rhizopogon* basidia to forcible discharge spores preclude wind dispersal of spores. Spore dispersal is instead achieved through excavation and consumption of basidiomata by mammals. *Rhizopogon* species are a major component in the diet of small mammals (Maser and Maser 1988, Luoma et al. 2003, Izzo et al. 2005) and are also consumed by large mammals including deer (Ashkannejhad and Horton 2006) in western North America. Passage through the mammalian gut does not reduce viability of *Rhizopogon* spores and in some cases has been found to enhance their germination (Kotter and Farentinos 1984, Colgan and Claridge 2002).

Rhizopogon species are common members of the EM fungal community in northern coniferous forests and are frequently observed in both newly recruited and old growth stands (Molina et al. 1999). They are especially abundant colonizers of young trees and this trend is often observed following disturbance. Members of *Rhizopogon* subgenus *Villosuli* are the most common species found colonizing roots in young stands in western *P. menziesii* forests (Twieg et al. 2007) and are common upon the roots of *P. menziesii* seedlings growing in the sites of recent timber harvest (Luoma et al. 2006). Baar et al. (1999) found that *Rhizopogon* species comprised a major component of the EM community colonizing the roots of recruiting *Pinus* seedlings following a stand replacing wildfire. These results are especially striking given that *Rhizopogon* species were not detected upon the roots of mature trees in the same stand preceding the fire (Taylor and Bruns 1999). Soil derived from both *Pinus* and *Pseudotsuga* stands, when used as inoculum in container plant bioassays, produces abundant colonization by *Rhizopogon* species even when the EM community observed in mature stands lacked *Rhizopogon* species (Baar et al. 1999, Taylor and Bruns 1999, Kjøller et al. 2003, Murata et al. 2013). These findings reinforce the role of *Rhizopogon* species as members of the resistant spore bank community that colonizes seedlings following disturbance.

Genus *Pseudotsuga:* Taxonomy and systematic hypotheses:

Pseudotsuga is a small yet ecologically diverse genus of coniferous trees of family Pinaceae that are most closely related to Larches (*Larix spp*.) (Lin et al. 2010). *Pseudotsuga* species possess female seed cones with characteristic trilobate (three-pointed) bract scales that protrude well beyond seed scales (Figure 1.2). The needles of

Pseudotsuga species are typically soft and surround the entire axis of branches and the bark is thick and corky with deep vertical furrows in older trees. The taxonomic concepts of *Pseudotsuga* species in Asia and Mexico are disputed and modern taxonomic treatments of the genus recognize between 4 (Farjon 2010) and 8 (Hermann 1982) species. *Pseudotsuga* possesses a relatively small distribution for a genus of the family Pinaceae and are naturally distributed around the northern Pacific Rim with species in western North America, Japan, China and Taiwan (Farjon 2010). *Pseudostuga menziesii* (Douglas-fir) possesses the largest natural distribution of any member of the genus and it is planted extensively beyond its natural range as one of the most important sources of high quality timber in the world (Farjon 2010). Natural populations of *P. menziesii* range throughout western North America. The northern range limit extends from southern Alaska and northern British Columbia with populations reaching as far south as central California on the coast and into the southern Mexican state of Oaxaca on the interior (Debreczy and Racz 1995, Farjon 2010, Gugger et al. 2011). This range is subdivided into distinct populations of coastal *P. menziesii var. menziesii* (CA, OR, WA, BC), northern interior *P. menziesii var. glauca* (BC, MT, WA, OR, ID), southern interior *P. menziesii var. glauca* (UT, CO, AZ, NM) and Mexican *P. menziesii var. glauca* (Aagard et al. 1995, Li and Adams 1989, Viard et al. 2001, Gugger et al. 2010 and 2011). While *P. menziesii* is a major overstory tree of coniferous forests throughout the northern half of its range, the populations extending southward into Mexico are increasingly fragmented and exist as isolated stands subject to restricted gene flow and inbreeding depression (Vargas-Hernández et al. 2004, Mápula-larreta et al. 2007, Velasco-García et al. 2007).

Other species of *Pseudotsuga* are restricted to smaller ranges and do not typically form the monodominant stands characteristic of northern *P. menziesii* forests. The second North American species, *P. macrocarpa* (Vacey) Mayr (Big Cone Douglas Fir), is restricted to the coastal and transverse mountain ranges of southern California (Farjon 2010). *P. macrocarpa* typically grows on seaward facing slopes in a landscape mosaic consisting of mixed *Pinus-Quercus* forest and coastal chaparral. This habitat is subject to frequent and high intensity fire events and *P. macrocarpa* is well adapted to these conditions, possessing a thick corky bark and readily resprouting following fire damage (Lombardo et al. 2009). Asian species concepts for *Pseudotsuga* are not completely resolved but morphological, chemical and molecular evidence clearly support them as distinct from North American species (Strauss et al. 1990, Wei et al. 2010, Yabe et al. 2011). *P. japonica* (Shiras.) Beissn. (Japanese Douglas Fir) is distributed in isolated stands at middle elevation (500—650 m) with a limited range on the Japanese islands of Shikoku, Kyushu, and in the Kii peninsula regions of Honshu (Farjon 2010). *P. sinensis* Dode (Chinese Douglas Fir) is perhaps the most contentious of the species concepts for *Pseudotsuga.* Farjon (2010) recognizes only a single species with a varietal structure spanning its range in China, northern Vietnam and Taiwan. The varietal names of Farjon (2010) are similar to previous species epithets. The Taiwanese distribution of *Pseudotsuga* has been referred to as *P. wilsoniana* Hayata (Li and Fu 1997), but has been synonymized by Fajron (2010) with *P. sinensis* var. *sinensis* that also possesses a central Chinese distribution. Eastern Chinese populations have been referred to as *P. forrestii* Craib., the southern Chinese and Vietnamese populations as *P. brevifolia* Cheng and Fu and parts of the central Chinese population as *P. gausenii* Flous. All East Asian species

of *Pseudotsuga* possess fragmented distributions consisting of small populations at middle to high elevation. Central and southern Chinese populations typically exist at middle elevation (600–1200 m) while Taiwanese and western Chinese populations exist at much higher elevation (1000–2700 m, and up to 3300m, respectively) (Farjon 2010).

Strauss et al. (1990) produced the first molecular phylogeny of *Pseudotsuga* species from Japan, Taiwan, mainland China, and North America using restriction fragment analysis of nuclear, mitochondrial and chloroplast DNA. This study revealed patterns of monophyly within Asian and North American species groups. The closest relationship between the groups is shared by Japanese and North American species. Their findings support a North American origin for *Pseudotsuga* with subsequent migration around the Pacific Rim. Phylogenetic analysis of the ITS rDNA (Gernandt and Liston 1999), cpDNA, and the nuclear LEAFY gene (Wei et al. 2010) from similar *Pseudotsuga* populations supported the same sister relationship between Asian and North American clades but could not conclusively support a North American origin for the genus. However, a North American origin of *Pseudotsuga* is consistent with the greater age of the North American fossil record for the genus, which dates to 50 – 58 MYA as compared to 23 MYA for Asia (Hermann 1985, Yabe 2011).

The *Rhizopogon*-*Pseudotsuga* symbiosis:

The ITS phylogeny of Grubisha et al. (2002) found that EM associates of major conifer genera (e.g., *Pinus, Abies*) were distributed across multiple subgenera of *Rhizopogon*. As an exception to this trend, a host-specific relationship was resolved between the species *Pseudotsuga menziesii* (Mirb.) Franco (Pinales: Pinaceae) and *R.*

subg. *Villosuli.* The specificity and obligatory nature of the *R.* subgenus *Villosuli-Pseudotsuga* EM symbiosis suggests a single evolutionary origin of this relationship within genus *Rhizopogon* (Grubisha et al. 2002). Only one *Rhizopogon* species outside of *R.* subgenus *Villosuli*, *R. salebrosus* (=*R. subcaerulescens*) in *R.* subgenus *Amylopogon*, has been demonstrated to form an EM relationship with *P. menziesii* without the presence of an alternate host (Massicotte et al. 1994). However, this species forms more abundant mycorrhizae with *Pinus* species and is also capable of forming EM relationships with a range of Pinaceae hosts (Massicotte et al. 1994), arbutoid mycorrhizae with Ericaceae hosts, and Monotropoid mycorrhizae with mycoheterotrophic *Pterospora andromedea* (Ericaceae) hosts (Molina et al. 1997). In contrast, members of *R.* subgenus *Villosuli* have consistently been demonstrated to only form mycorrhizae with *P. menziesii* (Massicotte et al. 1994, Molina et al. 1997). However, the relationship between *R.* subgenus *Villosuli* and *Pseudotsuga* has previously been observed only from host trees in the coastal populations of *P. menziesii* var. *menziesii* (CA, OR, WA, BC) and northern interior populations of *P. menziesii* var. *glauca* (ID, MT, OR, WA, BC). The specificity observed in northern and coastal *P. menziesii* has yet to be observed for trees in the southern extent of its range and the *Rhizopogon* associates of *P. macrocarpa* and East Asian *Pseudotsuga* are unknown.

Within the Pacific Northwest (PNW), *P. menziesii* is a dominant overstory tree species estimated to be capable of forming EM associations with up to 2500 species of fungi (Molina et al. 1994). The broad EM compatibility of *P. menziesii* helps to maintain fungal biodiversity and provides a source of mycorrhizal inoculum for establishing trees through common mycorrhizal networks (Twieg et al. 2007, Simard et al. 2009a). These

preestablished mycorrhizal networks have significant effects upon tree survival, establishment, and competitive ability (Simard and Durall 2004, Selosse et al. 2006, Whitfield 2007, Simard et al. 2009a) and because *Rhizopogon* spp. are prevalent members of *P. menziesii* EM communities, they are likely to represent a large fraction of the mycorrhizal benefits realized by their hosts. The regenerative capacity of *P. menziesii* forests has decreased over the last decade in response to a changing climate (Hamman and Wang 2006, Klenner et al. 2008) and conservation of isolated Mexican populations is of particular concern (Norma Oficial Mexicana NOM-059-SEMARNAT-2010 2010). *Pseudotsuga japonica* is listed as a "vulnerable" species according to the Red List of Threatened Plants of Japan (http://www.biodic.go.jp/english/rdb/rdb_e.html) and Japanese, Taiwanese and Chinese *Pseudotsuga* species are listed as "vulnerable" by the IUCN Red List of Threatened Species (http://www.iucnredlist.org/**).** Understanding of the *Pseudotsuga-Rhizopogon* symbiosis will provide information useful in *Pseudotsuga* forest conservation and enhance the management of harvested stands under the pressure of global climate change (Simard 2009b).

Rhizopogon vinicolor and *Rhizopogon vesiculosus*: Ecology and population genetics.

Rhizopogon vinicolor A.H. Smith and *R. vesiculosus* A.H. Smith are cryptic sister species belonging to *Rhizopogon* subgenus *Villosuli* (Kretzer et al. 2003) and they are obligate EM symbionts of *Pseudotsuga menziesii* (Molina et al. 1994). The symbiosis between these fungi and *P. menziesii* has previously been studied in British Columbia (Beiler et al. 2010) and Oregon (Dunham et al. 2013, Kretzer et al. 2005, Kretzer et al.

2003). The results of these studies have shown that these species are sympatric in distribution throughout the PNW and that they occur at similar frequencies on a landscape scale. The sporocarps and EM root tip morphologies of these two species are very similar, but they display markedly different life histories. *R. vesiculosus*, on average, creates much larger genets than *R. vinicolor* (Beiler et al. 2010, Dunham et al. 2013, Kretzer et al. 2005, Kretzer et al. 2003) and is distributed deeper into the soil profile when the two species co-occur in horizontal space (Beiler et al. 2012). Fixation index (F_{ST}) measures have indicated that *R. vesiculosus* shows moderate population differentiation at the watershed scale (Kretzer et al. 2005) and this trend is reinforced on a local stand-level scale where spatial autocorrelation analyses reveal a pattern of inbreeding in *R. vesiculosus* within a range of 120 meters (Dunham et al. 2013). In contrast, *R. vinicolor* was found to display a pattern of random mating on a stand-level scale (Dunham et al. 2013) and F_{ST} values did not differ from zero on a watershed-level scale (Kretzer et al. 2005). In the most recent study (Dunham et al. 2013), it was hypothesized that several factors may be driving the trends discussed above, but uneven rates of fruitbody production per genotype at the landscape scale is the most plausible hypothesis. However, the work conducted to date cannot support or refute the hypotheses that *R. vesiculosus* and *R. vinicolor* possess different mating systems or that resource competition is driving patterns of genotype distribution.

Introduction to mating in Agaricomycetes.

In all Agaricomycetes studied to date a common set of mating type homolog genes has been identified that function in the recognition and anastomosis of

monokaryotic hyphae and spores (Heitman et al. 2013). These genes are distributed across two genomic regions, which are commonly referred to as the A-locus and B-locus (Figure 1.3). The A-locus comprises one or more pairs of homeodomain containing genes each of which encode monomers of a heterodimeric transcription factor protein. Functional heterodimers of A-locus proteins function in regulation of nuclear pairing, clamp cell formation and several other processes involved with the maintenance of dikaryotic mycelium (Casselton and Olesnicky 1998). The B-locus contains several cassettes of genes that encode lipopeptide pheromone precursors and pheromone receptor proteins. Pheromone receptors are G-coupled 7-transmembrane proteins embedded in the cell membrane and upon recognition of compatible lipopeptide pheromones initiate the process of septal dissolution, nuclear migration, and fusion of clamp cells (Casselton and Olesnicky 1998). While both A and B locus genes are critical for successful mating in Agaricomycetes they continue to function in maintenance of mycelium throughout the remainder of the dikaryotic phase of the lifecycle (Casselton and Olesnicky 1998).

Much of the knowledge concerning mating recognition genetics in Agaricomycetes is based upon classical genetics studies of the model saprobic fungi *Coprinopsis cinerea* and *Schizophyllum commune* (Cassleton and Kues 2007). This research extended prior understanding of fungal mating gained through the study of the genetic pathways that govern sexual compatibility in the model yeasts *Saccharomyces cerevisiae* (Ascomycota: Saccharomycetales) (Caldwell et al. 1995) and *Cryptococcus neoformans* (Basidiomycota: Tremellales) (Alspaugh et al. 1997), as well as the plant pathogen *Ustilago maydis* (Basidiomycota: Ustilaginales) (Bölker et al. 1992). More

recent studies utilizing molecular genetics have greatly expanded our knowledge of the specific genes that regulate mating in model Agaricomycetes (O'Shea et al. 1998. Wendland et al. 1995) and have identified mating type homologs in several additional mushroom forming Agaricomycetes including *Pleurotus djamor* (James et al. 2004), *Coprinellus disseminatus* (James et al. 2006), and *Pholiota nameko* (Aimi et al. 2005). Current research has leveraged next generation sequencing technologies in order to generate complete fungal genomes and a detailed genetic map is now available for the study of mating type associated genes in both *C. cinerea* (Stajich et al. 2010) and *S. commune* (Ohm et al. 2010), as well as a myriad other Agaricomycetes (Grigoriev et al. 2014). A recent review of the genomes available for the order Polyporales characterized many of the genes associated with mating type recognition in this order of largely wood decay fungi (James et al. 2013). Despite the availability of many genomes and their demonstrated utility in unraveling the genetic architecture of fungal mating systems, very little knowledge of mating incompatibility systems in EM Agaricomycetes has been developed.

Mating-type characterization in Agaricomycetes has traditionally been performed by observation of the compatibility ratio between in-vitro crosses of sibling monokaryotic haploid strains. This methodology is less tractable with EM Agaricomycetes because many species possess spores that will not readily germinate in culture, making the isolation of monokaryons unfeasible. Genomic approaches to mating-type characterization in EM fungi provide an additional insight into this problem by investigating EM fungal mating-type genetics in-silico. To date the only EM Agaricomycete to have its mating type loci characterized with these new techniques is the

mushroom *Laccaria bicolor* (Niculita-Hirzel et al. 2008). This fungus possesses a known mating-type system (heterothallic tetrapolar, 25% self compatible) (Fries et al. 1984). Genomic characterization of *L. bicolor* mating-type loci revealed many similarities with the mating-type loci of better characterized tetrapolar model Agaricomycetes fungi and demonstrate the utility of genomic methods for the study of fungal mating systems (Niculita-Hirzel et al. 2008). James et al. (2013) have utilized similar genomic scale approaches to characterize the unknown mating-type system of the saprobic Polyporales fungus *Wolfiporia cocos.* This fungus has also proven resistant to culture based methods of mating-type determination and the results of James et al. (2013) suggest that their methods might also be applicable to EM Agaricomycetes.

Global phylogeography of EM fungi

Many EM fungi possess wind born spores that are expected to achieve long distance dispersal. From this knowledge one can reach the assumption that EM fungi have a panmictic population structure that is only limited by availability of appropriate substrate or host. This view of fungal dispersal patterns was prevalent well into the 20th century but with the advent of molecular genetics we now know that fungi are subject to many of the same geographic dispersal barriers that drive vicariance events in other taxa (Peay et al. 2010). A prime example of this paradigm shift can be found in the case of *Amanita muscaria* L., a wind-dispersed EM host generalist and one of the most broadly distributed EM Agaricomycetes in the world. It is the most recognized mushroom throughout global folklore and popular culture and its “red cap with white spots” are a familiar sight in cultural iconography as well as forests across the northern hemisphere.

Because of its broad distribution and cultural familiarity, this mushroom has been held as an example of an ecologically plastic EM fungus with global long-range dispersal. However, this view has been challenged by molecular phylogenetics (Geml et al. 2008), which demonstrate both intercontinental and intracontinental genetic disjunctions. Geml et al. (2008) demonstrate that the phylogeographic structure of *A. muscaria* is correlated with habitat, host and geographic region suggesting a high degree of regional endemism and multiple cryptic speciation events. Even finer scale population structure has been demonstrated for the wind-dispersed EM Agaricomycete *Cantharellus formosus* Corner. Spatial autocorrelation analysis of genotypic distribution in this fungus found a non-random grouping of genotypes with genetic isolation by distance at a range of 400 meters (Dunham et al. 2006). The landscape genetics of a species is an increasingly important consideration in conservation and management strategies (Segelbacher et al. 2010). Molecular studies underscore landscape genetics as an important concern to the conservation of EM fungi and the forest habitats they support.

Long distance dispersal of species must occur for observations of closely related species to be made in geographically distant locations. For wind-dispersed species these events can easily be attributed to rare long distance establishment via long-range travel of spores in air. For hypogeous fungi, which are reliant upon mycophagy for dispersal, a disjunct distribution of closely related species is less easily explained. Despite perceived barriers to long distance dispersal in hypogeous fungi these events do occur even over very large geographic scales. Hosaka et al. (2008) used molecular data from northern and southern hemisphere members of the hypogeous EM order Hysterangiales to infer the historical migration and evolutionary patterns of the group. Their work resolved an

Australian or Eastern Gondwanan origin of the group but neither the northern nor southern hemisphere taxa of the order formed a monophyletic group. This finding suggests that long distance dispersal, even over trans-oceanic boundaries, must be possible even for EM fungi with hypogeous fruiting habits.

Outline of the dissertation chapters

The research presented in this dissertation is centered upon *Rhizopogon* subgenus *Villosuli* and the EM relationship it shares with members of the genus *Pseudotsuga*. The topics covered in this work progress from microevolutionary community and population level interactions to macroevolutionary systematic hypotheses for *Rhizopogon*. This work provides insight into ecological mechanisms which structure communities and populations of *Rhizopogon.* Further, it tests systematic and geographic hypotheses for the origins and evolution of the *Rhizopogon-Pseudotsuga* EM symbiosis. Chapter 2 presents an experimental examination of ecological mechanisms responsible for the spatial partitioning of *P. menziesii* EM root tips between *R. vinicolor* and *R. vesiculosus*. This work utilizes a series of container plant experiments to test the hypotheses that either competitive interactions or niche differentiation drive spatial partitioning of *R. vinicolor* and *R. vesiculosus* root tip colonization. The results of these experiments indicate that it is resource heterogeneity which drives spatial partitioning between these *Rhizopogon* species, rather than the properties of particular resource patches. Where *R. vinicolor* succeeds in competing with *R. vesiculosus* the level of its success is directly correlated with the number of *R. vinicolor* multilocus genotypes present. These findings suggest for

the first time that competitive success in an EM fungus is directly proportional to its rate of reproductive success.

In collaboration with the United States Department of Energy Joint Genome institute we have sequenced the draft genome of a single isolate for both *R. vinicolor* and *R. vesiculosus.* Chapter 3 takes a comparative genomics approach to characterize the mating-type loci of *R. vinicolor* and *R. vesiculosus*. Both fungi possess a structure at their mating-type loci that is comparable to that of other members of the Boletales. Comparison of fungal genomes with known mating type reveals that the genetic diversity and organization of these loci is correlated with mating-type in the Boletales. These findings suggest that *R. vinicolor* may possess a mating-type system with less potential for inbreeding than that of *R. vesiculosus* and that the differential population structure of these species may be driven by differences in mating-type system.

Chapter 4 addresses the systematics and phylogeography of *R.* subgenus *Villosuli* across the range of all *Pseudotsuga* species. This research tests the hypothesis that the *Rhizopogon-Pseudotsuga* symbiosis evolved only once between the common ancestor of *R.* subgenus *Villosuli* and the common ancestor of all *Pseudotsuga* species. The results of 5 years of international fieldwork are used to characterize the first species of *Rhizopogon* known to associate with *Pseudotsuga* species in Asia, Mexico, and southern California. The phylogeography of *R.* subgenus *Villosuli* is inferred using a genome scale dataset that includes *Rhizopogon* specimens from the range of all *Pseudotsuga* species. This work supports a single evolutionary origin in North America for the *Rhizopogon-Pseudotsuga* EM symbiosis.

REFERENCES:

Aagaard, JE, Vollmer SS, Sorensen FC, and Strauss SH. 1995. Mitochondrial DNA Products among RAPD Profiles Are Frequent and Strongly Differentiated between Races of Douglas-Fir. *Molecular Ecology* 4 (4): 441–46. doi:10.1111/j.1365-294X.1995.tb00237.x.

Aimi, Tadanori, Rie Yoshida, Mariko Ishikawa, Dapeng Bao, and Yutaka Kitamoto. 2005. Identification and Linkage Mapping of the Genes for the Putative Homeodomain Protein (hox1) and the Putative Pheromone Receptor Protein Homologue (rcb1) in a Bipolar Basidiomycete, *Pholiota nameko*. *Current Genetics* 48 (3): 184–94. doi:10.1007/s00294-005-0012-7.

Anderson, Ian C., David R. Genney, and Ian J. Alexander. 2014. Fine-Scale Diversity and Distribution of Ectomycorrhizal Fungal Mycelium in a Scots Pine Forest. *New Phytologist* 201 (4): 1423–30. doi:10.1111/nph.12637.

Ashkannejhad, Sara, and Thomas R. Horton. 2006. Ectomycorrhizal Ecology under Primary Succession on Coastal Sand Dunes: Interactions Involving *Pinus contorta*, Suilloid Fungi and Deer. *New Phytologist* 169 (2): 345–54. doi:10.1111/j.1469-8137.2005.01593.x.

Alspaugh, J. Andrew, John R. Perfect, and Joseph Heitman. 1997. *Cryptococcus neoformans* Mating and Virulence Are Regulated by the G-Protein A Subunit GPA1 and cAMP. *Genes & Development* 11 (23): 3206–17. doi:10.1101/gad.11.23.3206.

Baar, J., T. R. Horton, A. M. Kretzer, and T. D. Bruns. 1999. Mycorrhizal Colonization of *Pinus muricata* from Resistant Propagules after a Stand-Replacing Wildfire. *New Phytologist* 143 (2): 409–18. doi:10.1046/j.1469-8137.1999.00452.x.

Baier, Roland, Jan Ingenhaag, Helmut Blaschke, Axel Göttlein, and Reinhard Agerer. 2006. Vertical Distribution of an Ectomycorrhizal Community in Upper Soil Horizons of a Young Norway Spruce (*Picea abies* [L.] Karst.) Stand of the Bavarian Limestone Alps. *Mycorrhiza* 16 (3): 197–206. doi:10.1007/s00572-006-0035-z.

Beiler KJ, Durall DM, Simard SW, Maxwell SA and Kretzer AM. 2010. Architecture of the wood-wide web: *Rhizopogon spp*. genets link multiple Douglas-fir cohorts *New Phytologist* 185(2) 543-553.

Beiler, Kevin J., Suzanne W. Simard, Valerie LeMay, and Daniel M. Durall. 2012. Vertical Partitioning between Sister Species of *Rhizopogon* Fungi on Mesic and Xeric Sites in an Interior Douglas-Fir Forest. *Molecular Ecology* 21 (24): 6163–74. doi:10.1111/mec.12076.

Bidartondo, M. I., and T. D. Bruns. 2002. Fine-Level Mycorrhizal Specificity in the Monotropoideae (Ericaceae): Specificity for Fungal Species Groups. *Molecular Ecology* 11 (3): 557–69. doi:10.1046/j.0962-1083.2001.01443.x.

Bidartondo, Martin I., Bastian Burghardt, Gerhard Gebauer, Thomas D. Bruns, and David J. Read. 2004. Changing Partners in the Dark: Isotopic and Molecular Evidence of Ectomycorrhizal Liaisons between Forest Orchids and Trees. *Proceedings of the Royal Society of London B: Biological Sciences* 271 (1550): 1799–1806. doi:10.1098/rspb.2004.2807.

Bidartondo, Martin I., David J. Read, James M. Trappe, Vincent Merckx, Roberto Ligrone, and Jeffrey G. Duckett. 2011. The Dawn of Symbiosis between Plants and Fungi. *Biology Letters* 7 (4): 574–77. doi:10.1098/rsbl.2010.1203.

Bölker, Michael, Martin Urban, and Regine Kahmann. 1992. The a Mating Type Locus of *U. maydis* Specifies Cell Signaling Components. *Cell* 68 (3): 441–50. doi:10.1016/0092-8674(92)90182-C.

Bruns, Thomas D., Martin I. Bidartondo, and D. Lee Taylor. 2002. Host Specificity in Ectomycorrhizal Communities: What Do the Exceptions Tell Us? *Integrative and Comparative Biology* 42 (2): 352–59. doi:10.1093/icb/42.2.352.

Bruns, Thomas D., Kabir G. Peay, Primrose J. Boynton, Lisa C. Grubisha, Nicole A. Hynson, Nhu H. Nguyen, and Nicholas P. Rosenstock. 2009. Inoculum Potential of *Rhizopogon* Spores Increases with Time over the First 4 Yr of a 99-Yr Spore Burial Experiment. *New Phytologist* 181 (2): 463–70. doi:10.1111/j.1469-8137.2008.02652.x.

Brundrett, Mark C. 2002. Coevolution of Roots and Mycorrhizas of Land Plants. *New Phytologist* 154 (2): 275–304. doi:10.1046/j.1469-8137.2002.00397.x.

Caldwell, G. A., F. Naider, and J. M. Becker. 1995. Fungal Lipopeptide Mating Pheromones: A Model System for the Study of Protein Prenylation. *Microbiological Reviews* 59 (3): 406–22.

Casselton LA, Kües U. 2007. Sex in Fungi: Molecular Determination and Evolutionary Implications, eds Heitman J, Kronstad JW, Taylor JW, Casselton LA (American Association for Microbiology Press, Washington), pp 283–300.

Casselton, Lorna A., and Natalie S. Olesnicky. 1998. Molecular Genetics of Mating Recognition in Basidiomycete Fungi. *Microbiology and Molecular Biology Reviews* 62 (1): 55–70.

Colgan III, Wes, and Andrew W. Claridge. 2002. Mycorrhizal Effectiveness of *Rhizopogon* Spores Recovered from Faecal Pellets of Small Forest-Dwelling Mammals. *Mycological Research* 106 (3): 314–20. doi:10.1017/S0953756202005634.

Cullings KW, Szaro TM , Bruns TD. 1996. Evolution of extreme specialization within a lineage of ectomycorrhizal epiparasites. *Nature* 379:63-66

Crane PR, Friis EM, Pedersen KR. 2000. The origin and early diversification of Angiosperms. In: *Shaking the Tree: Readings from Nature in the History of Life*. Ed. Gee, H. University of Chicago Press.

Debreczy Z. and Racz I. 1995. New species and varieties of conifers from Mexico. *Phytologia* 78:217–243.

Dickie, Ian A., Bing Xu, and Roger T. Koide. 2002. Vertical Niche Differentiation of Ectomycorrhizal Hyphae in Soil as Shown by T-RFLP Analysis. *New Phytologist* 156 (3): 527–35. doi:10.1046/j.1469-8137.2002.00535.x.

Dunham, Susie M., Thomas E. O'Dell, and Randy Molina. 2006. Spatial Analysis of within-Population Microsatellite Variability Reveals Restricted Gene Flow in the Pacific Golden Chanterelle (*Cantharellus formosus*). *Mycologia* 98 (2): 250–59. doi:10.3852/mycologia.98.2.250.

Dunham, Susie M., Alija Bajro Mujic, Joseph W. Spatafora, and Annette M. Kretzer. 2013. Within-Population Genetic Structure Differs between Two Sympatric

Sister-Species of Ectomycorrhizal Fungi, Rhizopogon Vinicolor and R. Vesiculosus. *Mycologia* 105 (4): 814–26. doi:10.3852/12-265.
Farjon A. 2010. *A handbook of the world's conifers*. Brill Leiden-Boston.
Fries, Elias Magnus. 1817. *Symbolae Gasteromycorum*. Lundae: Ex officina Berlingiana.
Fries, Nils, and Gregory M. Mueller. 1984. Incompatibility Systems, Cultural Features and Species Circumscriptions in the Ectomycorrhizal Genus *Laccaria* (Agaricales). *Mycologia* 76 (4): 633–42. doi:10.2307/3793220.
Fries, Nils, Klaus Serck-Hanssen, Lena Häll Dimberg, and Olof Theander. 1987. Abietic Acid, and Activator of Basidiospore Germination in Ectomycorrhizal Species of the Genus *Suillus* (Boletaceae). *Experimental Mycology* 11 (4): 360–63. doi:10.1016/0147-5975(87)90024-7.
Geml, József, Rodham E. Tulloss, Gary A. Laursen, Nina A. Sazanova, and D. L. Taylor. 2008. Evidence for Strong Inter- and Intracontinental Phylogeographic Structure in *Amanita muscaria*, a Wind-Dispersed Ectomycorrhizal Basidiomycete. *Molecular Phylogenetics and Evolution* 48 (2): 694–701. doi:10.1016/j.ympev.2008.04.029.
Gernandt, David S., and Aaron Liston. 1999. Internal Transcribed Spacer Region Evolution in *Larix* and *Pseudotsuga* (Pinaceae). *American Journal of Botany* 86 (5): 711–23.
Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F, Smirnova T, Nordberg H, Dubchak I, Shabalov I. 2014. MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Research* 42(1):D699-704.
Grubisha, Lisa C., James M. Trappe, Randy Molina, and Joseph W. Spatafora. 2002. Biology of the Ectomycorrhizal Genus Rhizopogon. VI. Re-Examination of Infrageneric Relationships Inferred from Phylogenetic Analyses of ITS Sequences. *Mycologia* 94 (4): 607–19.
Gugger, Paul F., Shinya Sugita, and Jeannine Cavender-Bares. 2010. Phylogeography of Douglas-Fir Based on Mitochondrial and Chloroplast DNA Sequences: Testing Hypotheses from the Fossil Record. *Molecular Ecology* 19 (9): 1877–97. doi:10.1111/j.1365-294X.2010.04622.x.
Gugger, Paul F., Antonio González-Rodríguez, Hernando Rodríguez-Correa, Shinya Sugita, and Jeannine Cavender-Bares. 2011. Southward Pleistocene Migration of Douglas-Fir into Mexico: Phylogeography, Ecological Niche Modeling, and Conservation of 'rear Edge' Populations. *New Phytologist* 189 (4): 1185–99. doi:10.1111/j.1469-8137.2010.03559.x.
Hamman A, Wang T. 2006. Potential effects of climate change on ecosystem and tree species distribution in British Columbia. *Ecology* **87:** 2773–2786.
Heitman, Joseph, Sheng Sun, and Timothy Y. James. 2013. Evolution of Fungal Sexual Reproduction. *Mycologia* 105 (1): 1–27. doi:10.3852/12-253.
Hermann RK 1982. The genus *Pseudotsuga*: historical records and nomenclature. Forest Research Laboratory, Special Publication 2a. Oregon State University, Corvallis, OR.

Hermann RK. 1985. The Genus *Pseudotsuga*: Ancestral History and Past Distribution, Forest Research Laboratory, Special Publication 2b. Oregon State University, Corvallis, OR.

Hibbett, David S., and P. Brandon Matheny. 2009. The Relative Ages of Ectomycorrhizal Mushrooms and Their Plant Hosts Estimated Using Bayesian Relaxed Molecular Clock Analyses. *BMC Biology* 7 (1): 13. doi:10.1186/1741-7007-7-13.

Hillebrand, Helmut. 2004. On the Generality of the Latitudinal Diversity Gradient. *The American Naturalist* 163 (2): 192–211. doi:10.1086/381004.

Horton, Thomas R., and Thomas D. Bruns. 1998. Multiple-Host Fungi Are the Most Frequent and Abundant Ectomycorrhizal Types in a Mixed Stand of Douglas Fir (*Pseudotsuga menziesii*) and Bishop Pine (*Pinus muricata*). *New Phytologist* 139 (2): 331–39. doi:10.1046/j.1469-8137.1998.00185.x.

Horton TR, Cazares E, Bruns TD. 1998. Ectomycorrhizal, vesicular-arbuscular and dark septate fungal colonization of bishop pine (*Pinus muricata*) seedlings in the first 5 months of wildfire. *Mycorrhizae* 8: 11–18.

Horton, Thomas R, Thomas D Bruns, and V Thomas Parker. 1999. Ectomycorrhizal Fungi Associated with *Arctostaphylos* Contribute to *Pseudotsuga menziesii* Establishment. *Canadian Journal of Botany* 77 (1): 93–102. doi:10.1139/b98-208.

Horton, T. R., and T. D. Bruns. 2001. The Molecular Revolution in Ectomycorrhizal Ecology: Peeking into the Black-Box. *Molecular Ecology* 10 (8): 1855–71.

Hosaka, Kentaro, Michael A. Castellano, and Joseph W. Spatafora. 2008. Biogeography of Hysterangiales (Phallomycetidae, Basidiomycota). *Mycological Research, Phylogeography and Biogeography of Fungi,* 112 (4): 448–62. doi:10.1016/j.mycres.2007.06.004.

Hosford, D. R., and J. M. Trappe. 1988. A Preliminary Survey of Japanese Species of *Rhizopogon*. *Transactions of the Mycological Society of Japan*. 29:63-72

Humphreys, Claire P., Peter J. Franks, Mark Rees, Martin I. Bidartondo, Jonathan R. Leake, and David J. Beerling. 2010. Mutualistic Mycorrhiza-like Symbiosis in the Most Ancient Group of Land Plants. *Nature Communications* 1 (November): 103. doi:10.1038/ncomms1105.

Izzo, Antonio D., Marc Meyer, James M. Trappe, Malcolm North, and Thomas D. Bruns. 2005. Hypogeous Ectomycorrhizal Fungal Species on Roots and in Small Mammal Diet in a Mixed-Conifer Forest. *Forest Science* 51(3)243-254

James, Timothy Y., Shian-Ren Liou, and Rytas Vilgalys. 2004. The Genetic Structure and Diversity of the A and B Mating-Type Genes from the Tropical Oyster Mushroom, *Pleurotus djamor*. *Fungal Genetics and Biology: FG & B 41* (8): 813–25. doi:10.1016/j.fgb.2004.04.005.

James, Timothy Y., Prayook Srivilai, Ursula Kües, and Rytas Vilgalys. 2006. Evolution of the Bipolar Mating System of the Mushroom *Coprinellus disseminatus* From Its Tetrapolar Ancestors Involves Loss of Mating-Type-Specific Pheromone Receptor Function. *Genetics* 172 (3): 1877–91.doi:10.1534/genetics.105.051128.

James, Timothy Y., Sheng Sun, Wenjun Li, Joseph Heitman, Hsiao-Che Kuo, Yong-Hwan Lee, Frederick O. Asiegbu, and Åke Olson. 2013. Polyporales Genomes

Reveal the Genetic Architecture Underlying Tetrapolar and Bipolar Mating Systems. *Mycologia* 105 (6): 1374–90. doi:10.3852/13-162.

Kennedy, Peter G., Dylan P. Smith, Tom R. Horton, and Randy J. Molina. 2012. *Arbutus menziesii* (Ericaceae) Facilitates Regeneration Dynamics in Mixed Evergreen Forests by Promoting Mycorrhizal Fungal Diversity and Host Connectivity. *American Journal of Botany* 99 (10): 1691–1701. doi:10.3732/ajb.1200277.

Klenner W, Walton R, Arsenault A, Kramseter L. 2008. Dry forests in the southern interior of British Columbia: historic disturbances and implications for restoration and management. *Forest Ecology & Management* **206:** 1711–1722.

Kjøller, Rasmus, and Thomas D. Bruns. 2003. *Rhizopogon* Spore Bank Communities within and among California Pine Forests. *Mycologia* 95 (4): 603–13. doi:10.2307/3761936.

Kotter, Martha M., and Robert C. Farentinos. 1984. Formation of Ponderosa Pine Ectomycorrhizae after Inoculation with Feces of Tassel-Eared Squirrels. *Mycologia* 76 (4): 758–60. doi:10.2307/3793237.

Kretzer AM, Dunham S, Molina R, Spatafora JW. 2003. Microsatellite markers reveal the below ground distribution of genets in two species of *Rhizopogon* forming tuberculate ectomycorrhizas on Douglas fir. *New Phytologist* 161(1) 313-320

Kretzer AM, Dunham S, Molina R, Spatafora JW. 2005. Patterns of vegetative growth and gene flow in *Rhizopogon vinicolor* and *R. vesiculosus* (Boletales, Basidiomycota) *Molecular Ecology* 14(8) 2259-2268

Li, Peng, and W. T. Adams. 1989. Range-Wide Patterns of Allozyme Variation in Douglas-Fir (*Pseudotsuga menziesii*). *Canadian Journal of Forest Research* 19 (2): 149–61. doi:10.1139/x89-022.

Li, Nan, and Li-kuo Fu. 1997. Notes on Gymnosperms I. Taxonomic Treatments of Some Chinese Conifers. *Novon* 7 (3): 261–64. doi:10.2307/3391941.

Lidgard, Scott, and Peter R. Crane. 1988. Quantitative Analyses of the Early Angiosperm Radiation. *Nature* 331 (6154): 344–46. doi:10.1038/331344a0.

Lin, Ching-Ping, Jen-Pan Huang, Chung-Shien Wu, Chih-Yao Hsu, and Shu-Miaw Chaw. 2010. Comparative Chloroplast Genomics Reveals the Evolution of Pinaceae Genera and Subfamilies. *Genome Biology and Evolution* 2 (January): 504–17. doi:10.1093/gbe/evq036.

Lepage, B., R. Currah, R. Stockey, and G. Rothwell. 1997. Fossil Ectomycorrhizae from the Middle Eocene. *American Journal of Botany* 84 (3): 410–410.

Lombardo KJ, Swetnam TW, Baisan CH, Borchert MI. 2009. Using bigcone Douglas-fir fire scars and tree rings to reconstruct interior chaparral fire history. *Fire Ecology* 5(3): 35–56

Luoma, Daniel L, Christopher A Stockdale, Randy Molina, and Joyce L Eberhart. 2006. The Spatial Influence of *Pseudotsuga menziesii* Retention Trees on Ectomycorrhiza Diversity. *Canadian Journal of Forest Research* 36 (10): 2561–73. doi:10.1139/x06-143.

Mápula-larreta, M., J. López-Upton, J. J. Vargas-Hernández, and A. Hernández-Livera. 2007. Reproductive Indicators in Natural Populations of Douglas-Fir in

Mexico. *Biodiversity and Conservation* 16 (3): 727–42. doi:10.1007/s10531-005-5821-y.

Martín, MP. 1996. *The Genus Rhizopogon in Europe*. Barcelona, Spain: BCG. pp. 173

Maser C, Maser Z. 1988. Interactions among squirrels, mycorrhizal fungi, and coniferous forests in Oregon. *Western North American Naturalist* **48** (3): 358–369

Massicotte, Hugues B., Randy Molina, Daniel L. Luoma, and Jane E. Smith. 1994. Biology of the Ectomycorrhizal Genus, *Rhizopogon* II: I. Patterns of host-fungus specificity following spore inoculation of diverse hosts grown in monoculture and dual culture. *New Phytologist* 126 (4):677–90.doi:10.1111/j.1469-8137.1994.tb02962.x.

Miller JR., Charles N. 1976. Early Evolution in the Pinaceae. *Review of Palaeobotany and Palynology, Patterns in Gymnosperm Evolution*, 21 (1): 101–17. doi:10.1016/0034-6667(76)90024-5.

Miller, Steven L., Pilar Torres, and Terry M. McClean. 1993. Basidiospore Viability and Germination in Ectomycorrhizal and Saprotrophic Basidiomycetes. *Mycological Research* 97 (2): 141–49. doi:10.1016/S0953-7562(09)80235-9.

Massicotte, Hugues B., Randy Molina, Daniel L. Luoma, and Jane E. Smith. 1994. Biology of the Ectomycorrhizal Genus, *Rhizopogon*. *New Phytologist* 126 (4): 677–90. doi:10.1111/j.1469-8137.1994.tb02962.x.

Molina, Randy, and James M. Trappe. 1982. Patterns of Ectomycorrhizal Host Specificity and Potential among Pacific Northwest Conifers and Fungi. *Forest Science* 28(3):423-458

Molina, Randy, and James M. Trappe. 1994. Biology of the Ectomycorrhizal Genus, *Rhizopogon*: I. Host associations, host-specificity and pure culture syntheses *New Phytologist* 126 (4): 653–75. doi:10.1111/j.1469-8137.1994.tb02961.x.

Molina, Randy, Jane E. Smith, Donaraye Mckay, and L. H. Melville. 1997. Biology of the Ectomycorrhizal Genus, *Rhizopogon*. III. Influence of Co-Cultured Conifer Species on Mycorrhizal Specificity with the Arbutoid Hosts *Arctostaphylos uva-ursi* and *Arbutus menziesii*. *New Phytologist* 137 (3): 519–28.

Molina, R., J. M. Trappe, L. C. Grubisha, and J. W. Spatafora. 1999. *Rhizopogon*. In *Ectomycorrhizal Fungi Key Genera in Profile*, edited by Dr John W. G. Cairney and Dr Susan M. Chambers, 129–61. Springer Berlin Heidelberg.

Murata, Masao, Akihiko Kinoshita, and Kazuhide Nara. 2013. Revisiting the Host Effect on Ectomycorrhizal Fungal Communities: Implications from Host–fungal Associations in Relict *Pseudotsuga japonica* Forests. *Mycorrhiza* 23 (8): 641–53. doi:10.1007/s00572-013-0504-0.

Niculita-Hirzel, Hélène, Jessy Labbé, Annegret Kohler, François Le Tacon, Francis Martin, Ian R. Sanders, and Ursula Kües. 2008. Gene Organization of the Mating Type Regions in the Ectomycorrhizal Fungus *Laccaria bicolor* Reveals Distinct Evolution between the Two Mating Type Loci. *New Phytologist* 180 (2): 329–42. doi:10.1111/j.1469-8137.2008.02525.x.

Ohm, Robin A., Robert Riley, Asaf Salamov, Byoungnam Min, In-Geol Choi, and Igor V. Grigoriev. 2014. Genomics of Wood-Degrading Fungi. *Fungal Genetics and Biology:* FG & B 72 (November): 82–90. doi:10.1016/j.fgb.2014.05.001.

O'Shea, Suzanne F., Pushpalata T. Chaure, John R. Halsall, Natalie S. Olesnicky, Andreas Leibbrandt, Ian F. Connerton, and Lorna A. Casselton. 1998. A Large Pheromone and Receptor Gene Complex Determines Multiple B Mating Type Specificities in *Coprinus cinereus*. *Genetics* 148 (3): 1081–90.

Peay, Kabir G., Thomas D. Bruns, Peter G. Kennedy, Sarah E. Bergemann, and Matteo Garbelotto. 2007. A Strong Species–area Relationship for Eukaryotic Soil Microbes: Island Size Matters for Ectomycorrhizal Fungi. *Ecology Letters* 10 (6): 470–80. doi:10.1111/j.1461-0248.2007.01035.x.

Peay, Kabir G., Martin I. Bidartondo, and A. Elizabeth Arnold. 2010. Not Every Fungus Is Everywhere: Scaling to the Biogeography of Fungal–plant Interactions across Roots, Shoots and Ecosystems. *New Phytologist* 185 (4): 878–82. doi:10.1111/j.1469-8137.2009.03158.x.

Pirozynski KA, Malloch DW; 1975. The origin of land plants: a matter of mycotropism. *Biosystems* 6, 153–164 (1975).

Segelbacher, Gernot, Samuel A. Cushman, Bryan K. Epperson, Marie-Josée Fortin, Olivier Francois, Olivier J. Hardy, Rolf Holderegger, Pierre Taberlet, Lisette P. Waits, and Stéphanie Manel. 2010. Applications of Landscape Genetics in Conservation Biology: Concepts and Challenges. *Conservation Genetics* 11 (2): 375–85. doi:10.1007/s10592-009-0044-5.

Simard SW and Durall DM, 2004. Mycorrhizal networks: a review of their extent, function, and importance. *Canadian Journal of Botany* 82: 1140-1165

Simard SW. 2009a. Mycorrhizal networks and complex systems: Contributions of soil ecology science to managing climate change effects in forested ecosystems. *Canadian Journal of Soil Science* 89: 369-382

Simard, Suzanne W. 2009b. The Foundational Role of Mycorrhizal Networks in Self-Organization of Interior Douglas-Fir Forests. *Forest Ecology and Management, IUFRO Conference on Biodiversity in Forest Ecosystems and Landscapes,* 258, *Supplement (December):* S95–107. doi:10.1016/j.foreco.2009.05.001.

Selosse, M-A, Richard F, He X, and Simard SW. 2006. Mycorrhizal networks: des liaisons dangereuses? *Trends in Ecology & Evolution* 21(11): 621-628.

Smith, SE., and David Read. Editors. 2008. *Mycorrhizal Symbiosis (Third Edition).* 1–9. London: Academic Press.

Smith, AH., and Zeller SM. 1966. Preliminary Account of the North American Species *of Rhizopogon. Memoirs of the New York Botanical Garden* 14(2):177 pgs

Stajich, Jason E., Sarah K. Wilke, Dag Ahrén, Chun Hang Au, Bruce W. Birren, Mark Borodovsky, Claire Burns, et al. 2010. Insights into Evolution of Multicellular Fungi from the Assembled Chromosomes of the Mushroom *Coprinopsis cinerea* (*Coprinus cinereus*). *Proceedings of the National Academy of Sciences* 107 (26): 11889–94. doi:10.1073/pnas.1003391107.

Strauss, Steven H., Allan H. Doerksen, and Joyce R. Byrne. 1990. Evolutionary Relationships of Douglas-Fir and Its Relatives (genus *Pseudotsuga*) from DNA Restriction Fragment Analysis. *Canadian Journal of Botany* 68 (7): 1502–10. doi:10.1139/b90-191.

Taylor, D. L., and T. D. Bruns. 1999. Community Structure of Ectomycorrhizal Fungi in a *Pinus muricata* Forest: Minimal Overlap between the Mature Forest and

Resistant Propagule Communities. *Molecular Ecology* 8 (11): 1837–50. doi:10.1046/j.1365-294x.1999.00773.x.

Taylor, T. N., W. Remy, H. Hass, and H. Kerp. 1995. Fossil Arbuscular Mycorrhizae from the Early Devonian. *Mycologia* 87 (4): 560–73. doi:10.2307/3760776.

Tedersoo, Leho, Tom W. May, and Matthew E. Smith. 2009. Ectomycorrhizal Lifestyle in Fungi: Global Diversity, Distribution, and Evolution of Phylogenetic Lineages. *Mycorrhiza* 20 (4): 217–63. doi:10.1007/s00572-009-0274-x.

Tedersoo, Leho, Mohammad Bahram, Märt Toots, Abdala G. Diédhiou, Terry W. Henkel, Rasmus Kjøller, Melissa H. Morris, et al. 2012. Towards Global Patterns in the Diversity and Community Structure of Ectomycorrhizal Fungi. *Molecular Ecology* 21 (17): 4160–70. doi:10.1111/j.1365-294X.2012.05602.x.

Teste, François P., Suzanne W. Simard, Daniel M. Durall, Robert D. Guy, Melanie D. Jones, and Amanda L. Schoonmaker. 2009. Access to Mycorrhizal Networks and Roots of Trees: Importance for Seedling Survival and Resource Transfer. *Ecology* 90 (10): 2808–22. doi:10.1890/08-1884.1.

Twieg B, Durall DM, Simard SW. 2007. Ectomycorrhizal fungal succession in mixed temperate forests. *New Phytologist* **176**: 437–447.

Vargas-Hernández JJ, López-Upton J, Reyes-Hernández VJ, Domínguez-Alvarez A, Mápula-Larreta M. 2004. Natural populations of Douglas fir in Mexico: current status and needs for conservation. In: Beaulieu J, ed. *Silviculture and the conservation of genetic resources for sustainable forest management: proceedings of the symposium of the North American forest commission, forest genetic resources and silviculture working groups, and the International Union of Forest Research Organizations (IUFRO).* Quebec City, QC, Canada, 26–36.

Velasco-García, MV, J López-Upton, G Angeles-Pérez, J Vargas-Hernández, and V Guerra-de la Cruz. 2007. *Pseudotsuga menziesii* Seed Dispersion in Populations of Central México. *Agrosciencia* 41 (1): 121–31.

Viard, F, Y A El-Kassaby, and K Ritland. 2001. Diversity and Genetic Structure in Populations of *Pseudotsuga menziesii* (Pinaceae) at Chloroplast Microsatellite Loci. *Genome* 44 (3): 336–44. doi:10.1139/g01-010.

Wang, B., and Y.-L. Qiu. 2006. Phylogenetic Distribution and Evolution of Mycorrhizas in Land Plants. *Mycorrhiza* 16 (5): 299–363. doi:10.1007/s00572-005-0033-6.

Wei XX, Yang ZU, Li Y, Wang XQ. 2010. Molecular phylogeny and biogeography of *Pseudotsuga* (Pinaceae): Insights into the floristic relationship between Taiwan and its adjacent areas. *Molecular Phylogenetics and Evolution.* 55(3):776-785.

Wendland, J, L J Vaillancourt, J Hegner, K B Lengeler, K J Laddison, C A Specht, C A Raper, and E Kothe. 1995. The Mating-Type Locus B Alpha 1 of *Schizophyllum commune* Contains a Pheromone Receptor Gene and Putative Pheromone Genes. *The EMBO Journal* 14 (21): 5271–78.

Whitfield J. 2007. Fungal roles in soil ecology: underground networking. *Nature* 449: 136-138

Yabe, Atsushi. 2011. *Pseudotsuga tanaii* Huzioka from the Earliest Miocene Shichiku Flora of Northeast Japan: Systematics and Ecological Conditions. *Paleontological Research* 15 (1): 1–11. doi:10.2517/1342-8144-15.1.001.

Zeller, Sanford M., and Carroll W. Dodge. 1918. *Rhizopogon* in North America. *Annals of the Missouri Botanical Garden* 5 (1): 1–36. doi:10.2307/2990021.

Japan Integrated Biodiversity Information System
http://www.biodic.go.jp/english/rdb/rdb_e.html

The IUCN Red List of Threatened Species
http://www.iucnredlist.org/

Norma Oficial Mexicana NOM-059-SEMARNAT-2010, Protección ambiental-Especies nativas de México de flora y fauna silvestres-Categorías de riesgo y especificaciones para su inclusión, exclusión o cambio-Lista de especies en riesgo.. Mexico City: Secretaría de Medio Ambiente y Recursos Naturales. 2010.

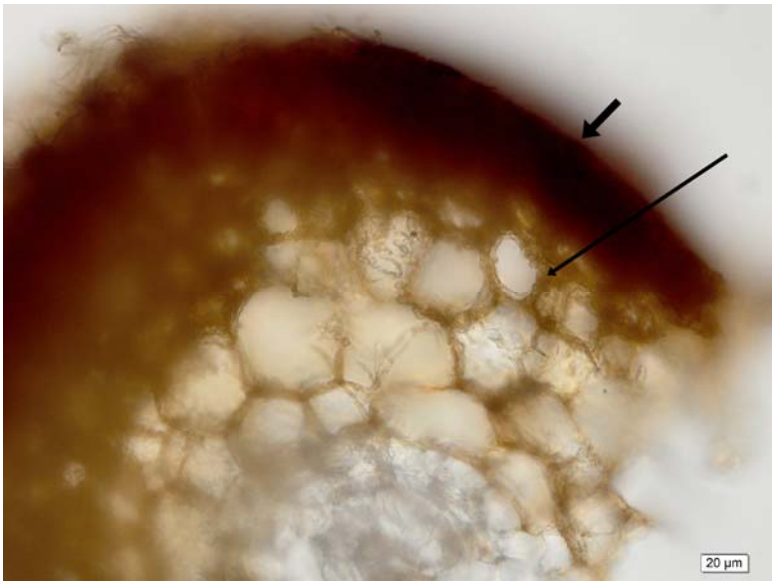


Figure 1.1: Cross section of *Rhizopogon* EM root tip with *Pseudotsuga sinensis.* The dark fungal hyphae of the Hartig net (thin arrow) and fungal mantle (thick arrow) are visible.

Figure 1.2: The genera *Rhizopogon* and *Pseudotsuga*. A) Freshly excavated basidiome of *R. vinicolor* showing fruiting habit and peridium. B) Bisected immature basidiome of *R. togasawariana* showing loculate gleba and rhizomorphs. (Photo Credit: Dr. Kentaro Hosaka) C) *Pseudotsuga menziesii* var. *menziesii.* Note short needles born individually around the entire stem and characteristic seed cones with trilobite bract scales.

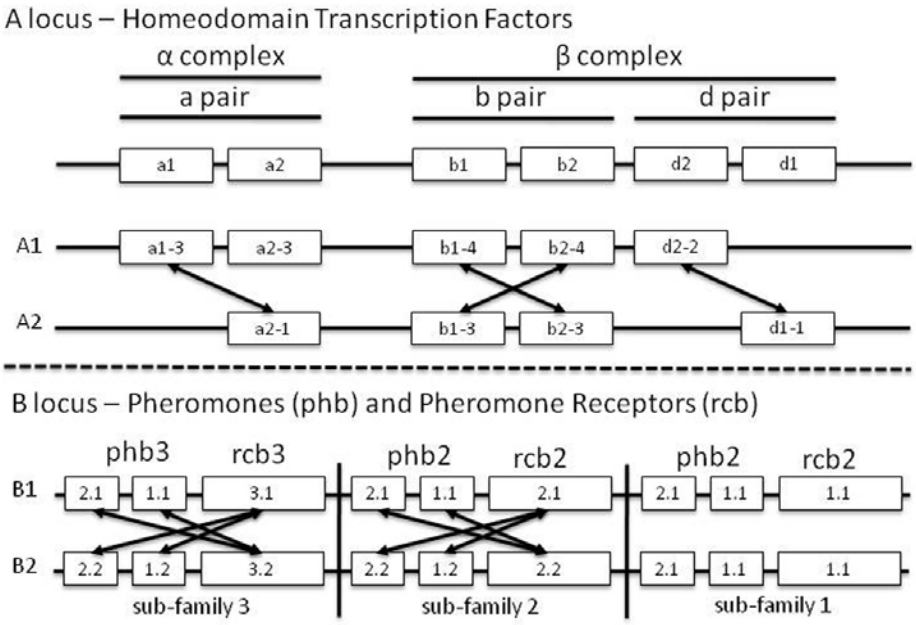


Figure 1.3: A cartoon depicting the A locus and B locus protein coding genes of *Coprinopsis cinerea,* a model fungus in the study of Agaricomycete mating-type systems. For a mating event to proceed between monokaryotic strains a successful pairing of both A locus and B locus proteins is required. A locus homeodomain transcription factor genes are arranged in multiple pairs in and β complexes (subloci). A successful mating event between monokaryotic strains "A1" and "A2" requires at least one pairing of two unlike alleles from the same pair as denoted by arrows. B locus pheromone (phb) and pheromone receptor (rcb) genes are arranged in functional cassettes. A successful mating event between monokaryotic strains "B1" and "B2" requires at least one pairing of unlike pheromone and pheromone receptor alleles from the same cassette as denoted by arrows. Figure adapted from Casselton and Olesnicky (1998).

Chapter 2. Spatial heterogeneity and genotypic diversity affects competitive interactions and resource partitioning between sister species of ectomycorrhizal fungi (*Rhizopogon vinicolor* and *Rhizopogon vesiculosus*).

Alija Bajro Mujic, Peter Kennedy, Daniel Durall, Joseph W. Spatafora

Target Journal: New Phytologist

ABSTRACT:

Vertical resource partitioning is a well documented phenomenon in ectomycorrhizal (EM) fungi and has been proposed as one of strongest explanatory factors for the maintenance of fungal diversity in forest ecosystems. Despite its prevalence in the literature and proposed ecological importance, very little is understood concerning the biotic and abiotic factors that underlie these patterns of niche partitioning in EM fungi. In this study we conduct a series of container plant experiments to experimentally test the role of spatial heterogeneity of soil resources upon competitive outcomes and resource partitioning in the EM sister species *Rhizopogon vinicolor* Smith and Zeller and *R. vesiculosus* Smith and Zeller. In natural forests these fungi are known to partition access to receptive host roots along a clear vertical gradient. We collected upper (organic) and lower (mineral) soils from a site where both species of fungi are known to coexist in near equal frequency and distribution. These soils were used in container plantings of the host tree, *Pseudotsuga menziesii* (Mirab.) Franco (Douglas-fir), either as a homogenized mix to remove the effect of resource heterogeneity or in stratified combination to mimic natural field conditions. We found that neither species had a preference for colonizing roots in either the upper or lower soil stratum despite significant differences in the carbon and nitrogen content of these strata. Under homogeneous soil conditions *R. vesiculosus* was the clear competitive superior and was able to totally exclude *R. vinicolor*. However, *R. vinicolor* was able to persist and occasionally dominate the EM community upon reestablishment of soil heterogeneity through stratification. Where *R. vinicolor* and *R. vesiculosus* were both found within a community, the competitive success of *R. vinicolor* was directly proportional to the

number of *R. vinicolor* multilocus genotypes present indicating that genotypic diversity and colonization by spore inoculum provide a competitive advantage for *R. vinicolor.* Taken together, our findings suggest that vertical niche partitioning in *R. vinicolor* and *R. vesiculosus* is likely to be an indirect interaction of soil resource heterogeneity and competitive strategies of these fungi rather than the edaphic properties of soil strata.

INTRODUCTION:

Niche differentiation based upon resource specialization is one of the oldest explanations for the avoidance of competition and maintenance of diversity (Armstrong and McGehee 1980). This hypothesis is well demonstrated to function in maintaining the diversity of vertebrate animal communities (Schoener 1974) but has been difficult to demonstrate as an explanation of diversity in plant communities, with only a few terrestrial plant systems showing clear patterns of resource partitioning (McKane et al. 2002, Bever et al. 2010). Recent advances in molecular tools have enabled research that supports resource partitioning as a potential explanation for diversity in communities of ectomycorrhizal (EM) fungi (Bruns 1995). EM fungal communities are thought to compete for two primary resources: 1) the fine roots of compatible EM host plants that supply the fungus with an organic carbon source (photosynthates) and 2) soil nutrients that are foraged by the extramatrical hyphae which connect colonized EM root tips to the surrounding substrate (Bruns 1995).

Spatial heterogeneity of resources is well documented as a driver of species diversity in natural communities of animals (Roth 1976, Ricklefs 1977 and references therein) and plants (Ricklefs 1977) and there is mounting evidence which suggests that

spatial heterogeneity of resources is a driver of diversity in EM fungal communities as well. Spatial resource partitioning in EM fungi has been observed between soil micro-sites due to heterogeneous distribution of organic debris (Tedersoo et al. 2003) and many studies have found evidence for vertical partitioning of EM fungal taxa between forest soil strata (Dickie et al. 2002, Rosling et al. 2003, Baier et al. 2006, Scattolin et al. 2008, Beiler et al. 2012, Anderson et al. 2014). These findings suggest that the differences in edaphic factors between soil micro-sites and strata establish spatial heterogeneity within the environment and that some EM fungal taxa possess fundamental niches, which are defined by particular patch types. It has been hypothesized that such niche differentiation can allow for a large diversity of fungi to coexist even in stands with low host tree diversity (Dickie et al. 2002, Anderson et al. 2014). A high diversity and abundance of EM fungi can provide many benefits to forest health. Colonization of roots by EM fungal communities enhances the survivorship of young trees and increases forest ecosystem resilience to disturbance (Trappe and Strand 1969, Cripps et al. 2004, Baier et al. 2006, Luoma et al. 2006, Simard 2009, Smith and Read 2008). An increase in the diversity of a forest's EM fungal community is correlated with increased nutrient uptake by trees (Baxter and Dighton 2001) and has been proposed to increase plant productivity (Read and Perez-Moreno 2003; van der Heijden and Sanders 2002). These effects of EM fungi are important modulators of resource availability in plant communities and may themselves drive resource partitioning and community structure amongst plants (Bever et al. 2010).

Most studies investigating the vertical partitioning of resources in EM fungi have taken a broad view of natural systems by sampling the complete EM fungal community.

However, recent research performed by Beiler et al. (2012) mapped the fine scale spatial structure of EM root tips colonized only by the sister species *Rhizopogon vinicolor* Smith and Zeller and *R. vesiculosus* Smith and Zeller (Boletales, Basidiomycota). This study found patterns of vertical resource partitioning between the two species with *R. vesiculosus* occupying deeper soil strata and *R. vinicolor* occupying shallower soil strata where the two species co-occurred in a vertical column of soil. These *Rhizopogon* species share a sympatric range within the Pacific Northwest of North America (Kretzer et al. 2005, Dunham et al. 2013, Beiler et al. 2010, 2012) and are obligate EM symbionts of a single host species, *Pseudotsuga menziesii* (Mirb.) Franco (Douglas-fir) (Molina and Trappe 1982, Molina et al. 1999). Genus *Rhizopogon* produce basidiospores within hypogeous sporocarps that are submerged in the upper organic horizon of the soil and dispersal is achieved through excavation and consumption of these sporocarps by mammals (Izzo et al. 2005). *R. vinicolor* and *R. vesiculosus* are both components of the EM fungal community colonizing the roots of *P. menziesii* at all stand ages(Luoma et al. 1991, Smith et al. 2002), but are especially abundant in young stands (Twieg et al. 2007) and following stand replacing disturbances such as fire or logging (Baar et al. 1999, Luoma et al. 2006). When the species co-occur at the same site they are often sampled as both EM root tips and sporocarps at near equal frequencies (Kretzer et al. 2005, Beiler et al. 2010, Dunham et al. 2013). The sympatric distribution, close evolutionary relationship, and obligate reliance upon a single host species suggest high levels of competition for resources between these two species. However, equal abundances of both species at a single site suggest that these species possess differences in fundamental niche or near equal competitive abilities. Competitive interactions between species can be

important in structuring EM fungal communities (Koide et al. 2005, Kennedy et al. 2007) but the role of competition in establishing patterns of vertical niche partitioning are unknown. Vertical partitioning of EM root tips as observed by Beiler et al. (2012) is consistent with the hypothesis that vertical partitioning of resources can function to maintain community diversity and suggests a possible explanation for how interspecies competition is minimized and coexistence achieved.

The goal of the current study is to investigate environmental factors that may drive the observed patterns of vertical resource partitioning in *R. vinicolor* and *R. vesiculosus* and to perform an experimental test of spatial heterogeneity and competition as drivers of community diversity. To this end, we performed a series of container plant experiments to separate the effects of competition and depth-associated edaphic factors upon these taxa in single or dual inoculation treatments. There are several life history traits of *R. vinicolor* and *R. vesiculosus* that are relevant to the design of these experiments. *Rhizopogon* species are classified as "long distance" exploration types (Agerer 2001) and produce smooth EM root tips with few but highly differentiated rhizomorphs. *R. vinicolor* and *R. vesiculosus* both produce "long distance" planar mat-like colonies but the pervasiveness of *R. vesiculosus* hyphae is greater than that of *R. vinicolor* at the small scale (20 cm range) with the maximum diameter of its hyphal aggregations growing up to two times the diameter of those produced by *R. vinicolor* (Beiler et al. 2012). *R. vesiculosus* also produces larger average genets (~18 meter diameter) than *R. vinicolor* (~6 meter diameter) (Kretzer et al. 2005, Beiler et al. 2010, Dunham et al. 2013). Larger clone size and more pervasive exploratory hyphae both suggest that *R. vesiculosus* is more efficient at colonizing root tips through vegetative

expansion. EM fungi capable of long-range growth in the soil are likely to compete favorably for root tips when root tip density is low (Peay et al. 2011). Thus, it stands to reason that the differences in genet size and hyphal exploration between *R. vesiculosus* and *R. vinicolor* could affect competition and resource partitioning across a root density gradient.

Given the life history traits discussed above we set out to test three hypotheses: 1) The fundamental niche of *R. vinicolor* and *R. vesiculosus* do not overlap and differ such that resources/conditions in the soil better suit the growth of each species at different soil depths. 2) The fundamental niches of *R. vinicolor* and *R. vesiculosus* overlap and their realized niches are determined through competitive interactions. 3) *R. vinicolor* and *R. vesiculosus* show differing root colonization strategies that are driven by differences in hyphal exploration types. Through our container plant experiments we found that the physical and chemical properties of the organic and mineral soil horizons do not directly affect the fundamental niches of these *Rhizopogon* species at the scale investigated. *R. vesiculosus* is the competitive superior, able to completely exclude *R. vinicolor,* under conditions where resource distribution is spatially homogenous. However, where our planting treatments mimicked natural conditions of heterogeneous soil stratification *R. vinicolor* competes successfully and persists as a component of the EM community. Additionally, when the two species do coexist the proportion of the community represented by *R. vinicolor* is directly proportional to the number of unique *R. vinicolor* multilocus genotypes (MLGs) present. To the knowledge of the authors, this is the first study to demonstrate a correlation between an increased rate of mating and competitive success in an EM fungus.

MATERIALS AND METHODS:

Field site description:

Our container plant experiments were designed to mimic the conditions of naturally occurring soil from *P. menziesii* forests where *R. vinicolor* and *R. vesiculosus* are known to co-occur. All field materials used in this study were collected at a site on the north slope of Marys Peak in the Oregon Coast Range (44°32.1′N, 123°32.1′W, maximum elevation = 516 meters). Ecological and population genetic studies of *R. vinicolor* and *R. vesiculosus* indicate that both species occur on this site at near equal frequencies (Kretzer et al. 2005, Dunham et al. 2013). Site conditions are described in detail in Dunham et al. (2013).

Materials collection and preparation:

On average, *R. vinicolor* produces genets that are distributed at shallower depths (median sample depth = 4 – 6 cm) than those of *R. vesiculosus* (median sample depth = 8 – 13 cm) (Beiler et al. 2012). Thus, we chose to sample soil according to depth rather than dividing soil samples directly into their component horizons. In the spring of 2012 we sampled soil from 11 locations spanning our collection site. At each location we harvested three soil cores (15 cm x 5 cm) spaced 0.5 meters apart. Cores were divided into an upper (0 – 7.5 cm depth) and lower fraction (7.5 – 15 cm depth) at the time of harvest. Each core fraction was passed through a 2 mm USA standard testing sieve (no. 10, Fisher Scientific Company, Pittsburg PA) and fine roots (<2 mm diameter) of *P. menziesii* were washed, dried at 37 C, and quantified by weighing upon a microscale

(Mettler-Toledo, Columbus, OH). The sieved soil fractions were then pooled into upper and lower soil depth classes. Pooled sifted soil was layered 1.5 inches thick in trays and autoclaved four times for 1.5 hours at each session, with 24 hours incubation at 25 C between autoclaving. A portion of sterilized upper depth class soils were mixed 1:1 with sterilized lower depth class soils to create a third soil type that blended the properties of both depth classes. Sterile soil from the upper, lower, and blended depth classes were mixed 1:1 with sterile coarse sand to generate a total of three unique soil mixes for use in pot culture experiments.

Field collections of *Rhizopogon* sporocarps occurred between May and August 2011. Basidiospores of *Rhizopogon* species darken from a hyaline state as they mature (Smith and Zeller 1966) and thus sporocarps were qualified as mature if the interior spore bearing tissue (gleba) was soft and dark brown in color. Four mature sporocarps of both *R. vinicolor* and *R. vesiculosus* were used in preparing inoculum. These sporocarps were chosen from collections separated by at least twenty meters in the field to increase the probability that they were produced by different fungal individuals. The surface tissue (peridium) of sporocarps was removed and a small section of gleba tissue was subjected to DNA extraction for species identification. Remaining gleba tissue was submerged in 30 ml of deionized water (dH_2O), homogenized using an immersion blender, and filtered through cheesecloth to remove tissue fragments. The resulting spore slurry was brought to 50 ml volume and stored at 4 C until needed. A total of 8 slurries, each containing the spores of a single sporocarp, were created. Spores in each slurry were quantified using a Bright-Line hemocytometer (Hausser Scientrific, Horsham, PA) and slurries of each

species were pooled in equal spore proportions to create a master inoculation solution for each species.

Seeds of *P. menziesii* were obtained from Silvaseed Company (Roy, WA). Surface sterilization of seeds was first performed to remove any endogenous EM fungal inoculum. Seeds were placed in a 10% solution of hydrogen peroxide (H_2O_2) for 15 minutes with stirring at 400 rpm. Seeds were then rinsed twice in dH_2O and soaked in a fresh solution of dH_2O for 24 hours with stirring at 200 rpm. Any seed that was found floating on the surface after soaking was discarded and the remaining seeds were prepared for germination by cold stratification. Seeds were wrapped in autoclaved cheesecloth and soaked in dH_2O for another 24 hours. Excess water was removed from the cheesecloth seed pouch by shaking vigorously. The seed pouch was then sealed into a zip seal plastic bag and ventilation was provided by three straws inserted through the seal of the bag, extending from outside the bag to the seed pouch. The ventilated seed bag was incubated at 4 C for 28 days.

Planting treatments: setup, maintenance, and harvest:

We tested the three hypotheses of this study by performing three container plant experiments that manipulated depth associated edaphic properties and competitive interactions. The experimental unit of replication was a single "Cone-tainer"^{TM} planting cell (SC7 stubby, Stuewe and Sons, Tangent, OR) planted with a single *P. menziesii* seedling. The bottom of each planting cell was lined with synthetic pillow stuffing and 80 ml soil mix was added to the cell. Three cold stratified seeds of *P. menziesii* were placed on top of the soil mix and each cell was topped with 10 ml autoclaved coarse sand.

Planting treatments manipulated soil stratification by varying the soil depth class present in the lower 40 ml and upper 40 ml container volumes. Inoculation of the upper and/or lower container volumes was achieved by premixing *Rhizopogon* spore slurries into the soil mix used to fill that volume. Each planting cell replicate received a total of 1×10^7 spores from either one *Rhizopogon* species or both species in composite.

A diagrammatic description of experimental treatments, replication, and setup is detailed in Figure 2.1. A brief explanation of each experiment and its goals is as follows. Experiment 1 manipulated soil stratification conditions of the upper and lower soil depth classes with single species inoculations to determine the role of edaphic properties in the delimitation of *Rhizopogon* fundamental niches. Experiment 1 contained four treatments, which tested each species in isolation of competition with soil stratification that either mimicked natural conditions or restricted growth to a single soil depth class. Experiment 1 was designed to investigate the role of edaphic properties in the delimitation of *Rhizopogon* fundamental niches and its treatments are hereafter referenced as treatments 1.1, 1.2, 1.3 and 1.4. Experiment 2 used blended depth class soils with both single and dual species inoculation treatments to determine the role of competition in the delimitation of *Rhizopogon* realized niches. A negative control treatment of uninoculated blended depth class soil was included in experiment 2 to ensure that endogenous EM fungal inoculum in field soil was destroyed by autoclaving. Experiment 2 contained four treatments, which are hereafter referenced as treatments 2.1, 2.2, 2.3 and 2.4. Experiment 3 combined both soil depth class stratification and dual species inoculations to investigate the competitive behavior of the two *Rhizopogon* species under conditions closer to those

encountered in the field. Experiment 3 contained one treatment that is hereafter referenced as treatment 3.1.

Planting treatments for all experiments were established on May 21 – 22, 2012 and incubated in a growth chamber at 20 C with a 12-hour light/dark cycle. From the time of planting until the time of seedling germination replicate containers were kept evenly moist by hand watering. At the time of germination each replicate conetainer was thinned to a single seedling and watering commenced thereafter on a 3-day interval. Conetainers were placed in rows of 7 in a 98-cell plant rack (RL98, Stuewe and Sons, Tangent, OR). Blank rows of 7 were left between planting treatments to avoid cross contamination. Racks were periodically rotated within the growth chamber and treatments were rotated within each rack to eliminate effects associated with environmental heterogeneity.

Conetainers were destructively harvested from December 13 – 14, 2012. Each seedling was cut at the soil surface and all aerial parts of the seedling were dried at 37 C for 48 hours and quantified using a microscale. The soil and root ball of each replicate was cut in half between the upper and lower 40ml volume of soil. The soil from both the upper and lower volumes of all replicates within a treatment was pooled in zip seal plastic bags and stored at 4 C until further processing. Roots from the upper and lower volume were washed in tap water, stored in dH_2O at 4 C, and processed within 2 weeks. Roots from the upper and lower volumes of each planting cell were examined with stereomicroscopy and all EM root tips were harvested into 1.5 ml microcentrifuge tubes, submerged in dH_2O, lyophilized, quantified with a microscale, and stored at -20 C until further processing. Remaining fine root (< 2 mm diameter) mass from the upper and

lower volumes was separated from the major root axis and roots were dried at 37 C for 48 hours and quantified using a microscale.

Molecular data collection:

Root tips morphologies of *R. vinicolor* and *R. vesiculosus* are highly similar and molecular identification is necessary to distinguish these species on *P. menziesii* roots. Lyophilized EM root tips from the upper and lower soil volumes of each replicate were cleaned and individual root tips were selected using a dissecting stereoscope. EM root tips and parent sporocarps were identified to species and genotyped using the methods and markers developed by Kretzer et al. (2000, 2003, 2005), with slight modifications. Briefly, the species identity of root tips and sporocarps was confirmed by amplifying the ITS rDNA region using the primers ITS1F and ITS4 (White et al. 1990, Gardes and Bruns 1993) and then digesting PCR products using the enzyme *Alu*I. Restriction digests were visualized and scored directly on 2% agarose gels stained with ethidium bromide. Six microsatellite loci were used to identify the multilocus genotypes (MLGs) of *R. vinicolor* (Rv15, Rv46, Rv53, Rve3.21, Rve2.77, and Rve1.34) and five microsatellite loci were used to identify MLGs of *R. vesiculosus (*Rve2.10, Rve2.14, Rv02, Rve1.34, and Rve2.44). DNA was extracted from EM root tips using the "dilution protocol" of the PHIRE Plant Direct PCR Kit (Thermo Fisher Scientific, Waltham, MA) following the manufacturer's protocol. DNA was extracted from sporocarp tissue using the fastDNA kit (MP Biomedicals, Santa Ana, CA) following manufacturer's protocol. DNA extractions from EM root tips were diluted 1:10 with molecular grade dH_2O and PCR of all extractions was carried out using the PHIRE Plant Direct PCR Kit following the

manufacturer's protocol. Four EM root tips (two tips from the upper and lower volumes) were identified to species from each replicate of treatments 1.1, 1.2, 1.3, 1.4, 2.1, and 2.2. Twenty root tips (ten tips from the upper and lower volumes) were identified to species from each replicate of treatments 2.3 and 3.1. Genotyping of EM root tips was performed for up to ten root tips (five tips from the upper and lower volumes) chosen arbitrarily from each replicate of treatments 2.1, 2.2, and 3.1.

Collection of environmental parameters:

We collected nutrient and pH data of live soils and autoclaved soils for both the upper (0 – 7.5 cm) and lower (7.5 – 15 cm) sieved soil depth classes. Total nitrogen and carbon of soils was measured at the Crop and Soil Central Analytical Laboratory at Oregon State University using a CNS-2000 Macro Analyzer (Leco Corporation, St. Joseph, MI). Measurements of soil pH were made by diluting soil 1:2 in sterile deionized water, shaking thoroughly and incubating at room temp for 20 minutes before measurements were made using a sympHony pH meter (VWR, Radnor, PA) and Orion Ross pH electrode (ThermoFisher Scientific, Waltham, MA). Eight replicate measurements of C and N and six replicate pH measurements were made from both live and autoclaved sieved soil from the pooled upper and lower soil depth classes. Post experiment, harvested soils were sieved to removed coarse sand and single C and N measurements were made from upper depth class soils used in treatments 1.1, 1.2, and 3.1 as well as lower depth class soils used in treatments 1.1, 1.3, 1.4, and 3.1.

Soil moisture data was collected at three time points (05/22/2013, 06/22/2013, 09/10/2013) during the spring and summer of 2013. During each collection trip three soil

cores were collected as previously described from a subset of five of the same collection sites used for initial soil collection. Each core was divided into upper and lower depth classes and each core fraction was weighted upon a scale, dried at 49 C for 48 hours, and then weighted again to determine total moisture content.

RESULTS:

Root density and percent colonization:

We quantified the density of fine roots (<= 2 mm) present in the upper and lower volumes of field soil cores and conetainer replicates. Root density was calculated as the milligrams of fine root mass divided by the ml of soil volume sampled. To determine root density in the field 147 ml of soil was sampled from both the upper (0 – 7.5 cm) and lower (7.5 – 15 cm) fractions of each field soil core and root density measures of experimental replicates were made by sampling 40 ml of soil in both upper and lower experimental replicate volumes of all treatments. Root density was highly variable both within and between experimental treatments and ranged from 0.5 mg/ml to 3.77 mg/ml among all replicate depth volumes. Root density measures collected from field soil cores were also variable both within and between soil depth classes and ranged from 0 mg/ml to 10.12 mg/ml. There was no significant difference in root density between the combined dataset of the upper or lower experimental replicate volumes and field soil core volumes (ANOVA, $p = 0.434$). However, the root density of the upper field soil depth class was significantly greater than that of lower field soil depth class within each core (paired t-test, $p = 0.02$). This trend was, in general, not observed in any of the

experimental treatments and there was no significant difference between the upper and lower replicate volumes within treatment except in treatment 1.2 which showed significantly increased root density in the lower soil volume (paired t-test, $p = 0.004$, mean of upper volume= 1.38 mg/ml, mean of lower volume = 2.63 mg/ml).

We quantified the level of EM colonization for the upper and lower volumes of each conetainer replicate. The percent of root mass that had undergone EM colonization was calculated as the weight of all EM root tips divided by the total combined mass of all EM root tips and fine roots (< 2 mm) multiplied by 100. There was no significant difference in the EM colonization between treatments in any single-species stratified soil treatment (1.1, 1.2, 1.3, 1.4) (ANOVA with post hoc Tukey's HSD, all p values greater than 0.2). In mixed soil treatments (2.1, 2.2, 2.3) only the lower volume of treatment 2.1 and the upper volume of treatment 2.3 differed significantly in between treatment EM colonization (ANOVA with post hoc Tukey's HSD, $p = 0.01$). Despite a general lack of between treatment differences, a strong trend of difference in EM colonization was observed between the upper and lower replicate volumes within replicates. Significantly greater EM colonization was observed in the upper replicate volumes compared to lower replicate volumes within treatments 1.1 (paired t-test, $p = 0.006$), 1.4 (paired t-test, $p = 0.024$), 2.1 (paired t-test, $p = 0.01$), 2.2 (paired t-test, $p = 0.012$), and 3.1 (paired t-test, $p = 0.017$). Differences in pairwise EM colonization between upper and lower soil replicate volumes were observed irrespective of species and soil-type and there was no correlation between the percent EM colonization and root density in any experimental replicate volume for either *R. vinicolor* (Simple linear regression model, $p = 0.15$, $R^2 = 0.027$) or *R. vesiculosus* (Simple linear regression model, $p = 0.2$, $R^2 = 0.013$). A complete listing

of all EM colonization data recorded in this study can be found in Supplementary Table 2.S1.

Collection of environmental parameters:

A summary of the results from this sampling is shown in Figure 2.2c and Figure 2.2d. The upper and lower depth class soils of unsterilized field soil varied significantly in both [C] and [N] (C: $n=6$, $p < 0.001$; N: $n=6$, $p < 0.001$; two-tailed t-test). Sterilization treatment of field soil did produce a consistent decrease in total [C] for both the upper and lower soil depth classes, but the sterilized upper and lower soil depth classes still varied significantly in both [C] and [N] (C: $n=5$, $p = 0.002$; N: $n=5$, $p < 0.001$; two-tailed t-test). Average soil pH in the upper and lower depth classes of unsterilized field soil were found to be of approximately equal acidity (unsterilized upper soil pH = 5.19, n=6; unsterilized lower soil pH = 5.21, n=6) while sterilization treatments of field soil marginally changed the average pH of upper and lower depth class field soils by less than 0.2 pH (sterile upper soil pH = 5.06, n = 6; sterile lower soil pH = 5.39, n = 6). Soil moisture data was collected from field soil at three time points during the 2013 dry season. Five of the original eleven soil sample sites were selected for characterization of soil moisture. Average soil moisture was significantly greater in the upper soil depth class (0 – 7.5 cm) than in the lower soil depth class (7.5 –15 cm) at all three time points (two-tailed t-tests, T1 $p=0.0005$, T2 $p=0.034$ T3 $p =0.049$) with a decreasing trend throughout the dry season (Figure 2.2).

Molecular data collection: Species level identification by ITS RFLP:

Several of the seedlings in conetainer replicates died during the duration of the experiment and from the total 732 EM root tips intended for harvest, 541 were harvested and all were identified to species by RFLP analysis (Table 2.1). All single species treatments were found to be free of cross contamination and the fungus present on all EM root tips tested was either *R. vinicolor* or *R vesiculosus,* as expected by treatment (Figure 2.1, Figure 2.2). Sterilization treatments of field soil were found to be effective as all root tips harvested from the negative control treatment 2.4 were examined at 60X magnification and found to be completely without signs of EM symbiosis. In both of two species inoculation treatments (2.3, 3.1) *R. vesiculosus* was the most prevalent species detected on EM root tips (Table 2.1). *R. vesiculosus* was the dominant competitor and was the only species identified upon EM root tips in the blended soil, two species treatment (2.3). Colonization by *R. vesiculosus* alone across all replicates of treatment 2.3 varied significantly from a 1:1 colonization model assumed under equal rates of colonization for both *R. vinicolor* and *R. vesiculosus* (χ^2 goodness of fit = 180, $p < 0.001$, n =180). In contrast, *R. vinicolor* was found to represent from 0% to 59% of the total EM community within a replicate of the stratified soil two species treatment (3.1), while *R. vesiculosus* occupied the remaining community percentage. Given a 1:1 colonization model, *Rhizopogon vesiculosus* was still the dominant competitor in treatment 3.1 (χ^2 goodness of fit = 75.6, $p<0.001$, n=169) and it represented 83% of all root tips sampled from treatment 3.1 while *R. vinicolor* represented the remaining 17%. While *R. vesiculosus* was the dominant competitor in both mixed species treatments, the stratified soil treatment (3.1) showed significantly increased colonization by *R. vinicolor* over the

homogenized soil treatment (2.3) (χ^2 test of independence = 30.2, $p < 0.001$, n=349). No pattern of vertical resource partitioning was observed where *R. vinicolor* and *R. vesiculosus* co-occurred in the same replicate of treatment 3.1 as both species were found colonizing both upper and lower soil depth classes (Table 2.1).

Molecular data collection: Identification of MLGs

Both *R. vinicolor* and *R. vesiculosus* showed reduced allelic diversity from the parent sporocarp generation to the F1 EM root tip generation (Table 2.2). From the four sporocarps of each species used in inoculation, four unique MLGs were identified from the *R. vinicolor* sporocarps but only three unique MLGs were identified from the *R. vesiculosus* parent sporocarps. This indicates that the representation of unique MLGs in the *R. vesiculosus* spore slurry was not evenly distributed, with a single MLG providing 50% of the spores present. Neither *R. vinicolor* nor *R. vesiculosus* are known to produce asexual propagules or possess a high rate of secondary homothallism (Dunham et al. 2013). It follows that allelic recombination in an F1 population of these species occurs by the crossing of monokaryotic hyphae germinated from unique basidiospores. Alleles from the parent sporocarps AM-OR11-051 (*R. vinicolor*) and AM-OR11-058 (*R. vesiculosus*) were not detected in the F1 generation EM root tips and most likely did not contribute alleles in mating. These parents possessed only unique alleles at the loci Rv 53 and Rve 1.21, respectively, and these alleles were completely absent from our subsample of the F1 generation.

Both *R. vinicolor* and *R. vesiculosus* were commonly detected as multiple MLGs within a single conetainer replicate (Table 2.3). Some unique MLGs were detected in

multiple conetainer replicates, but this trend was significantly more prevalent (Wilcoxian rank sum test, p = 0.04) in *R. vesiculosus* with 7 MLGs present in multiple replicates and an average 6.6 repeated detections of the same MLG from independent replicates. In comparison, only 4 *R. vinicolor* MLGs were present in multiple replicates with an average 2.25 repeated detections of the same MLG from independent replicates. Of the 78 *R. vinicolor* and 154 *R. vesiculosus* EM root tips genotyped a total of 28 unique MLGs were detected for *R. vinicolor* and 8 were detected for *R. vesiculosus*. We used Punnett squares to calculate the probability of observing a given MLG as the expected frequency of that genotype given all potential pairings of alleles from parent sporocarps (Table 2.3). Using this approach we determined that the F1 generation of *R. vesiculosus* root tips could possess 10 possible MLGs and our sampling of 8 MLGs represented 80% of the total possible genetic diversity. In contrast, the F1 generation of *R. vinicolor* root tips could possess 297 possible MLGs and our sampling of 28 MLGs represented only 9% of the total possible genetic diversity. Despite the greater diversity of *R. vinicolor* MLGs observed across the sum of conetainer replicates, there was no significant difference between the mean genotypic diversity (t-test, expected heterozygosity (Hexp) in Table 2.3, p = 0.4) or evenness (t-test, E.5 in Table 2.3, p = 0.53) in individual conetainer replicates that contained *R. vinicolor* or *R. vesiculosus*. Figure 2.3 shows a plot comparing the number of MLGs for both species in each conetainer replicate volume (both upper and lower 40 ml) of treatment 3.1 against the percent of the EM community represented by that species in the same conetainer volume. This plot shows a positive linear relationship (p = 0.046, R^2 = 0.5115) between the number of *R. vinicolor* MLGs and the total percent of the community occupied by *R. vinicolor* (Figure 2.3). No

significant relationship was observed between the number of *R. vesiculosus* MLGs and the total percent of the community it occupied in treatment 3.1.

DISCUSSION:

Spatial heterogeneity of resources and competitive success:

We conducted a series of three container plant inoculation experiments using the host tree *Pseudotsuga menziesii* and the EM fungal species *R. vinicolor* and *R. vesiculosus.* These experiments were designed to investigate patterns of vertical resource partitioning that have been observed in the natural setting for these species (Beiler et al. 2012) and to experimentally test the role of spatial resource heterogeneity, or patchiness, in the maintenance of EM fungal diversity. In our first experiment, we inoculated each species of *Rhizopogon* individually into two field soil strata and quantified the amount of root colonization to determine if differences in fundamental niche naturally drive differential colonization by these species in the upper or lower soil stratum. Though the upper field soil stratum possesses significantly greater carbon and nitrogen than the lower stratum (Figure 2.2), we detected no difference in the amount of colonization by either species in either of the two strata in any treatment of experiment one. This finding suggests that both *R. vinicolor* and *R. vesiculosus* share a very similar fundamental niche that is not affected by the edaphic factors which differ between the upper and lower soil strata.

Our second experiment removed the effect of soil heterogeneity by homogenizing both the upper and lower soil strata into a single mixed soil class. We inoculated mixed

soil with both *Rhizopogon* species individually and in combination to investigate the role of interspecies competition upon root colonization. As was seen in the results of experiment one, we found no significant difference in the ability of either species to colonize root tips when inoculated individually into mixed soils. However, when both species were co-inoculated into homogenized soils we observed complete competitive exclusion of *R. vinicolor* by *R. vesiculosus*. The competitive superiority of *R. vesiculosus* is not likely to be the result of an emergent property of mixed soils. This is corroborated by the results of experiment one which showed that neither species displays a preference for colonizing host roots in either soil strata. Rather, it is likely that with significant overlap in their fundamental niche, *R. vinicolor* and *R. vesiculosus* vertically partition resources because of differing realized niches as a result of competitive exclusion. Our third experiment was designed to complement experiments one and two by combining interspecies competition with a heterogeneous soil medium stratified to mimic field conditions. With the reestablishment of soil resource patchiness in experiment 3, we no longer observed complete competitive exclusion of *R. vinicolor* by *R. vesiculosus.* Rather, *R. vinicolor* persisted and occasionally dominated the observed EM community but neither *R. vinicolor* nor *R. vesiculosus* showed any preference for a particular stratum and equal colonization of both the upper and lower soil strata was observed for both species (Table 2.1).

Taken together, the results of experiments one, two, and three suggest that under homogenous conditions *R. vesiculosus* is the competitive superior to *R. vinicolor* in a competitive hierarchy. In such a hierarchy, introduction of one or more EM fungal competitors which are competitively superior to *R. vesiculosus* but inferior to *R. vinicolor*

could function to maintain the relatively equal balance observed between these species in natural settings (Bruns 1995, Kennedy et al. 2007). However, results of our third experiment suggest that patchiness within the soil resource allows *R. vinicolor* to successfully compete with *R. vesiculosus.* Neither *R. vinicolor* nor *R. vesiculosus* display a preference for colonization within a particular soil stratum in any of our experiments and thus it is soil heterogeneity itself, rather than properties of a particular patch within the heterogeneous mosaic, that is the underlying requirement for vertical resource partitioning and the maintenance of diversity in these EM taxa.

Genotypic diversity and competitive success:

Our study demonstrates for the first time a positive linear correlation between competitive success and genotypic diversity as a function of mating in an EM fungus. When soil was stratified to mimic natural conditions, *R. vinicolor* was able to successfully compete with *R. vesiculosus* and its share of resources was positively correlated with the number of unique *R. vinicolor* MLGs present. No such correlation between competitive success and the number of unique MLGs was observed for *R. vesiculosus* in any treatment (Figure 2.3). It is important to note that these observations were made despite the fact that the mean genotypic diversity and evenness of *R. vinicolor* and *R. vesiculosus* did not differ significantly at the scale of an individual conetainer (Table 2.3). The correlation between increased genotypic diversity and competitive success in *R. vinicolor* is further strengthened by the limited proportion (9%) of potential *R. vinicolor* genetic diversity captured by our relatively small sample size (n = 78). We predict that this trend would gain stronger support with larger sample sizes that capture a

greater proportion of the potential population diversity. Our observations of higher total genotypic diversity in *R. vinicolor* are consistent with previous observations from the field site used in this study which found that *R. vesiculosus* and *R. vinicolor* were detected with near equal frequencies within a 16 hectare plot, but that *R. vesiculosus* was represented by fewer MLGs and showed a pattern of inbreeding while *R. vinicolor* was breeding randomly (Dunham et al. 2013).

Competitive resource partitioning and root colonization strategies:

Our findings also indicate that differing root colonization strategies may facilitate competitive partitioning of resources by *R. vinicolor* and *R. vesiculosus.* We found that with approximately twice the sample size, *R. vesiculosus* root tips (n = 154, 8 MLGs detected) possessed less than one third the number of MLGs that were detected in *R. vinicolor* root tips (n = 78, 28 MLGs detected). These observations are particularly striking given that even with a low number (10) of potential *R. vesiculosus* MLGs we failed to sample all possible genotypes and found a significant tendency to observe the same *R. vesiculosus* MLG across multiple conetainer replicates (Table 2.3). Our findings of lower total genetic diversity in *R. vesiculosus* considered along with the larger average genet size of *R. vesiculosus* (Kretzer et al. 2003, 2005, Dunham et al. 2013, Beiler et al. 2010, 2012) and its tendency to create more pervasive mycelium around host tree roots (Beiler et al. 2012) suggest that *R. vesiculosus* relies upon vegetative expansion to colonize new root tips. It is likely that *R. vesiculosus* utilizes a competitive strategy of root tip colonization where relatively few mating events upon root tips produce a few unique MLGs which colonize additional root tips through extensive vegetative growth. In

contrast, *R. vinicolor* seems best able to compete through direct colonization of root tips at the site of spore germination and mating which results in a more diverse population of smaller MLGs. Under such a scenario *R. vesiculosus* would gain the competitive advantage of additional carbon resources available from established connections with the host tree. In sufficient numbers, however, early colonizing MLGs of *R. vinicolor* could resist invading *R. vesiculosus* mycelium by priority effects. Priority effects have been demonstrated as a key factor in determining the outcome of competition between several other *Rhizopogon* species with the earliest colonizing species gaining the greatest competitive advantage (Kennedy and Bruns 2005, Kennedy et al. 2009). If only a small proportion of the root system is occupied by an early colonizing *Rhizopogon* species, however, the competitive advantage granted by priority effects can be reduced or eliminated because many receptive roots are still available for competitors (Kennedy et al. 2009). In cases where *R. vinicolor* possesses a greater number of unique MLGs it can be assumed that these MLGs arose through independent spore germination events upon different root tips. Thus, the observed correlation between competitive success and an increase in the number of unique MLGs in *R. vinicolor* would be consistent with the strengthening of priority effects through the colonization of a greater proportion of available root tips. It is also possible that the observed competitive advantage of *R. vesiculosus* is minimized when heterogeneous conditions in the soil matrix slow the growth of *R. vesiculosus* long enough to allow *R. vinicolor* to colonize a greater proportion of host roots by spores and thus strengthen the competitive advantage gained through priority effects.

Root density:

In light of *R. vinicolor's* apparent competitive inferiority and the lack of preference in either species for either depth class of soil, the persistence of *R. vinicolor* in upper field soil horizons (Beiler et al. 2012) suggests that secondary factors must influence the outcomes of competition between *R. vinicolor* and *R. vesiculosus*. We detected significantly greater root density in the upper soil stratum than the lower soil stratum of the field site used in this study. Differential root density between soil strata is a form of resource heterogeneity and the ability of a fungus to forage over long distances can increase its ability to compete for root tips when the density of root tips is low (Peay et al. 2011). In constrast, EM fungi which produce "short-range" or "contact" exploration type mycelia tend to be more concentrated in areas near the soil surface where root density is higher (Baier et al. 2006, Scattolin et al. 2008). The variability of root density between our soil collection localities implies that distribution of *P. menziesii* root tips available for colonization at our field site is highly heterogenous and varies between localized regions of root proliferation or absence. Despite variability between collection localities, each field soil core possessed significantly higher root density in its upper volume (0 –7.5 cm depth) than in its lower volume (7.5 – 15 cm depth) and greater root density in upper field soil was correlated with greater mean moisture content throughout the dry season (Figure 2.2). Under field conditions, *R. vinicolor* is known to preferentially occupy upper soil strata with a range restricted to the upper 10 cm when in competition with *R. vesiculosus* (Beiler et al. 2012). A general preference of EM fungi for growth in upper soil strata is to be expected given that root density and abundance of EM fungal species typically decreases with increasing soil depth and this trend has been

found consistent across many forest types found on multiple continents (Baier et al. 2006, Dickie et al. 2002, Anderson et al. 2014). Previous surveys (Zak 1971, Harvey et al. 1979) have documented that *R. vinicolor* and *R. vesiculosus* EM root tips are most abundant in the upper organic soil stratum of *Pseudotsuga menziesii* forests. Our observations of greater average moisture content and root density in the upper 7.5 cm of field soil suggest that the upper soil layer at our field site is a resource rich and favorable habitat for EM colonization. Because moisture content and root density did not vary significantly in any of our experimental treatments we cannot extend our experimental results to considerations of moisture or root density in the field. However, a competitive strategy that is reliant upon colonization by spores in *R. vinicolor* would be consistent with a competitive realized niche within microsites of greater root density and moisture.

Comparison of experimental and laboratory conditions:

Here we have determined the effects of soil resource heterogeneity and genetic diversity upon competitive interactions between *R. vinicolor* and *R. vesiculosus*. However, other aspects of our experimental findings must be interpreted with caution because our container plant experiments differed from field conditions in several ways. Firstly, our container plants were kept evenly moist throughout the duration of the experiment while our observations found that moisture levels differed significantly between soil depth classes through the season under natural conditions. Secondly, root density was greater in the upper soil depth class than in the lower soil depth class under field conditions and no such trend was demonstrated in our container plants. Finally, soil temperature in natural field soil varies with both depth and season (Page et al. 1982) and

all soil within our plant containers was held at constant temperature throughout the duration of the experiment. Moisture, root density, and temperature are likely to affect competition between *Rhizopogon* species in the wild and these phenomenon have yet to be explored. In addition to incongruencies between environmental conditions in experimental and field settings there are potential biological effects related to the physical properties of the planting containers themselves. We observed a significant trend of greater EM colonization in the upper volume than in the lower volume of many experimental replicates. Because this trend was not correlated with soil type or root density we can best explain the increased success of both species to colonize upper replicate volumes as an experimental effect related to container properties rather than an ecological principle with application to natural settings.

A final consideration to the comparability of our experimental system and the natural system is the young age class of all trees in the experimental setting. *Rhizopogon* species are well documented as disturbance response organisms that are frequent colonizers of tree seedling roots. EM colonization of *P. menziesii* roots by *R. vinicolor* and *R. vesiculosus* occurs at all forest age classes, but is especially abundant in young stands (Twieg et al. 2007). *Rhizopogon* species are one of the most common EM colonizers of both bioassay and naturally recruited seedlings following forest disturbance (Jones et al. 1997, Baar et al. 1999, Luoma et al. 2006). Furthermore, outplanted seedlings in mature stands of compatible EM hosts (Taylor and Bruns, 1999) and container bioassay seedlings planted in soil from such forest stands (Murata et al. 2013) can be heavily colonized by *Rhizopogon* species even when the roots of mature trees showed no evidence of colonization by *Rhizopogon* species. Thus, it is possible that the

rates of EM colonization by *R. vinicolor* and *R. vesiculosus* were affected by the young age of the host trees in our study.

Conclusions:

Though vertical partitioning of resources has been well documented in EM fungal communities, we know very little concerning the mechanisms that drive this phenomenon and how they might affect EM community diversity. In this study, we present experimental evidence indicating that heterogeneity of soil resources maintains community balance between the EM sister species *R. vinicolor* and *R. vesiculosus*. A recent exhaustive survey of fungal taxa within a boreal forest of EM host trees found that niche differentiation based upon soil horizon and understory plant community is ubiquitous between closely related fungal species (Taylor et al. 2014) and it was suggested that this effect is due either to a lack of genetic constraints upon habitat preference or strong selective pressure for character displacement. Our results demonstrate that vertical niche partitioning in *R. vinicolor* and *R. vesiculosus* is in part the result an interaction between soil resource heterogeneity and competitive strategy rather than specialization upon particular soil resource conditions. It follows that vertical niche partitioning amongst EM fungi could be an indirect rather than direct effect of soil heterogeneity and this poses an additional explanation for niche partitioning between closely related species that compete for the same resources. The biological stochasticity introduced by resource heterogeneity may itself be a driving factor in the adaptation and speciation of fungi and this phenomenon could generate fungal diversity at multiple ecological scales. This hypothesis is consistent with a recent global survey of fungal

diversity, which found that despite strong latitudinal trends in fungal patterns of diversity, there was significant variation between sites within regions (Tedersoo et al. 2014). Future studies that investigate resource heterogeneity in the parameters of both degree and scale may yield a greater understanding of how the indirect effects of patchy resources generate and maintain fungal diversity.

REFERENCES:

Agerer, Reinhard. 2001. Exploration Types of Ectomycorrhizae. *Mycorrhiza* 11 (2): 107–14. doi:10.1007/s005720100108.

Anderson, Ian C., David R. Genney, and Ian J. Alexander. 2014. Fine-Scale Diversity and Distribution of Ectomycorrhizal Fungal Mycelium in a Scots Pine Forest. *New Phytologist* 201 (4): 1423–30. doi:10.1111/nph.12637.

Armstrong, RA., McGehee R. 1980. Competitive Exclusion."*The American Naturalist* 115 (2): 151–70.

Baar, J., T. R. Horton, A. M. Kretzer, and T. D. Bruns. 1999. Mycorrhizal Colonization of *Pinus muricata* from Resistant Propagules after a Stand-Replacing Wildfire. *New Phytologist* 143 (2): 409–18. doi:10.1046/j.1469-8137.1999.00452.x.

Baier, Roland, Jan Ingenhaag, Helmut Blaschke, Axel Göttlein, and Reinhard Agerer. 2006. Vertical Distribution of an Ectomycorrhizal Community in Upper Soil Horizons of a Young Norway Spruce (*Picea abies* [L.] Karst.) Stand of the Bavarian Limestone Alps. *Mycorrhiza* 16 (3): 197–206. doi:10.1007/s00572-006-0035-z.

Baxter, James W., and John Dighton. 2001. Ectomycorrhizal Diversity Alters Growth and Nutrient Acquisition of Grey Birch (*Betula populifolia*) Seedlings in Host–symbiont Culture Conditions. *New Phytologist* 152 (1): 139–49. doi:10.1046/j.0028-646x.2001.00245.x.

Beiler KJ, Durall DM, Simard SW, Maxwell SA and Kretzer AM. 2010. Architecture of the wood-wide web: *Rhizopogon spp*. genets link multiple Douglas-fir cohorts *New Phytologist* 185(2) 543-553.

Beiler, Kevin J., Suzanne W. Simard, Valerie LeMay, and Daniel M. Durall. 2012. Vertical Partitioning between Sister Species of *Rhizopogon* Fungi on Mesic and Xeric Sites in an Interior Douglas-Fir Forest. *Molecular Ecology* 21 (24): 6163–74. doi:10.1111/mec.12076.

Bever, James D., Ian A. Dickie, Evelina Facelli, Jose M. Facelli, John Klironomos, Mari Moora, Matthias C. Rillig, William D. Stock, Mark Tibbett, and Martin Zobel. 2010. Rooting Theories of Plant Community Ecology in Microbial Interactions. *Trends in Ecology & Evolution* 25 (8): 468–78. doi:10.1016/j.tree.2010.05.004.

Bruns, Thomas D. 1995. Thoughts on the Processes That Maintain Local Species Diversity of Ectomycorrhizal Fungi. In *The Significance and Regulation of Soil Biodiversity*, edited by H. P. Collins, G. P. Robertson, and M. J. Klug, 63–73. *Developments in Plant and Soil Sciences* 63. Springer Netherlands.

Cripps, C. L. 2004. Fungi in Forest Ecosystems: Systematics, Diversity, and Ecology. *Memoirs of the New York Botanical Garden Vol. 89*, xvi + 364 pp.

Dickie, Ian A., Bing Xu, and Roger T. Koide. 2002. Vertical Niche Differentiation of Ectomycorrhizal Hyphae in Soil as Shown by T-RFLP Analysis. *New Phytologist* 156 (3): 527–35. doi:10.1046/j.1469-8137.2002.00535.x.

Dunham, Susie M., Alija Bajro Mujic, Joseph W. Spatafora, and Annette M. Kretzer. 2013. Within-Population Genetic Structure Differs between Two Sympatric Sister-Species of Ectomycorrhizal Fungi, *Rhizopogon vinicolor* and *R. vesiculosus*. *Mycologia* 105 (4): 814–26. doi:10.3852/12-265.

Gardes M, Bruns TD. 1993. ITS primers with enhanced specificity for basidiomycetes — application to the identification of mycorrhizae and rusts. *Molecular Ecology* 2: 113–118.

Harvey, A. E., M. J. Larsen, and M. F. Jurgensen. 1979. Comparative Distribution of Ectomycorrhizae in Soils of Three Western Montana Forest Habitat Types. *Forest Science* 25(2):350–358

Izzo, Antonio D., Marc Meyer, James M. Trappe, Malcolm North, and Thomas D. Bruns. 2005. Hypogeous Ectomycorrhizal Fungal Species on Roots and in Small Mammal Diet in a Mixed-Conifer Forest. *Forest Science* 51(3)243-254

Jones MD, Durall DM, Harniman SMK, Classen DC, Simard SW. 1997. Ectomycorrhizal diversity on *Betula papyrifera* and *Pseudotsuga menziesii* seedlings grown inthe greenhouse or outplanted in single-species and mixed plots in southern British Columbia. *Canadian Journal of Forest Research* 27:1872–1889.

Kennedy, Peter G., and Thomas D. Bruns. 2005. Priority Effects Determine the Outcome of Ectomycorrhizal Competition between Two *Rhizopogon* Species Colonizing *Pinus muricata* Seedlings. *New Phytologist* 166 (2): 631–38. doi:10.1111/j.1469-8137.2005.01355.x.

Kennedy, Peter G., Sarah E. Bergemann, Sara Hortal, and Thomas D. Bruns. 2007. Determining the Outcome of Field-Based Competition between Two *Rhizopogon* Species Using Real-Time PCR. *Molecular Ecology* 16 (4): 881–90. doi:10.1111/j.1365-294X.2006.03191.x.

Kennedy, Peter G., Kabir G. Peay, and Thomas D. Bruns. 2009. Root Tip Competition among Ectomycorrhizal Fungi: Are Priority Effects a Rule or an Exception? *Ecology* 90 (8): 2098–2107. doi:10.1890/08-1291.1.

Koide, Roger T., Bing Xu, Jori Sharda, Ylva Lekberg, and Nancy Ostiguy. 2005. Evidence of Species Interactions within an Ectomycorrhizal Fungal Community. *New Phytologist* 165 (1): 305–16. doi:10.1111/j.1469-8137.2004.01216.x.

Kotliar, Natasha B., and John A. Wiens. 1990. Multiple Scales of Patchiness and Patch Structure: A Hierarchical Framework for the Study of Heterogeneity. *Oikos* 59 (2): 253–60. doi:10.2307/3545542.

Kretzer, Annette M., Randy Molina, and Joseph W. Spatafora. 2000. Microsatellite Markers for the Ectomycorrhizal Basidiomycete *Rhizopogon vinicolor*. *Molecular Ecology* 9 (8): 1190–91. doi:10.1046/j.1365-294x.2000.00954-12.x.

Kretzer AM, Dunham S, Molina R, Spatafora JW. 2003. Microsatellite markers reveal the below ground distribution of genets in two species of *Rhizopogon* forming tuberculate ectomycorrhizas on Douglas fir. *New Phytologist* 161(1) 313-320

Kretzer AM, Dunham S, Molina R, Spatafora JW. 2005. Patterns of vegetative growth and gene flow in *Rhizopogon vinicolor* and *R. vesiculosus* (Boletales, Basidiomycota) *Molecular Ecology* 14(8) 2259-2268

Luoma D, Frenkel RE, Trappe JM. 1991. Fruiting of hypogeous sporocarps in Oregon Douglas-fir forests: seasonal and habitat variation. *Mycologia* 83:335-353.

Luoma, Daniel L, Christopher A Stockdale, Randy Molina, and Joyce L Eberhart. 2006. The Spatial Influence of *Pseudotsuga menziesii* Retention Trees on Ectomycorrhiza Diversity. *Canadian Journal of Forest Research* 36 (10): 2561–73. doi:10.1139/x06-143.

McKane, Robert B., Loretta C. Johnson, Gaius R. Shaver, Knute J. Nadelhoffer, Edward B. Rastetter, Brian Fry, Anne E. Giblin, et al. 2002. Resource-Based Niches Provide a Basis for Plant Species Diversity and Dominance in Arctic Tundra. *Nature* 415 (6867): 68–71. doi:10.1038/415068a.

Molina, R., J. M. Trappe, L. C. Grubisha, and J. W. Spatafora. 1999. *Rhizopogon*. In *Ectomycorrhizal Fungi Key Genera in Profile*, edited by Dr John W. G. Cairney and Dr Susan M. Chambers, 129–61. Springer Berlin Heidelberg.

Murata, Masao, Akihiko Kinoshita, and Kazuhide Nara. 2013. Revisiting the Host Effect on Ectomycorrhizal Fungal Communities: Implications from Host–fungal Associations in Relict *Pseudotsuga japonica* Forests. *Mycorrhiza* 23 (8): 641–53. doi:10.1007/s00572-013-0504-0.

Nei, Masatoshi. 1978. Estimation of Average Heterozygosity and Genetic Distance from a Small Number of Individuals. *Genetics* 89 (3): 583–90.

Page AL, Miller RH, Keeney DR. Eds. 1982. *Methods of Soil Analysis*. American Society of Agronomy, Inc. Soil Science Society of American, Inc. Madison, Wisconsin USA.

Peay, Kabir G., Peter G. Kennedy, and Thomas D. Bruns. 2011. Rethinking Ectomycorrhizal Succession: Are Root Density and Hyphal Exploration Types Drivers of Spatial and Temporal Zonation? *Fungal Ecology* 4 (3): 233–40. doi:10.1016/j.funeco.2010.09.010.

Read, D. J., and J. Perez-Moreno. 2003. Mycorrhizas and Nutrient Cycling in Ecosystems – a Journey towards Relevance? *New Phytologist* 157 (3): 475–92. doi:10.1046/j.1469-8137.2003.00704.x.

Ricklefs, Robert E. 1977. Environmental Heterogeneity and Plant Species Diversity: A Hypothesis. *The American Naturalist* 111 (978): 376–81.

Rosling, A., R. Landeweert, B. D. Lindahl, K.-H. Larsson, T. W. Kuyper, A. F. S. Taylor, and R. D. Finlay. 2003. Vertical Distribution of Ectomycorrhizal Fungal Taxa in a Podzol Soil Profile. *New Phytologist* 159 (3): 775–83. doi:10.1046/j.1469-8137.2003.00829.x.

Roth, Roland R. 1976. Spatial Heterogeneity and Bird Species Diversity. *Ecology* 57 (4): 773–82. doi:10.2307/1936190.

Scattolin, Linda, Lucio Montecchio, Elena Mosca, and Reinhard Agerer. 2008. Vertical Distribution of the Ectomycorrhizal Community in the Top Soil of Norway Spruce Stands. *European Journal of Forest Research* 127 (5): 347–57. doi:10.1007/s10342-008-0209-7.

Schoener, TW. 1974. Resource Partitioning in Ecological Communities. *Science* 185 (4145): 27–39. doi:10.1126/science.185.4145.27.

Simard, Suzanne W. 2009. The Foundational Role of Mycorrhizal Networks in Self-Organization of Interior Douglas-Fir Forests. *Forest Ecology and Management*, IUFRO Conference on Biodiversity in Forest Ecosystems and Landscapes, 258, Supplement (December): S95–107. doi:10.1016/j.foreco.2009.05.001.

Smith, AH., and Zeller SM. 1966. Preliminary Account of the North American Species *of Rhizopogon. Memoirs of the New York Botanical Garden* 14(2):177 pgs

Smith JE, Moina R, Huso MMP, Luoma DL, McKay D, Castellano MA, Lebel T,

Valachuvic Y. 2002. Species richness, abundance, and composition of hypogeous and epigeous ectomycorrhizal fungal sporocarps in young, rotation-age, and old-growth stands of Douglas-fir (*Pseudotsuga menziesii*) in the Cascade range of Oregon, U.S.A. *Canadian Journal of Botany* 80:186-204

Smith, Sally E., and David Read. 2008. *Mycorrhizal Symbiosis (Third Edition)* London: Academic Press.

Taylor, D. L., and T. D. Bruns. 1999. Community Structure of Ectomycorrhizal Fungi in a *Pinus muricata* Forest: Minimal Overlap between the Mature Forest and Resistant Propagule Communities. *Molecular Ecology* 8 (11): 1837–50. doi:10.1046/j.1365-294x.1999.00773.x.

Taylor, D. Lee, Teresa N. Hollingsworth, Jack W. McFarland, Niall J. Lennon, Chad Nusbaum, and Roger W. Ruess. 2013. A First Comprehensive Census of Fungi in Soil Reveals Both Hyperdiversity and Fine-Scale Niche Partitioning. *Ecological Monographs* 84 (1): 3–20. doi:10.1890/12-1693.1.

Tedersoo, Leho, Urmas Kõljalg, Nils Hallenberg, and Karl-Henrik Larsson. 2003. Fine Scale Distribution of Ectomycorrhizal Fungi and Roots across Substrate Layers Including Coarse Woody Debris in a Mixed Forest. *New Phytologist* 159 (1): 153–65. doi:10.1046/j.1469-8137.2003.00792.x.

Tedersoo, Leho, Mohammad Bahram, Sergei Põlme, Urmas Kõljalg, Nourou S. Yorou, Ravi Wijesundera, Luis Villarreal Ruiz, et al. 2014. Global Diversity and Geography of Soil Fungi. *Science* 346 (6213): 1256688. doi:10.1126/science.1256688.

Trappe, James M., and Robert F. Strand. 1969. Mycorrhizal Deficiency in a Douglas-Fir Region Nursery. *Forest Science* 15(4):381–389

Twieg B, Durall DM, Simard SW. 2007. Ectomycorrhizal fungal succession in mixed temperate forests. *New Phytologist* **176**: 437–447.

Van Der Heijden MGA, Sanders IR, ***eds.*** 2002. *Mycorrhizal Ecology. Ecological studies* 157. Berlin, Germany: Springer Verlag

White TJ, Bruns TD, Lee SW, Taylor JW. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis MA, Gelfand DH, Sninsky JJ, White TJ, (Eds.), *PCR Protocols--a Guide to Methods and Applications.* Academic Press, San Diego, CA. Pp 315–322.

Zak, B. 1971. Characterization and Classification of Mycorrhizae of Douglas Fir. II. *Pseudotsuga menziesii* + *Rhizopogon vinicolor*. *Canadian Journal of Botany* 49 (7): 1079–84. doi:10.1139/b71-154.

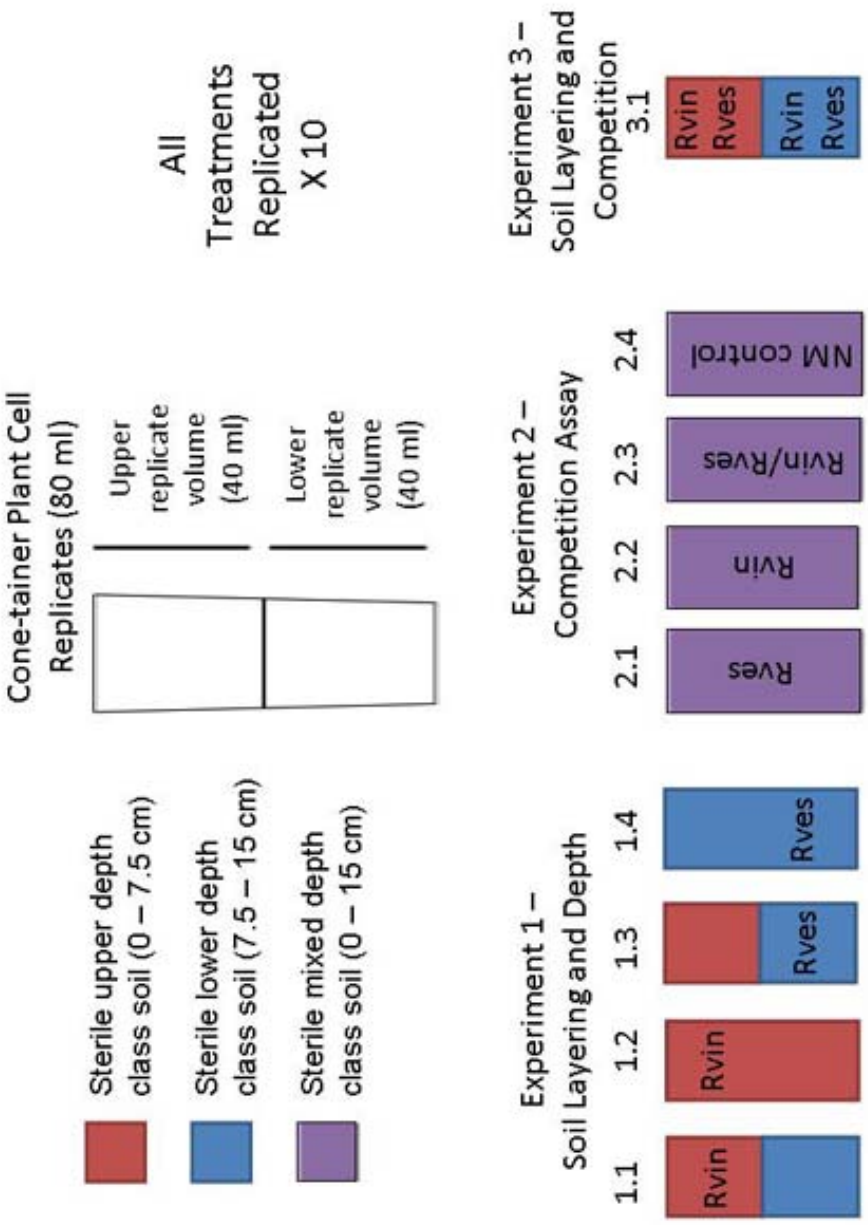


Figure 2.1: A diagrammatic representation of experimental treatments and replicate design. Three experiments are diagrammed and each treatment within an experiment is depicted showing the soil type contained within its upper and lower replicate volume. Each treatment is numbered by the superscript number above its diagram in the format (experiment).(treatment). Each treatment is internally labeled with the species inoculated within the replicate volume where spores were inoculated. Rvin = *Rhizopogon vinicolor*, Rves = *Rhizopogon vesiculosus*, NM = non-mycorrhizal

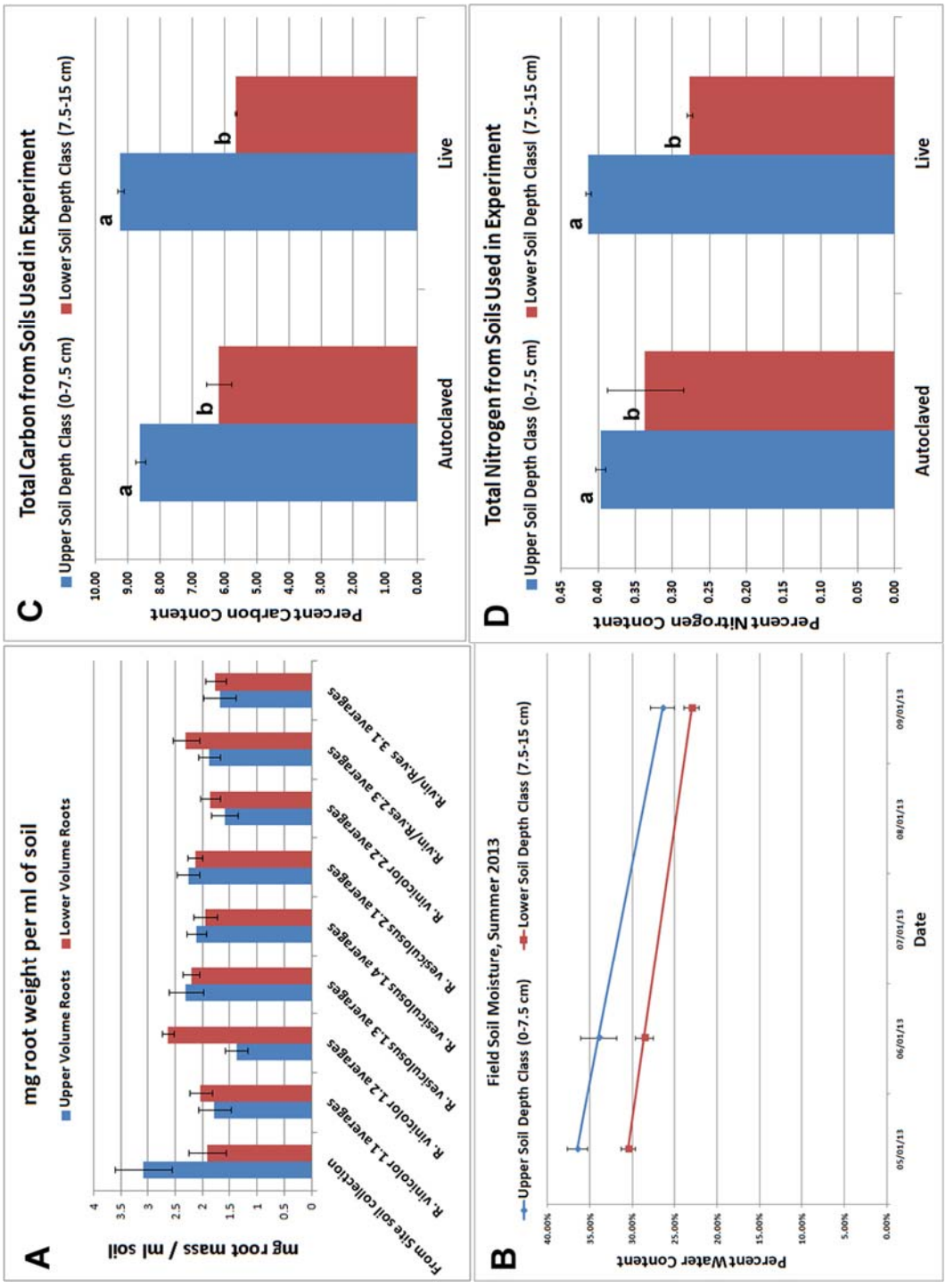


Figure 2.2: Environmental parameters: A) Root density in field and experiment, B) Soil moistureof field soil, C) C values of autoclaved and live field soils, D) N values of autoclaved and live field soils. Error bars depict standard error. Superscript letters above values indicate statistical similarity groups.

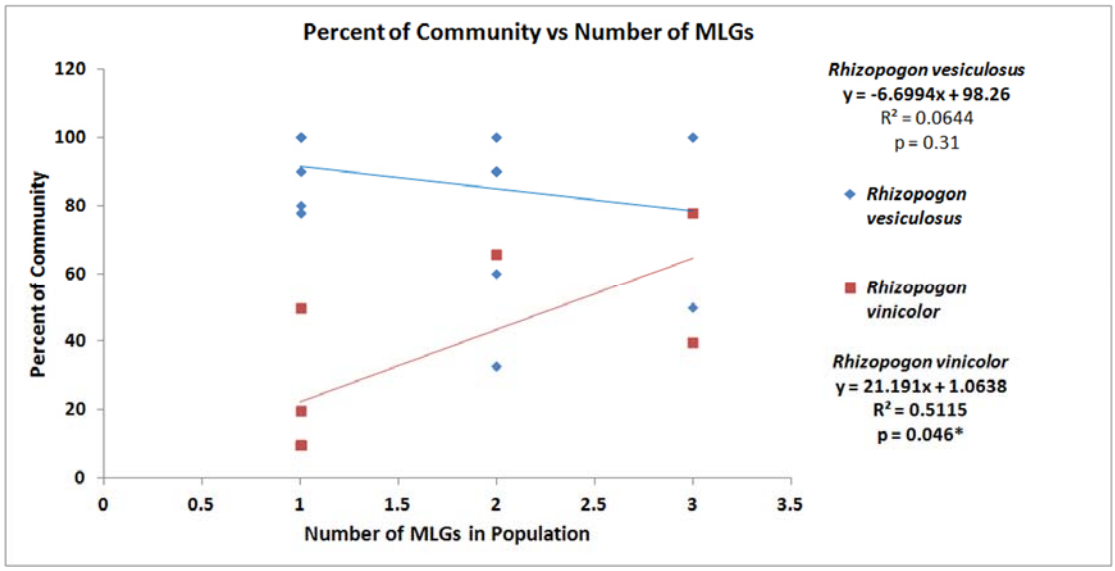


Figure 2.3: Linear regression showing the correlation between the number of individuals (MLGs) present in a replicate and the percent of the community that is represented by that species in that replicate.

Table 2.1: Species identity of all EM root tips harvested from experimental treatments

Species Identity of Ectomycorrhizal Root Tips

Single Species Treatments

Experiment 1 - Stratified Soils

Treatments 1.1 & 1.2 (*R. vinicolor*)	# *R. vinicolor* tips	# *R. vesiculosus* tips
1.1_T - ALL REPLICATES	14	0
1.1_B - ALL REPLICATES	14	0
1.2_T - ALL REPLICATES	18	0
1.2_B - ALL REPLICATES	18	0
totals	**64**	**0**
percent of all sampled root t	**100.00%**	**0.00%**

Experiment 2 - Mixed Soils

Treatment 2.1 (*R. vesiculosus*)	# *R. vinicolor* tips	# *R. vesiculosus* tips
2.1_T - ALL REPLICATES	0	16
2.1_B - ALL REPLICATES	0	16
totals	**0**	**32**
% of all root tips	**0.00%**	**100.00%**

Treatments 1.1 & 1.2 (*R. vesiculosus*)	# *R. vinicolor* tips	# *R. vesiculosus* tips
1.3_T - ALL REPLICATES	0	16
1.3_B - ALL REPLICATES	0	16
1.4_T - ALL REPLICATES	0	20
1.4_B - ALL REPLICATES	0	20
totals	**0**	**72**
% of all root tips	**0.00%**	**100.00%**

Treatment 2.2 (*R. vinicolor*)	# *R. vinicolor* tips	# *R. vesiculosus* tips
2.2_T - ALL REPLICATES	12	0
2.2_B - ALL REPLICATES	12	0
totals	**24**	**0**
% of all root tips	**100.00%**	**0.00%**

Two Species Treatments

Experiment 2 - Mixed Soils

Treatment replicate	# *R. vinicolor* tips	# *R. vesiculosus* tips
2.3_1T	0	10
2.3_1B	0	10
2.3_3T	0	10
2.3_3B	0	10
2.3_4T	0	10
2.3_4B	0	10
2.3_5T	0	10
2.3_5B	0	10
2.3_6T	0	10
2.3_6B	0	10
2.3_7T	0	10
2.3_7B	0	10
2.3_8T	0	10
2.3_8B	0	10
2.3_9T	0	10
2.3_9B	0	10
2.3_10T	0	10
2.3_10B	0	10
totals	**0**	**180**
% of all root tips	**0.00%**	**100.00%**
Tips in upper replicate vol.	**0**	**90**
Tips in lower replicate vol.	**0**	**90**

Experiment 3 - Stratified Soils

Treatment replicate	# *R. vinicolor* tips	# *R. vesiculosus* tips
3.1_1T	0	3
3.1_1B	0	10
3.1_2T	0	10
3.1_2B	5	5
3.1_3T	0	10
3.1_3B	0	10
3.1_4T	0	10
3.1_4B	0	9
3.1_5T	1	9
3.1_5B	1	9
3.1_6T	4	6
3.1_6B	7	2
3.1_7T	6	3
3.1_7B	1	9
3.1_8T	2	8
3.1_8B	0	10
3.1_9T	0	10
3.1_9B	1	8
totals	**28**	**141**
% of all root tips	**16.57%**	**83.43%**
Tips in upper replicate vol.	**13**	**69**
Tips in lower replicate vol.	**15**	**72**

Table 2.2: A table showing the allelic diversity and other statistics by locus for both the sporocarps and the EM root tips. 1-D = Simpson's index, Hexp = Nei's Unbiased Genotypic diversity (an analog of expected heterozygosity) (Nei 1978)

Rhizopogon vesiculosus **(F1, root tips)**

locus	allele	1-D	Hexp	Evenness
2.1	2	0.4	0.8	0.83
2.14	1			
2	1			
1.21	2	0.36	0.72	0.77
1.34	1			
2.44	2	0.5	1	1
mean	**1.5**	**0.21**	**0.84**	**0.87**

Rhizopogon vinicolor (F1, root tips)

locus	allele	1-D	Hexp	Evenness
15	3	0.57	0.86	0.88
46	3	0.66	0.99	0.99
53	2	0.49	0.98	0.98
3.21	2	0.26	0.52	0.65
2.77	2	0.49	0.97	0.97
mean	**2.4**	**0.49**	**0.86**	**0.9**

Rhizopogon vesiculosus **(Parent, sporocarps)**

locus	allele	1-D	Hexp	Evenness
2.1	2	0.47	0.94	0.94
2.14	2	0.22	0.44	0.61
2	2	0.22	0.44	0.61
1.21	3	0.41	0.61	0.63
1.34	2	0.22	0.44	0.61
2.44	2	0.47	0.94	0.94
mean	**2.17**	**0.33**	**0.63**	**0.72**

Rhizopogon vinicolor **(Parent, sporocarps)**

locus	allele	1-D	Hexp	Evenness
15	3	0.41	0.61	0.63
46	3	0.53	0.8	0.78
53	4	0.66	0.88	0.81
3.21	2	0.22	0.44	0.61
2.77	2	0.22	0.44	0.61
mean	**2.8**	**0.41**	**0.63**	**0.69**

Table 2.3: Number of multilocus genotypes (MLGs) by treatment, the total number MLGs per species, and the number of replicates each MLG occurs within. Hexp = Nei's Unbiased Genotypic diversity (an analog of expected heterozygosity) (Nei 1979), E.5 = Evenness, E.5 (Grunwald et al. 2003), Ia = Index of Association, rbarD = adjusted Index of Association. The mating compatibility systems of *R. vinicolor* and *R. vesiculosus* are unknown but their closest relative to have a determined mating system (*Rhizopogon rubescens*) is demonstrated as bipolar (Kawai et al. 2008). The probability of observing (P(obs)) the sampled multilocus genotypes of both species are given for an uncorrected homothallic (all self crosses 100% compatible), corrected bipolar (self crosses 50% compatible), and corrected tetrapolar (self crosses 25% compatible).

Rhizopogon vesiculosus

By Treatment

Treatment	N	MLG	Hexp	E.5	rbarD	% of community
2.1_2	7	3	0.714286	0.912771516	0.321583836	100
2.1_3	10	4	0.711111	0.802078543	-0.069673301	100
2.1_4	9	4	0.694444	0.748368959	0.222222222	100
2.1_5	10	4	0.777778	0.898778719	0.027434884	100
2.1_6	9	1	0	NA	NA	100
2.1_7	10	4	0.822222	0.974622302	0.085645932	100
2.1_8	9	2	0.222222	0.589715507	NA	100
2.1_10	10	3	0.511111	0.692792356	0.607957292	100
3.1_1	8	2	0.535714	0.940856087	NA	100
3.1_2	10	5	0.822222	0.848900591	-0.109739537	100
3.1_3	10	3	0.644444	0.88039073	0.731386854	100
3.1_4	10	3	0.511111	0.692792356	0.576132567	75
3.1_5	9	3	0.666667	0.895275499	0.252810403	90
3.1_6	7	3	0.666667	0.8330782	-0.180906807	69
3.1_7	8	2	0.428571	0.794949052	1	45
3.1_8	10	1	0	NA	NA	90
3.1_9	10	2	0.2	0.57142984	NA	95
Total	**156**	**10**	**0.805376**	**0.769936604**	**0.196089818**	

By Multilocus Genotype

MLG	genotype	P(obs) (homothallic)	P(obs) (bipolar)	P(obs) (tetrapolar)	# Tips sampled	# Replicates
ves_MLG1	452266331122	0.2500	0.1875	0.1250	15	4
ves_MLG2	452266331112	0.0833	0.1250	0.1667	53	11
ves_MLG3	452266231112	0.0833	0.1250	0.1667	17	7
ves_MLG4	442266331112	0.0833	0.1250	0.1667	24	7
ves_MLG5	442266331111	0.0417	0.0313	0.0208	2	1
ves_MLG6	442266231112	0.0833	0.1250	0.1667	29	9
ves_MLG7	442266231111	0.0833	0.0625	0.0417	4	3
ves_MLG8	442266221111	0.0417	0.0313	0.0208	10	2
				Total	**154**	

Table 2.3: Continued

Rhizopogon vinicolor

By Treatment

Treatment	N	MLG	Hexp	E.5	rbarD	% of community
2.2_1	9	6	0.888889	0.86556543	0.129275099	100
2.2_5	9	2	0.388889	0.756410006	1	100
2.2_6	9	4	0.805556	0.930847293	0.653859383	100
2.2_7	8	6	0.892857	0.830815525	0.171437984	100
2.2_9	10	1	0	NA	NA	100
2.2_10	8	3	0.464286	0.629576252	0.5	100
3.1_1	0	0				0
3.1_2	0	0				0
3.1_3	0	0				0
3.1_4	5	1	0	NA	NA	25
3.1_5	1	1	NA	NA	NA	10
3.1_6	9	6	0.888889	0.86556543	0.212044267	59
3.1_7	8	3	0.464286	0.629576252	0.936818839	38
3.1_8	2	1	0	NA	NA	10
3.1_9	1	1	NA	NA	NA	5
Total	**79**	**29**	**0.934761**	**0.664551186**	**0.055528253**	

By Multilocus Genotype

MLG	genotype	P(obs) (homothallic)	P(obs) (bipolar)	P(obs) (tetrapolar)	# Tips sampled	# Replicates
vin_MLG1	6644442224	0.00128082	0.000849	0.000507357	3	1
vin_MLG2	6622442224	0.00128082	0.000849	0.000507357	3	1
vin_MLG3	6622342224	0.002561639	0.001699	0.001014713	1	1
vin_MLG4	6622332224	0.00128082	0.000849	0.002029427	1	1
vin_MLG5	4614342224	0.00128082	0.001699	0.002029427	1	1
vin_MLG6	4614332224	0.00128082	0.001699	0.002029427	3	1
vin_MLG7	1644342224	0.005123279	0.003397	0.002029427	15	2
vin_MLG8	1624442224	0.005123279	0.003397	0.002029427	1	1
vin_MLG9	1624342224	0.012808197	0.010191	0.008117707	2	2
vin_MLG10	1624332224	0.007684918	0.006794	0.00608828	1	1
vin_MLG11	1622442224	0.002561639	0.001699	0.001014713	3	1
vin_MLG12	1614342324	0.00128082	0.001699	0.002029427	1	1
vin_MLG13	1614342222	0.003842459	0.005096	0.00608828	1	1
vin_MLG14	1614332222	0.003842459	0.005096	0.00608828	2	1
vin_MLG15	1612342324	0.00128082	0.001699	0.000507357	6	1
vin_MLG16	1612342322	0.00128082	0.001699	0.000507357	8	2
vin_MLG17	1612342224	0.003842459	0.005096	0.00608828	7	3
vin_MLG18	1612332222	0.003842459	0.005096	0.00608828	1	1
vin_MLG19	1414342322	0.00128082	0.001699	0.002029427	1	1
vin_MLG20	1412342324	0.00128082	0.001699	0.002029427	1	1
vin_MLG21	1412342224	0.00128082	0.001699	0.002029427	1	1
vin_MLG22	1412332324	0.00128082	0.001699	0.002029427	6	1
vin_MLG23	1144332224	0.00128082	0.000849	0.002029427	2	1
vin_MLG24	1122332224	0.003842459	0.005096	0.00608828	2	1
vin_MLG25	1114342222	0.003842459	0.005096	0.00608828	1	1
vin_MLG26	1112342224	0.003842459	0.005096	0.00608828	2	1
vin_MLG27	1112332324	0.00128082	0.001699	0.000507357	1	1
vin_MLG28	1112332224	0.003842459	0.005096	0.00608828	1	1

Supplementary Table 2.S1: All EM colonization measurements made in this study**.**

Species: *R. vinicolor*

Experiment.Treatment_ Replicate	T-tips	B-tips	T-roots	B-roots	Percent Colonization (T)	Percent Colonization (B)
1.1_1	0.0263	0.002	0.0379	0.0645	40.97	3.01
1.1_2	0.0118	0.0197	0.0136	0.0725	46.46	21.37
1.1_6	0.0117	0.0058	0.1122	0.1213	9.44	4.56
1.1_7	0.0225	0.0177	0.0322	0.0547	41.13	24.45
1.1_8	0.0166	0.0176	0.452	0.0571	3.54	23.56
1.1_9	0.0399	0.0086	0.0528	0.0555	43.04	13.42
1.1_10	0.0291	0.0048	0.0465	0.0694	38.49	6.47
1.1 averages	**0.0225571**	**0.010886**	**0.106743**	**0.070714**	**31.87**	**13.83**

Species: *R. vinicolor*

Experiment.Treatment_ Replicate	T-tips	B-tips	T-roots	B-roots	Percent Colonization (T)	Percent Colonization (B)
1.2_1	0.0047	0.0205	0.0638	0.0656	6.86	23.81
1.2_2	0.0007	0.0062	0.0612	0.0924	1.13	6.29
1.2_3	0.0217	0.0259	0.0718	0.0792	23.21	24.64
1.2_5	0.0428	0.0458	0.0262	0.0541	62.03	45.85
1.2_6	0.0098	0.0298	0.0355	0.0922	21.63	24.43
1.2_7	0.0005	0.0043	0.0538	0.097	0.92	4.24
1.2_8	0.0101	0.0243	0.0184	0.0795	35.44	23.41
1.2_9	0.0162	0.0059	0.0216	0	42.86	100.00
1.2_10	0.0071	0.0356	0.0128	0.0913	35.68	28.05
1.2 averages	**0.0126222**	**0.022033**	**0.040567**	**0.072367**	**25.53**	**31.19**

Species: *R. vesiculosus*

Experiment.Treatment_ Replicate	T-tips	B-tips	T-roots	B-roots	Percent Colonization (T)	Percent Colonization (B)
1.3_2	0.0361	0.0197	0.0496	0.0671	42.12	22.70
1.3_3	0.0632	0.0054	0.0874	0.053	41.97	9.25
1.3_4	0.0485	0.0137	0.0352	0.0627	57.95	17.93
1.3_6	0.0431	0.0138	0.0932	0.0815	31.62	14.48
1.3_7	0	0.0308	0.0701	0.0566	0.00	35.24
1.3_8	0.0189	0.0161	0.0463	0.0747	28.99	17.73
1.3_9	0.0024	0.0178	0.0426	0.0789	5.33	18.41
1.3_10	0.0115	0.0161	0.0893	0.0996	11.41	13.92
1.3 averages	**0.0279625**	**0.016675**	**0.064213**	**0.071763**	**27.42**	**18.71**

Species: *R. vesiculosus*

Experiment.Treatment_ Replicate	T-tips	B-tips	T-roots	B-roots	Percent Colonization (T)	Percent Colonization (B)
1.4_1	0.0053	0	0.0371	0.0475	12.50	0.00
1.4_2	0.0267	0.0252	0.0451	0.0947	37.19	21.02
1.4_3	0.0592	0.021	0.0597	0.036	49.79	36.84
1.4_4	0.0252	0.013	0.056	0.0621	31.03	17.31
1.4_5	0.01	0.0004	0.1029	0.0679	8.86	0.59
1.4_6	0	0.0125	0.0656	0.0613	0.00	16.94
1.4_7	0.0145	0.0157	0.074	0.0801	16.38	16.39
1.4_8	0.0465	0.0139	0.0348	0.0359	57.20	27.91
1.4_9	0.0356	0.0143	0.0622	0.1131	36.40	11.22
1.4_10	0.0244	0.0136	0.0593	0.0525	29.15	20.57
1.4 averages	**0.02474**	**0.01296**	**0.05967**	**0.06511**	**27.85**	**16.88**

Supplementary Table 2.S1: Continued

Species: *R. vesiculosus*

Experiment.Treatment_ Replicate	T-tips	B-tips	T-roots	B-roots	Percent Colonization (T)	Percent Colonization (B)
2.1_2	0.0019	0.0006	0.1124	0.0978	1.66	0.61
2.1_3	0.0345	0.0159	0.0433	0.0446	44.34	26.28
2.1_4	0.0198	0.0115	0.0427	0.0765	31.68	13.07
2.1_5	0.0288	0.0195	0.0381	0.0779	43.05	20.02
2.1_6	0.0368	0.0582	0.0438	0.049	45.66	54.29
2.1_7	0.0345	0.0113	0.0745	0.0727	31.65	13.45
2.1_8	0.0413	0.0076	0.0829	0.0601	33.25	11.23
2.1_10	0.0421	0.0189	0.0473	0.0619	47.09	23.39
2.1 averages	**0.0299625**	**0.017938**	**0.060625**	**0.067563**	**34.80**	**20.29**

Species: *R. vinicolor*

Experiment.Treatment_ Replicate	T-tips	B-tips	T-roots	B-roots	Percent Colonization (T)	Percent Colonization (B)
2.2_1	0.0197	0.0232	0.0268	0.0642	42.37	26.54
2.2_5	0.0127	0.0031	0.0486	0.0386	20.72	7.43
2.2_6	0.0475	0.0135	0.0488	0.0534	49.33	20.18
2.2_7	0.0257	0.019	0.0616	0.0667	29.44	22.17
2.2_9	0.0252	0.0275	0.0296	0.0545	45.99	33.54
2.2_10	0.0167	0.034	0.0197	0.0472	45.88	41.87
2.2 averages	**0.0245833**	**0.02005**	**0.039183**	**0.0541**	**38.95**	**25.29**

Species: *R. vinicolor / R. vesiculosus*

Experiment.Treatment_ Replicate	T-tips	B-tips	T-roots	B-roots	Percent Colonization (T)	Percent Colonization (B)
2.3_1	0.0291	0.0323	0.0237	0.0513	55.11	38.64
2.3_3	0.0227	0.0468	0.0314	0.0659	41.96	41.53
2.3_4	0.0154	0.0196	0.0504	0.0281	23.40	41.09
2.3_5	0.0305	0.0233	0.0244	0.0739	55.56	23.97
2.3_6	0.0598	0.0168	0.0332	0.0706	64.30	19.22
2.3_7	0.0128	0.0014	0.0482	0.0462	20.98	2.94
2.3_8	0.0486	0.0347	0.0724	0.0652	40.17	34.73
2.3_9	0.0544	0.0544	0.0203	0.0839	72.82	39.33
2.3_10	0.038	0.0536	0.0591	0.0612	39.13	46.69
2.3 averages	**0.0345889**	**0.031433**	**0.040344**	**0.0607**	**45.94**	**32.02**

Species: *R. vinicolor / R. vesiculosus*

Experiment.Treatment_ Replicate	T-tips	B-tips	T-roots	B-roots	Percent Colonization (T)	Percent Colonization (B)
3.1_1	**0.0005**	0.0087	0.0638	0.098	0.78	8.15
3.1_2	0.0116	0.0039	0.0343	0.0403	25.27	8.82
3.1_3	0.0078	0.0011	0.1297	0.0409	5.67	2.62
3.1_4	0.0137	0.0175	0.0116	0.0375	54.15	31.82
3.1_5	0.0032	0.0074	0.0327	0.0749	8.91	8.99
3.1_6	0.0058	0.0022	0.0798	0.0957	6.78	2.25
3.1_7	0.016	0.0043	0.0393	0.0664	28.93	6.08
3.1_8	0.0245	0.0019	0.0778	0.0606	23.95	3.04
3.1_9	0.0225	0.0111	0.033	0.0626	40.54	15.06
3.1 averages	**0.0117333**	**0.006456**	**0.055778**	**0.0641**	**21.67**	**9.65**

Chapter 3. Comparative mating type genomics of the ectomycorrhizal truffles *Rhizopogon vinicolor* and *Rhizopogon vesiculosus* (Basidiomycota: Boletales)

Alija Bajro Mujic, Joseph W. Spatafora

Target Journal: G3: Genes | Genomes | Genetics

ABSTRACT:

Ectomycorrhizal (EM) fungi pose a particularly challenging system for the study of reproductive biology because many of these fungi cannot easily be cultured under laboratory conditions. To overcome traditional barriers to the study of EM mating systems we have utilized Illumina whole genome shotgun sequencing. Here we present the draft genome sequences of the EM sister species *Rhizopogon vinicolor* Smith and Zeller and *R. vesiculosus* Smith and Zeller (Basidiomycota, Boletales). We have characterized gene content and organization of a 400 kb region surrounding the homeodomain transcription factor genes of the mating type A-locus and a 80 kb region surrounding the pheromone receptor and pheromone precursor genes of the mating type B-locus. We found that both *R. vinicolor* and *R. vesiculosus* possess a single pair of homeodomain transcription factor homologs. Both *Rhizopogon* species contain homologs of the pheromone receptor and pheromone precursor genes but *R. vinicolor* possesses a more diverse set of these homologs. Comparison of *Rhizopogon* genomes with available reference genomes from the Boletales, Agaricales, and Polyporales revealed broad synteny of the A-locus region within the Boletales but several genomic rearrangements across orders. Our analysis of the B-locus region of Boletales genomes found a correlation between genetic diversity at the B-locus region and mating type with tetrapolar species possessing more diverse gene content than bipolar species. Genome assemblies for *Rhizopogon* species were produced from dikaryotic tissue and an analysis of haplotypic SNPs within each genome revealed a greater degree of heterozygosity in *R. vinicolor*. The *Rhizopogon* genomes presented here are the first genomes available for Basidiomycota truffles and we demonstrate a novel technique for characterizing genome

scale heterozygosity using genomes assembled from a combination of genetically distinct haplotypes.

INTRODUCTION:

The mating event of fungi in Basidiomycota involves the anastomosis of monokaryons and the resultant formation of a dikaryon (Heitman et al. 2013). Because the spores of most species of ectomycorrhizal (EM) fungi cannot be germinated and maintained as monokaryons in culture, they pose a challenging system for the study of fungal reproductive biology. Some progress has been made with cultivable EM Agaricomycetes and those species that have been studied can generally be categorized into one of three mating types: secondary homothallic (a form of selfing), as in *Laccaria altaica* (Fries and Mueller 1984); heterothallic bipolar (50% rate of outcrossing), as in *Suillus luteus*, *S. granulatus,* and *Rhizopogon roseolus* (Fries and Neumann 1990, Kawai et al. 2008); and heterothallic tetrapolar (75% rate of outcrossing), as in *Hebeloma cylindrosporum* (Debaud and Gay 1987). This range of mating systems encompasses the major forms of genetic outcrossing systems known in fungi of Basidiomycota (Heitman et al. 2013).

EM Agaricomycetes represent a diverse group of fungi that have independently developed the EM habit at least six to eight times (Hibbett and Matheny 2009). Such a diverse group of fungi would be expected to display differential evolutionary histories and population structure between species. Because of the lack of knowledge concerning the mating systems of EM fungi much of the previous work in fungal population genetics has focused upon functional and ecological traits of fungi and left concerns of

reproductive biology largely neglected (Douhan et al. 2011). The recent development of next generation sequencing technologies has allowed for novel approaches to the study of fungal genetics. In this study we leverage these technologies to investigate the mating systems of the EM Agaricomycetes *Rhizopogon vinicolor* and *Rhizopogon vesiculosus* (Boletales, Basidiomycota), which have been the focus of population genetics and ecology studies of EM fungi (Kretzer et al. 2003, Kretzer et al. 2005, Beiler et al. 2010, Beiler et al. 2012, Dunham et al. 2013).

Early studies investigating the population structure of EM fungi were performed with fungi that had been characterized for their mating recognition systems (Dahlberg and Stenlid 1994, De la Bastide, 1994). However, the lack of knowledge concerning the mating systems of most EM Agaricomycetes makes extrapolation of these findings to other EM fungi difficult. Many studies have been able to address EM Agaricomycetes mating systems indirectly by characterizing population genetic structure (Douhan et al. 2011), but a precise understanding of the genetics that governs the mating system of EM fungi and how that relates to population structure remains elusive. In this study we present the draft genomes of the EM Agaricomycetes *Rhizopogon vinicolor* and *Rhizopogon vesiculosus* (Boletales, Basidiomycota). We utilize these genomes to characterize the genomic regions that govern mating recognition in these fungi and draw comparisons between these genomic regions with better characterized mating systems within Agaricomycetes.

Rhizopogon vinicolor and *R. vesiculosus* are EM sister species that are members of Boletales, a larger group of forest fungi. Species of *Rhizopogon* produce hypogeous fruiting bodies, also called false truffles, which achieve spore dispersal when they are

excavated and consumed by mammals (Izzo et al. 2005). Both species share a sympatric distribution in the Pacific Northwest of North America (Chapter 4 of this dissertation) and grow in association with only a single ectomycorrhizal host species, *Pseudotsuga menziesii* (Molina 1994). When both species co-occur in a stand of *Pseudotsuga* they can often be detected at near equal frequencies as both EM root tips and fruiting bodies (Kretzer et al. 2003, Kretzer et al. 2005, Dunham et al. 2013). Despite equal frequency of detection these species display different life histories and population structure at the landscape scale. *R. vesiculosus* produces larger average genets (Kretzer et al. 2003 and 2005, Beiler et al. 2012, Dunham et al. 2013), producing more and larger sporocarps per genet than *R. vinicolor*, and shows patterns of inbreeding within a range of 120 meters (Dunham et al. 2013). The low degree of population structure observed for *R. vesiculosus* at the scale of a single locality extends to the watershed scale with populations of *R. vesiculosus* showing decreased levels of population differentiation than *R. vinicolor* (Kretzer et al. 2005).

Several hypotheses have been proposed to explain the patterns of population structure observed in *Rhizopogon* species. Differential rates of secondary homothallism, i.e., the production of heterokarotic binucleate basidiospores (Horton 2006), is one mechanism, but both species produced binucleate spores at near equal and relatively low rates (1-2%), which are typical of outcrossing EM Agaricomycetes (Dunham et al. 2013). It is possible that *R. vesiculosus* is more likely to mate with close relatives because of its larger genet size and higher production of sporocarps per genet (Dunham et al. 2013) or that *R. vinicolor* is under selective pressure from interspecies competition with *R. vesiculosus* to outcross more readily (Chapter 2 of this dissertation). However, without

further knowledge of the mating system operating in these two fungi it is impossible to conclusively determine the source of observed patterns of population structure in natural populations of these fungi.

Here we report the first genomes of truffle-forming Basidiomycota and characterize the mating type loci of *Rhizopogon* vinicolor and *R. vesiculosus.* Both species were found to possess the homeodomain transcription factors, pheromone receptors and lipopeptide pheromones typical of the A and B loci in other members of Agaricomycetes. *Rhizopogon vinicolor* was found to possess a greater diversity of both pheromone receptor and pheromone precursor genes than *R. vesiculosus*. In addition, *R. vinicolor* possessed multiple paralogs of the gene STE14, which is demonstrated to function in post-translational modification of lipopeptide mating pheromones (Caldwell et al. 1995). The genomic regions surrounding the A-locus homeodomain proteins of both *R. vinicolor* and *R. vesiculosus* were largely syntenous. A class wide comparison of the synteny of genomic regions around the A locus, however, reveals large sections (100kbp – 200kbp) of highly conserved region oriented 5' of homeodomain protein pairs, but that a unique transposition of a large genomic fragment oriented 3' to the homeodomain proteins is a derived state of some members of Agaricales. Organization and gene content of the B-locus in *R. vinicolor* was similar to that of tetrapolar Boletales and *R. vesiculosus* showed similarities to both bipolar and tetrapolar Boletales species. Rates of non-synonymous SNP mutation were used as an estimate of heterozygosity within the individual dikaryon. Results of SNP analyses revealed lower rates of heterozygosity in *R. vesiculosus* consistent with patterns of inbreeding observed in natural populations.

MATERIALS AND METHODS:

Culture conditions and tissue harvest:

Tissue cultures of *R. vinicolor* and *R. vesiculosus* were derived from fresh sporocarps collected during the summer of 2011 from a field site located on Mary's Peak in the Oregon coast range. Coordinates and ecological properties of this site are detailed in Dunham et al. (2013). Sporocarps were cultured under laboratory conditions as follows: Each sporocarp was cleaned of adhering debris using a damp cloth and then divided into two hemispheres using a flame sterilized scalpel. A small section (1 mm^3) of clean dikaryotic tissue from the internal basidiospore bearing tissue (gleba) was then transferred to 60 mm petri dishes of modified Melner Norkranz media (MMN) (CaCl2 * 2(H2O) 0.05g, NaCl 0.025g, MgSO4 * 7(H2O) 0.15g, (NH4)2 * HPO4 0.25g, KH2PO4 0.5g, FeCl2 * 6(H2O) 0.012g, Thiamine HCL 1mg, Glucose 10g, Malt Extract 3g, MES Buffer 4.88g, Agar 20g, H2O to 1L) using a flame sterilized scalpel. Cultures were incubated at 20°C until growth was observed. Several immature (gleba white to yellow in color) and early maturity (gleba light brown) sporocarp of each species were brought into culture and the most vigorous strain of each species was selected for genome sequencing. The sporocarps used to derive cultures were accessioned into the fungal herbarium collections of Oregon State University under the accession numbers OSC # 147973 (*R. vinicolor*, AM-OR11-026) and OSC # 148003 (*R. vesiculosus*, AM-OR11-056). The original basidiome (OSC#148003, Collector's number: AM-OR11-056) used to derive the *R. vesiculosus* culture used in genome sequencing was destructively sampled for

production of spore slurries used in competition bioassays described in chapter 2 and this collection is no longer available.

The two cultures selected for genome sequencing were then transferred to specialized culture conditions to facilitate the harvest of mycelium for DNA and RNA extraction. For DNA extraction small sections (1 mm^3) of *R. vesiculosus* agar cultures were transferred to 100 ml of sterile MMN broth. Cultures were incubated at 25°C for 1 – 4 weeks with short daily applications of stirring to promote oxygenation and colony fragmentation. Once *R. vesiculosus* cultures reached 1 cm diameter fungal tissue was separated from nutrient broth using a Buchner flask and Whatman filter paper. Agar residue from the original inoculation plugs were removed by dissected and remaining fungal tissue was blotted dry between two sheets of filter paper. Harvested tissue was flash frozen in liquid nitrogen, lyophilized, and stored at -80°C until the time of DNA extraction. Culture and harvest of *Rhizopogon vinicolor* for DNA extraction was performed on MMN agar plates topped with presoaked and sterilized cellophane membrane (BioRad, Hercules, CA) and small sections (1 mm^3) of *R. vinicolor* agar culture were transferred to the top of the cellophane membrane. Cellophane MMN cultures were incubated at 20°C for 1-2 weeks and tissue was harvested by directly peeling mycelium from the cellophane membrane. *R. vinicolor* tissue was flash frozen, lyophilized and stored at -80°C as described for *R. vesiculosus.*

Production of tissue for RNA extraction was performed on four distinct culture media to promote a diverse population of RNA transcripts. Tissue production for RNA was performed for both species using solid agar media and cellophane membrane upon four media formulations: Potato Dextrose Agar (PDA) (BD Difco, Franklin Lakes, New

Jersey), standard MMN (formula described above), MMN media omitting MES buffer, and MMN media omitting B-vitamin supplement. Cultures were allowed to grow no longer than 7 days at 20°C from the time of initial growth after transfer. Cultures were then harvested by peeling mycelium from the cellophane membrane, flash frozen in liquid nitrogen, and stored at -80°C without lyophylization. Frozen tissue was stored for no more than 2 months before RNA extraction was performed.

DNA and RNA extraction:

DNA was extracted from lyophilized tissue following a CTAB phenol:chloroform:isoamyl alcohol (P:C:IAA) protocol. Lyophilized tissue was pulverized either by using an autoclaved mortar and pestle or in a 1.5 ml microcentrifuge tube using autoclaved micropestles. All grinding was conducted after chilling dry tissue with liquid nitrogen. Approximately 100 µl of finely ground fungal tissue was suspended in 2% CTAB buffer, incubated at 65°C for 60 minutes and then mixed 1:1 with 25:24:1 P:C:IAA to extract DNA. Samples were centrifuged and the aqueous layer was separated and treated with 10 mg/ml RNAse A (multiple suppliers) for 1 hour at 35°C. RNAse treated samples were extracted a second time by mixing 1:1 with 24:1 C:IAA. These final extracts were centrifuged again and DNA was precipitated from the supernatant by addition of 2 volumes 95% ethanol and 1/10 volume 3M sodium acetate. DNA extractions were quantified using a Qubit flourometer (Life Technologies, Grand Island, NY) and multiple extracts of each isolate were combined and again precipitated with 95% ethanol to produce a single concentrated DNA solution for genome sequencing.

RNA was extracted from flash frozen tissue following a combined TRIzol and RNeasy cleanup protocol. Frozen tissue was thawed directly in TRIzol reagent (Life Technologies, Grand Island, NY) in a lysing matrix D (MP biomedicals, Santa Ana, CA) bead beating tube. Samples were homogenized in a GenoGrinder 2000 device (Spex Sample Prep, Metuchen, NJ) using two rounds of shaking at 1750 RPM for 30 seconds with a cool down period of 1 minute between runs. Homogenized samples were combined with 0.2 volume chloroform and centrifuged. RNA was precipitated from supernatant using 0.5 volumes of isopropyl alcohol and RNA was resuspended in molecular biology grade water. The RNA solution was then cleaned using the RNeasy Mini Kit (Qiagen, LOCATION) following manufacturer protocols including the optional on column DNase treatment. Final RNA extractions were quantified using a Qubit flourometer and quality was accessed using a NanoDrop spectrophotometer (Thermo Scientific, Wilmington, DE). RNA extractions from each of the four growth conditions were combined in equal proportions for each isolate to create a single pooled RNA extraction for transcriptome sequencing.

Genome sequencing and assembly:

The genome and transcriptome of both *R. vinicolor* and *R. vesiculosus* were sequenced using the Illumina HiSeq next generation sequencing platform. Illumina library construction, sequencing, and annotation of the *R. vinicolor* and genome and transcriptome were conducted at the United States Department of Energy Joint Genome Institute (DOE-JGI) (Walnut Creek, CA). Genome and transcriptome assembly and annotation followed standard DOE-JGI procedures (Grigoriev et al. 2006). Briefly, a 270

bp insert DNA library was prepared for sequencing with paired-end 150mers (2X150bp) on an Illumina HiSeq 2000. Genome assembly was performed using VELVET 1.19 (Zerbino and Birney, 2008) and AllPaths-LG (Gnerre et al. 2011) and transcriptome assembly was performed using TRINITY (Grabherr et al. 2011). Gene prediction and annotation of the *R. vinicolor* genome utilized the DOE-JGI gene annotation pipeline (Grigoriev et al. 2006, http://genome.jgi.doe.gov/programs/fungi/FungalGenomeAnnotationSOP.pdf). Assemblies, annotations, and EST data are available for download and review in the DOE-JGI fungal genome portal MycoCosm (Grigoriev et al. 2014).

Genome and transcriptome sequencing, assembly, and annotation of *R. vesiculosus* were performed at Oregon State University. The Illumina DNA library used for genome sequencing was prepared using the NEBnext DNA Master Mix Kit for Illumina (New England Biolabs, Ipswich, MA) following the manufacturer's protocol with size selection for a 420 bp library fragment size performed by agarose gel electrophoresis size selection. An RNASeq library was prepared for use in transcriptome sequencing using the Illumina TruSeq RNA sample preparation kit (Illumina Inc., San Diego, CA) following the manufacturer protocol. All Illumina sequencing of *R. vesiculosus* was performed on an Illumina HiSeq 2000 device at the Oregon State University Center for Genome Research and Biocomputing (CGRB). The genomic DNA library was sequenced on a full flow cell lane using a 2X100bp cycle and Illumina version 2 chemistry. The RNAseq library was sequenced on 1/8 of a flowcell lane using a 1X50bp cycle and Illumina version 3 chemistry. Raw Illumina reads were trimmed and quality filtered using the fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and

custom Perl scripts. De novo genome assembly of DNA reads was performed using VELVET 1.0 (Zerbino and Birney, 2008) and de novo transcriptome assembly was performed using TRINITY. Gene annotation of *R. vesiculosus* was performed using the MAKER pipeline (Cantarel et al. 2008) with reference EST and protein homology data downloaded from the DOE-JGI MycoCosm portal for four closely related members of order Boletales (*Rhizopogon vinicolor*, *Boletus edulis*, *Suillus brevipes*, and *Suillus luteus*) and the aformentioned *R. vesiculosus* transcriptome data.

Synteny analyses:

In order to determine the organization of genetic regions flanking the homeodomain (A-locus) and pheromone receptor/precursor genes (B-locus) of *R. vinicolor* and *R. vesiculosus* we used a multistep reference-guided approach. We first identified contigs in the genome assembly of *R. vinicolor* and *R. vesiculosus* that contained genes with predicted mating function by using the NCBI BLASTP algorithm to search custom BLAST databases generated for the predicted gene models of *R. vinicolor* and *R. vesiculosus.* Genes of known mating type function in other fungi were used as queries in these BLAST searches. The homeodomain transcription factor (HD) genes of that function in mating recognition (A-locus proteins) of Agaricomycetes are consistently found in close proximity to the gene encoding the mitochondrial intermediate pepitidase (MIP) protein (James et al. 2013). Pheromone precursor (P/P) and receptor (P/R) genes, which function in Agaricomycete mating recognition (B-locus proteins), are typically found in close association with one another as functional “cassettes” (Casselton and Olesnicky, 1998). Reference sequences used as queries in A-locus protein BLAST

searches were the homeodomain transcription factor genes of *Rhizopogon roseolus* (Genbank Accession #'s: BAL45602, BAL45603) and MIP gene of *Suillus pictus* (Genbank accession #: AY179596). B-locus protein BLAST searches used pheromone receptor genes of *Coprinopsis cinerea* (Genbank accession #: AAO17256) and *Serpula lacrymans* (Genbank accession #: EGO31061) that are known to function in mating recognition (Stajich et al. 2010, Skrede et al. 2013) as well as a pheromone precursor gene of *C. cinerea* (*C. cinerea* phb3.2, Stajich et al. 2010). Assembly contigs of *R. vinicolor* and *R. vesiculosus* that contained both HD genes and the MIP gene were considered candidates for the mating type A-locus and contigs containing both P/R and P/P genes in close proximity were considered candidates for the mating type B-locus.

Once we identified candidate contigs containing the A and B mating type loci of *Rhizopogon* species we characterized protein content of these contigs and conducted BLAST searches for these proteins to identify homologous genomic regions in reference fungal genomes available on the DOE-JGI MycoCosm. Homologous proteins were identified in reference genomes as the top BLAST hit to *Rhizopogon* proteins with e scores greater than or equal to $1e^{-20}$. Sequence composition of P/P genes is highly divergent within and between taxa and while the transmembrane domains of P/R genes are relatively conserved in the Agaricomycetes, not all fungal species contain homologs of the same P/R genes (James et al. 2006, James et al. 2013). As such BLAST searches were not sufficient to identify all P/P and P/R genes contained within a genome. Thus, we performed manual searches of gene annotation information in the putative mating type regions identified by BLAST searches using the VISTA synteny browser available through the DOE-JGI MycoCosm. We also designed a custom Perl script (Appendix 1)

that uses regular expression pattern matching to search B-locus regions for open reading frames containing terminal –CAAX prenylation motifs that are typical of fungal mating type pheromones (Caldwell et al. 1995, Casselton and Olsenicky 1998). Once all genes in a genomic region were characterized homologous regions were identified in reference genomes as possessing conserved gene synteny with the A and B mating type regions identified in *Rhizopogon* genomes.

Reference genomes were chosen from both bipolar and tetrapolar species and covered as much taxonomic breadth within Agaricomycetes as possible while still retaining enough genomic synteny to allow for alignment. For the A-locus region we chose reference genomes from *Suillus luteus* (DOE-JGI taxon ID: Suilu1)*, Serpula lacrymans* (DOE-JGI taxon ID: SerlaS7_9_2, Eastwood et al. 2011), *Laccaria bicolor* (DOE-JGI taxon ID: Lacbi2, Martin et al. 2008), *Coprinopsis cinerea* (DOE-JGI taxon ID: Copci1, Stajich et al. 2010), and *Phanerochaete chrysosporium* (DOE-JGI taxon ID: Phchr2, Ohm et al. 2014) that are available for download and review in the DOE-JGI MycoCosm. Reference genomes for the B-locus region were the same as those chosen for the A-locus with the exclusion of *Phanerochaete chrysosporium* and the addition of *Paxillus involutus* (DOE-JGI taxon ID: Paxin1), *Suillus brevipes* (DOE-JGI taxon ID: Suibr1) and *Coniophora puteana* (DOE-JGI taxon ID: Conpu1, Floudas et al. 2012). For A-locus annotation a reference region of approximately 400 kilobase pairs centered on the HD genes was chosen from the genome assembly of *Serpula lacrymans*. All predicted proteins falling within this region were identified and subsequently used as BLAST queries against the *Rhizopogon* genomes and other fungal reference genomes. Mapping of protein coordinates and selection of protein sequences was either performed by hand

or with the use of custom Perl scripts (Appendix 1). Annotation of the B-locus was performed similarly to that of the A locus except a region of approximately 80 kilobase was selected for annotation centered around P/P and P/R genes in *S. lacrymans*. *Rhizopogon* genomes sequenced in this study and some of the reference genomes were fragmented into several contigs in the areas surrounding the A and B mating type loci. We conducted genomic alignments of putative A and B mating type locus regions using MAUVE (Darling et al. 2004) to facilitate the reconstruction of fragmented genomes and final synteny maps were visualized using CHROMOMAPPER (Niculita-Hirzel and Hirzel, 2008).

Phylogenetic analyses:

The P/R genes of Agaricomycetes possess sufficient sequence identity in their membrane spanning domains that orthology of proteins from multiple genomes cannot be fully assessed by use of the BLAST algorithm alone. To confirm the identity of P/R homologues we conducted phylogenetic analysis of all putative mating type P/R genes identified in this study along with several reference sequences (http://www.ncbi.nlm.nih.gov/genbank/) of P/R genes known to function in mating recognition in other Agaricomycetes. Alignment of raw amino acid sequence was performed using MUSCLE (Edgar 2004) and phylogenetic analysis was conducted using the maximum likelihood algorithm implemented in RAxML 7.04 (Stamatakis 2006). In the synteny analysis we also discovered several homologs of a putative isoprenyl cysteine methyltransferase (ICMT) gene, which functions in post translational modification of P/P genes. The copy number and placement of these ICMT genes within Boletales genomes

was correlated with mating type but orthology of these genes was unclear. To clarify the relationship of these ICMT gene homologs a phylogenetic analysis was conducted using the methods outlined here.

Analysis of SNPs:

The genomes of *R. vinicolor* and *R. vesiculosus* were sequenced using DNA extracted from dikaryotic fungal tissue and heterozygosity in the form of indels and single nucleotide polymorphisms (SNP) were visible in the assembly pile-up file generated by the VELVET assembler. We used the frequency of synonymous and non-synonymous SNPs detected in our genome assemblies as a measure of heterozygosity within the fungal individual sequenced in each genome. SNPs were identified by creating alignments of trimmed and quality controlled Illumina reads to the consensus sequence of the VELVET de novo assembly using BOWTIE (Langmead et al. 2009). Pileup files were generated from BOWTIE alignments using SAMtools (Li et al. 2009) and final SNP calls were performed using VARSCAN (Koboldt et al. 2009). SNPs were predicted by VARSCAN with a threshold p-value cutoff of 1e-4 with Bonferroni correction. The rate of synonymous versus non-synonymous mutations in each genome was determined by referencing VARSCAN output against predicted gene models using a custom Perl scripts (Appendix 1).

RESULTS:

Genome assembly statistics:

The genome of *R. vinicolor* assembled to a total length of 36.1 Mb with 2310 scaffolds containing 3018 contigs with a retained minimum contig length of 1000 bp. The average sequence coverage depth of the *R. vinicolor* assembly was 154X. Fifty percent of all nucleotides in the assembly were placed into 218 scaffolds (scaffold N50 = 218) with a minimum scaffold size of 43737 (raw N50 = 43737). The largest scaffold was 0.56 Mb and the DOE-JGI gene annotation pipeline identified 14469 *R. vinicolor* gene models. The genome of *R. vesiculosus* assembled to a total length of 42.2 Mb in 3765 contigs with a retained minimum contig length of 1000 bp. The average coverage depth of the *R. vesiculosus* assembly was 51X. The scaffold N50 of the *R. vesiculosus* assembly equaled 361 and the raw N50 equaled 30999. The MAKER gene annotation pipeline identified 13623 *R. vesiculosus* gene models. Both assemblies contained a high proportion of sequence data within predicted protein coding regions with 23.1 mb (63.9%) of the *R. vinicolor* genome and 20.6 mb (48.9.%) of the *R. vesiculosus* genome falling within the boundaries of predicted exons or introns. A complete listing of genome assembly statistics can be found in Table 3.1.

Gene content and synteny of the A-locus:

The organization of the genomic region, which flanks the A-locus homeodomain transcription factor proteins, was determined for several hundred kilobases in *R. vinicolor*, *R. vesiculosus*, *Suillus luteus*, *Serpula lacrymans*, *Laccaria bicolor*,

Coprinopsis cinerea, and *Phanerochaete chrysosporium*. All of the genomes examined possessed a single, divergently transcribed pair of HD proteins save for *C. cinerea* which possesses two pairs of convergently transcribed HD proteins in close proximity to one another (Figure 3.1, Stajich et al. 2010). Our analyses showed a high level of gene conservation between *Rhizopogon* genomes and reference genomes for a 400 Kbp region centered upon the HD proteins in *Rhizopogon.* The majority of predicted genes from this region in *Rhizopogon* species were also present in the vicinity of A-locus HD proteins within reference genomes. The genomic region beginning at approximately 100 kb 5' and extending to 30 kb 3' of HD proteins in both *Rhizopogon* species shows the highest level of gene conservation and synteny with homologous regions of all reference genomes (Figure 3.1). This region also possesses conserved gene content and synteny with homologous genomic regions in many previously characterized Agaricomycetes (James et al. 2013, James et al. 2004b, James et al. 2006). The HD proteins in both *Rhizopogon* genomes are flanked by the gene encoding the mitochondrial intermediate peptidase (MIP) protein and by gene encoding the beta-flanking protein (bfp); a gene organization that is consistent for nearly all characterized Agaricomycetes (James et al. 2004a, James 2007, Kües et al. 2011). Only some strains of *S. lacrymans* (Figure 3.1, Skrede et al. 2013) and *Schizophyllum commune* (Ohm et al. 2010) are known to vary from this pattern.

The genomic region beginning 35 kb 3' of HD proteins in both *Rhizopogon* species contains several blocks of genes that are conserved in reference genomes but possess translocated or inversed orientation relative to *Rhizopogon* species. The blocks of gene translocation and inversion possess synteny within the gene cluster and share a

common point of recombination approximately 100 kb 3' of the HD proteins in *Rhizopogon* species (Figure 3.2). A greater degree of synteny is shared by *Rhizopogon* species, *S. luteus,* and *P. chrysosporium* in this region than with the Agaricales reference genomes, *L. bicolor* and *C. cinerea.* In the Agaricales reference genomes a 120 Kbp region beginning at this point is translocated approximately 300 kb to the 5' side of HD proteins with several inversions of gene blocks within this translocation. The shared gene organization of the A-locus region in Boletales and Polyporales suggests that the gene organization of the A-locus is a derived trait of order Agaricales. A comparison of *Rhizopogon* and *Suillus* genomes with *S. lacrymans* 7.9 indicate that more recent recombination events have occurred in this region within order Boletales (Figure 3.2). *S. lacrymans* 7.9 possesses a major inversion of the same region which is translocated in Agaricales references and a small 5 kb translocation from the 5' to the 3' end of the inversion. This translocation begins approximately 10 kb 3' of the HD proteins and contains the gene encoding RPB2. In total, we mapped the genomic location of 159 predicted proteins from all genomes and several additional genes from reference genomes.

Gene Content and Synteny of the B-locus:

All genomes examined were found to contain a B-locus region containing P/R genes that shared high sequence identity and these genes were usually found in close proximity to P/P genes. P/P genes were not readily recognized by either the DOE-JGI or MAKER gene annotation pipelines and of all the P/P genes identified in this study approximately 50% were identified only by the use of custom Perl script. All the P/P

genes that were identified in annotation pipelines, however, were also identified by our script. P/P homolog genes of *Rhizopogon* species possess a C terminal –CAAX prenylation motif where (C = Cysteine residue, A = any aliphatic residue, X = any amino acid residue) as well as a 2 residue acidic "ER" domain approximately 10 – 15 amino acids upstream of the –CAAX motif (Table 3.2). The presence of these motifs in *C. cinerea* P/P genes is a demonstrated requirement for their function in mating (Riquelme et al. 2005).

All of the reference genomes selected for analysis in this study have been characterized as possessing either a bipolar or tetrapolar mating type system with the exception of *S. brevipes*. The bipolar species included were *C. puteana* (Ainsworth and Rayner 1990) and *S. luteus* (Fries and Neumann 1990) and the tetrapolar species were *P. involutus* (Fries 1985), *S. lacrymans* (Harmsen 1960), *L. bicolor* (Fries and Mueller 1984), and *C. cinerea* (Raper 1966). The average number of P/R genes found across all Boletales genomes was four and the average number of P/P genes was five (Table 3.3). Synteny between the B-locus regions is most conserved at the ordinal level with Boletales and Agaricales genomes all showing at least some synteny of P/R, P/P, and non-mating type genes with other members of their order. The major exceptions are *P. involutus*, which shows low synteny with other Boletales genomes surrounding the P/R and P/P genes, and *C. puteana*, which possesses a major inversion of the region 3' of P/R genes. There was only limited B-locus synteny observed between orders (Figure 3.3). The B-locus regions of *R. vesiculosus, R. vinicolor,* and *P. involutus* were present across multiple contigs and required reference to the other genomes in this analysis to assemble them confidently into scaffolds. Tetrapolar species of order Boletales tended to have a

greater number of both P/R and P/P genes than bipolar species and each P/R gene tended to be associated with a greater number of P/P genes than in bipolar species. Of the species with uncharacterized mating systems, *R. vinicolor* possessed a greater number of both P/R (6) and P/P (9) genes than *R. vesiculosus* (P/R 4, P/P 4) and *S. brevipes* shared very similar B-locus gene content and close synteny with its close bipolar relative, *S. luteus*.

Protein coding genes in proximity to B-locus P/R and P/P genes were generally conserved and syntenous across Boletales genomes. Of particular note is a gene immediately 3' of the P/R B-locus region (ICMTX). This gene possesses a predicted function as an isoprenyl cysteine methyltransferase (ICMT), which are a family of proteins responsible for the addition of a methyl group to the C terminal cysteine of mature pheromones (Caldwell et al. 1995). ICMTX is represented by a single copy just downstream of the P/R and P/P genes in *R. vesiculosus*, *S. brevipes* and the bipolar Boletales reference genomes, *S. luteus* and *C. puteana*. *C. puteana* possess an additional two ICMTX genes located 130 Kbp downstream of the B-locus region in association with an inverted and translocated B-locus gene cluster. *R. vinicolor* possesses two copies of ICMTX in immediate proximity to one another at the 3' end of the B-locus region as does the tetrapolar species *S. lacrymans.* In *R. vinicolor*, the B-locus region is broken into separate contigs at the location of the ICMTX pair and each contig contains a distinct allele of each ICMTX protein. The tetrapolar species *P. involutus* was also found to possess two copies of ICMTX in close proximity to the typical 3 prime terminal P/R gene of the Boletales B-locus (STE3.3). This block of genes in *P. involutus*, along with many other genes of conserved synteny in the B-locus region of other Boletales references, are

imbedded at the center of a separate and much larger contig than the contigs containing other *P. involutus* B-locus genes (Figure 3.3).

Phylogenetic Analysis of B-locus genes:

Phylogenetic inference of the evolutionary relationships between P/R genes (Figure 3.4) indicate that both *Rhizopogon* species possess P/R genes belonging to the two major P/R mating type families and three subfamilies identified in *C. cinerea* (Riquelme et al. 2005) and *Pleurotus* species (James et al. 2004b). Multiple paralogs of P/R families 1 and 2 were observed for *R. vinicolor* and the tetrapolar species, whereas *R. vesiculosus* possesses only one representative homolog in each of the four major clades. Because of the phylogenetic placement of novel *R. vinicolor* P/R paralogs in association with P/R paralogs from tetrapolar species it can be inferred that the reduced number of P/R genes observed in *R. vesiculosus*, *S. brevipes*, and the bipolar species represent a loss of P/R genes in these species rather than an expansion in *R. vinicolor* and the tetrapolar species. The only species found to lack at least one representative from each distinct clade of P/R genes were *S. brevipes* and the bipolar species *S. luteus* and *C. puteana.* Genes of known function in mating recognition in *S. lacrymans* (Skrede et al. 2013) and *C. cinerea* (O'shea et al. 1998, Riquelme et al. 2005) were found in all four of the major P/R gene clades.

We identified several ICMTX genes in close proximity to B-locus P/R and P/P genes in both *Rhizopogon* species and the Boletales reference genomes. ICMTX genes were not found in close proximity to the B-locus of Agaricales species but homologs of ICMTX genes were identified from Agaricales genomes by a BLAST query of *R.*

vinicolor ICMTX genes (JGI protein ID#s: 747032, 724114, 411247, 411249). Additional ICMT genes were identified from all genomes by a BLAST query of the ICMT pheromone maturation factor STE14 from *S. cerevisiae* (Genbank Accession #: P32584). The top scoring hits from BLAST searches were aligned with Boletales B-locus ICMTX genes and results of phylogenetic analysis are presented as an unrooted phylogram in Figure 3.5. The pairs of B-locus ICMTX genes present in *R. vinicolor* and in tetrapolar Boletales reference genomes all cluster within their individual species and the single ICMTX gene of *R. vesiculosus* (MAKER ID: 7665) is grouped with high confidence with only one allele of the two ICMTX genes of *R. vinicolor* (JGI ID: 724114). B-locus region ICMTX genes of the Boletales are not separated from other ICMT genes with high confidence. STE14 is grouped with 100 percent bootstrap support with ICMT homologs identified as its top BLAST hit from each genome save for *C. puteana*. Placement of the closest relative to STE14 from *C. puteana* with B-locus ICMTX homologs suggest that all B-locus ICMTX genes and those ICMT genes identified using BLAST are homologs of STE14. In addition, the two unique ICMTX genes located 130 Kbp downstream from P/R and P/P genes in *C. puteana* are grouped with the best BLAST hit to STE14 from *C. puteana,* which does not reside near the B-locus region (Figure 3.5). These two ICMTX genes are located at the terminus of a block of other B-locus genes that have been translocated and inverted from the original B-locus region of *C. puteana* and it is likely that they have been translocated as a unit.

SNP analysis and rates of heterozygosity within the individual

Although the *R. vinicolor* genome possessed only 251 additional gene models in comparison to *R. vesiculosus* it possessed 2.5 mb more sequence within gene models. VARSCAN2 predicted a total of 229571 SNPs in the *R. vesiculosus* genome and 392731 SNPs in the *R. vinicolor* genome. *R. vinicolor* possessed a greater proportion of its SNPs in predicted genes than *R. vesiculosus* with 82.4% of all SNPs within intergenic regions and 17.9% in genic regions. In comparison, *R. vesiculosus* had 88.6% of all SNPs in intergenic regions and 11.4% in genic regions. The rate of non-synonymous mutation was also greater in *R. vinicolor* with 6.4% of all SNPs causing non-synonymous mutations compared to 3.8% of all SNPs causing non-synonymous mutations in *R. vesiculosus*. A comparison of SNP ratios for all *R. vinicolor* and *R. vesiculosus* assemblies is presented in Table 3.4.

DISCUSSION:

Summary of Results:

In this study we report for the first genome sequences of basidiomycete false-truffles *R. vinicolor* and *R. vesiculosus* and characterize the gene content and organization of their complete A and B mating type loci. The A-locus HD genes and the B-locus P/R and P/P genes in these *Rhizopogon* species were revealed as homologs of genes which possess known mating type function in *C. cinerea* and *S. lacrymans.* We have also characterized the B-locus P/R and P/P homolog genes of the EM Boletales *P. involutus, S. brevipes*, and *S. luteus.* Gene content and synteny of the A-locus region surrounding

HD genes is largely conserved for all of the Boletales genomes analyzed and there is a significant degree of synteny surrounding the HD proteins between Boletales genomes and *P. chrysosporium* (Polyporales). Agaricales genomes show a derived translocation around HD proteins. The B-locus of all fungi observed showed greater variation in mating type gene content than A-locus regions and were more similar within orders than between them. Finally, we conducted a genome wide mapping of SNPs that were detectable in the dikaryotic assembly of both *Rhizopogon* species and found the *R. vinicolor* possessed a greater rate of both non-synonomous and synonymous SNP mutations.

Comparison of bipolar and tetrapolar B-locus structure in the Boletales:

We chose reference genomes from the order Boletales that were known to possess either a bipolar or tetrapolar mating system to determine if gene content or synteny of mating type loci might provide clues to the unknown mating systems of *Rhizopogon* species. No major differences were observed in gene content or synteny between Boletales genomes in the A-locus region immediately surrounding HD proteins. While *S. lacrymans* did possess large inversions near HD proteins, these inversions did not alter HD gene content or synteny. The most notable differences between the mating loci of bipolar and tetrapolar Boletales examined is seen in the gene content of the B-locus region. The bipolar species *S. luteus* and *C. puteana* showed a reduced diversity of both P/R and P/P genes and lacked a second copy of the ICMT genes which flank the 3' terminal Boletales P/R gene, STE3.3. Both bipolar species also lacked homologs within the STE 3.2 and STE3.1 P/R gene families and possessed at least two paralogs in the

STE3.4 family. The B-locus of *S. brevipes* possessed near identical gene content and synteny with *S. luteus* and this lends support for a bipolar mating system in *S. brevipes.* In contrast to the bipolar Boletales, the tetrapolar Boletales *P. involutus* and *S. lacrymans* both possessed homologs in all P/R families, multiple paralogs within many P/R families and possess a paired doublet of ICMTX genes flanking STE3.3. The B-locus of *R. vinicolor* shows the most similarity to that of the tetrapolar Boletales with multiple P/R paralogs in both the STE3.4 and STE3.3 families and a greater number of P/P genes flanking P/R genes. *R. vesiculosus* lacked paralogs within any of the P/R families, possessed fewer P/P genes and only a single ICMTX gene flanking STE3.3. Taken together, the features of the *R. vinicolor* B-locus are consistent with the features of tetrapolar species while *R. vesiculosus* possesses features, i.e., reduced genetic diversity, at its B-locus that is most similar to bipolar species.

It is hypothesized that bipolar mating systems in Agaricomycetes have been derived multiple times from tetrapolar ancestors (Hibbett and Donoghue 2001) and this transition has been observed in Basidiomycota through genetic linkage and cosegretation of the A and B mating type loci (Bakkeren and Kronstad 1994) and through a loss of specificity or function in B-locus P/R genes (James et al. 2006). The observation of fewer P/R and P/P genes in bipolar Boletales genomes compared to tetrapolar Boletales genomes is consistent with a loss of function and subsequent loss of genetic diversity. The reduction in B-locus P/R and P/P gene content in *R. vesiculosus* compared to *R. vinicolor* is not as drastic as the difference between *R. vinicolor* and the bipolar species *S. luteus* and *C. puteana.* This might indicate that *R. vesiculosus* has lost some mating type P/R specificity in recent evolutionary history and it is in the process of gene loss through

drift associated with reduced effective population size. The P/R and P/P genes of Agaricomycetes mating type B-locus are highly variable between haplotypes and lack or truncation of particular P/P genes has been observed for some *C. cinerea* B-locus haplotypes (Riquelme et al. 2005). Thus, it is also possible that we have selected a strain of *R. vesiculosus* for genome sequencing that simply possesses a deletion of P/R and P/P loci in both of its haplotypes or that our genome assembly was biased for only one haplotype that lacked P/R and P/P loci. However, it is more likely that *R. vesiculosus* truly possesses reductions in P/R and P/P genes since these reductions are also observed in the bipolar species *S. luteus* and *C. puteana.*

Conserved homology and function of pheromone receptor, pheromone precursor, and isoprenyl methyl transferase genes.

The P/P genes identified in *Rhizopogon* genomes contain both –CAAX and "ER" motifs that support their role in mating recognition. The –CAAX motif of fungal P/P genes is known to flag peptides for isoprenyl modification and the prenyl moiety incorporated into the modified protein is strongly influenced by the final "X" residue of the –CAAX motif. Alanine, serine, glutamine, cysteine, or methionine in this position targets a peptide for farnesylation and a leucine targets it for geranylgeranylation (Caldwell et al. 1995). All characterized fungal pheromones are prenylated with a farnesyl moiety (Casselton and Olesnicky 1998, Brown and Casselton 2001, Michaelis et al. 2012) and the P/P genes identified from *Rhizopogon* species are all terminated by residues that target them for farnesylation (Table 3.2). The P/P genes identified from *Rhizopogon* genomes possess highly variable amino acid sequences but they all share the

two residue acidic "ER" domain upstream of the –CAAX motif that is the site of N terminal proteolysis in mature fungal pheromones (Brown and Casselton 2001, Caldwell et al. 1995). The presence of an "ER" or "EH" doublet at the N terminus of mature pheromones is an important element in the activation of complementary pheromone receptors in *C. cinerea* (Brown and Casselton 2001, Olesnicky et al. 1999), though N terminal doublets of "QR", "DR", and "ED" have also been identified in active *C. cinerea* pheromones (Riquelme et al. 2005). All of the *Rhizopogon* P/P genes identified in this study possess acidic "ER" doublets at a position consistent with the potential N terminus of a mature pheromone (Table 3.2).

Much of our understanding of the structure and function of Basidiomycota mating type P/R and P/P genes is derived from the model Ascomycota yeast *S. cerevisiae*. For instance the lipopeptide pheromones identified as major mating signal molecules in the Basidiomycota species *S. commune* (Wendland et al. 1995), *Tremella mesenterica* (Sakagami et al. 1981), and *C. cinerea* (Olesniky et al. 1999) share common peptide domains and all undergo posttranslational modifications that produce mature lipophillic farnesylated pheromones. These pheromones are homologous in structure and function to the mating type a-factor pheromones known to induce mating in *S. cerevisiae* (Caldwell et al. 1995). Homologous transformation studies which transformed *S. commune* (Fowler et al. 1999) and *C. cinerea* (Olesniky et al. 1999) mating type pheromone precursor and pheromone receptor genes into *S. cerevisiae* have observed effective translation, post-translational modification, and secretion of transformant pheromones and *S. cerevisiae* mutants of this type are mating competent. We identified P/R genes in both *Rhizopogon* species and in Boletales reference genomes that group with high confidence into clades

containing these same P/R genes from *C. cinerea* known to function in mating type recognition. Because the structure and function of B-locus P/R and P/P genes is highly conserved across Ascomycota and Basidiomycota (=Dikarya), it is likely that the P/R and P/P genes we have identified from *Rhizopogon* genomes are homologous to those identified in *C. cinerea.*

Though a common system of pheromone signaling functions in the mating system of many Dikarya, it is unknown if the post translational modifications to these pheromones in Basidiomycota fungi are orchestrated by a homologous genetic pathway as the one which has been characterized in *S. cerevisiae* (Chen et al. 1997). There is evidence to support that endogenous post translational modification of farnesylated lipopeptide pheromones in *Ustilago hordei* (Basidiomycota: Ustilaginomycotina) (Kosted et al. 2000) as well as *C. cinerea* (Brown and Casselton 2001) are highly similar to those performed by *S. cervisiae* and are critical for the pheromone's ability to activate complementary receptors. Methylation of C terminal cysteine residue of mature pheromones is performed in *S. cerevisiae* by the ICMT gene STE14 and methylation is critical for excretion of mature pheromones from the cell (Caldwell et al. 1995, Chen et al. 1997). Similarly, Kosted et al. (2000) found that methylation of the C terminal cysteine of mating type pheromone in *U. hordei* served as a critical component in pheromone maturation and export. We identified putative gene duplicates of the ICMTX gene with INTERPRO predicted ICMT function in close proximity to the B-locus P/R and P/P genes of all Boletales genomes examined. Phylogenetic analysis of the ICMTX genes from all taxa in this study supports a homologous relationship between these genes and the STE14 gene of *S. cerevisiae*. Given the close linkage of ICMTX genes to the B-

locus in Boletales, the demonstrated role of cysteine methylation in mating pheromone maturation, and the ability of the endogenous *S. cerevisiae* pheromone maturation pathway to activate Agaricomycetes pheromones, it is likely that the ICMTX genes we have identified in *Rhizopogon* species are involved with mating pheromone maturation. Duplication of the ICMTX gene in *S. lacrymans, P. involutus,* and *R. vinicolor* may confer specialized or enhanced maturation of lipopeptide mating pheromones necessary for tetrapolar mating. Duplication of the ICMTX gene in *R. vinicolor* and the apparent loss of the duplicate in *R. vesiculosus* may also confer differential pheromone maturation pathways in these fungi. Gene duplication has been demonstrated as a potential source of genomic incompatability resulting in reproductive isolation between sister taxa (Lynch and Force 2000). It is unlikely, however, that speciation between *R. vinicolor* and *R. vesiculosus* was spurred by gene duplications in the mating pheromone pathway but rather that such duplication events could account for differential mating behavior. *P. involutus* also possessed duplicates of the ICMTX gene as a tightly associated pair. Given the observed pattern of conserved synteny of the region containing ICMTX in other Boletales genomes, the translocation of STE3.3 and the large block of containing the ICMTX pair onto a separate contig in *P. involutus* likely represents a derived recombinatorial state. This region is likely no longer in linkage with other B-locus proteins in *P. involutus*.

Gene conservation and synteny of the mating type A-locus

We chose reference genomes from Agaricales, Boletales, and Polyporales in order to make comparisons of the A-locus mating type region across the class Agaricomycetes.

Alignment of an additional genome from the Auriculariales (*Auricularia delicata*) was attempted but was determined to be too divergent from other genomes to be useful in our analysis. We found that gene content and synteny is reasonably well conserved for approximately 100 kb centered around the HD proteins of all genomes analyzed in these three orders. These findings are consistent with previous analyses comparing this region for many of the same genomes included in our analysis (James et al. 2013). However, when the region of interest was expanded to several hundred kilobase pairs flanking the HD proteins a previously undetected set of genomic rearrangements was detected in the genomes of *L. bicolor* and *C. cinerea*. Both of these Agaricales genomes possess a translocation of a 100—200 Kbp region from the 3' region of Boletales and *P. chrysosporium* A-loci to their 5' A-locus region. Though such a translocation has not previously been noted between Agaricales, Polyporales and Boletales A-loci, it has been noted between members of Agaricales. Van Peer et al. (2011) noted a translocation of genes from the same region 3' of HD proteins in *C. cinerea* and *L. bicolor* to the 5' A-locus region in *S. commune.* In addition to translocations in *C. cinerea* and *L. bicolor* we also noted a major inversion in this region of the A-locus in *S. lacrymans* relative to other genomes examined. Van Peer et al. (2011) noted such an inversion of this region in *Flammulina velutipes* (Agaricales) relative to *C. cinerea* and *L. bicolor*. Given our common findings of gene rearrangement with Van Peer et al. (2011) it seems likely that a common point of recombination for Agaricomycetes exists in this region of the A-locus that is centered approximately 35 – 100 kb 3' of the HD transcription factor proteins.

Polyporales is an earlier diverging member of the Agaricomycetes than Agaricales or Boletales (Hibbett 2006). The presence of a shared synteny between Boletales

genomes and *P. chrysosporium*, which is not shared by *L. bicolor* and *C. cinerea*, suggests that these Agaricales genomes possess a derived gene organization at the A-locus. Van Peer et al. (2011) noted genomic reorganizations at a similar location as found in our study but involved different genes than those that occur in homologous regions of Boletales and Polyporales studied here. This suggests that inversions and translocations of genes in this region are common in the evolutionary history of Agaricomycetes and that this region may represent a recombination hotspot shared amongst many lineages of Agaricomycetes. Strong balancing selection is known to function in the HD proteins of *C. cinerea* (May et al. 1999) and it is possible that a recombination hotspot near the HD proteins may function to break linkage of surrounding regions with HD proteins and to encourage development of new mating type specificities. Rearrangements at this recombination hotspot from an ancestral gene arrangement shared with *C. cinerea* have been implicated in the duplication of HD proteins in *F. velutipes* and *S. commune* (Van Peer et al. 2011) and such duplications could function to increase the diversity of mating specificities. It is likely that similar rearrangements in this region will be observed as more Agaricomycetes genomes are characterized at their A-locus.

Heterozygosity within the dikaryon, an inherent challenge of dikaryotic genome assembly

Because SNPs are evident in our dikaryotic *Rhizopogon* genome assemblies it is evident that many of the contigs assembled by VELVET are in fact chimeric sequences that combine the sequences of two homologous chromosomes. We utilized the presence of SNPs in this assembly as a means of gauging the rate of heterozygosity in each

individual. The genome assembly of *R. vesiculosus* is 6.1 mb larger than that of *R. vinicolor* and though the number of predicted gene models is similar between these assemblies the percent of the *R. vinicolor* assembly present in genic regions (63.9%) is much greater than in the *R. vesiculosus* assembly (48.9%). This indicates that we have assembled a greater proportion of the intergenic genome regions of *R. vesiculosus* than in *R. vinicolor*. It would be expected that heterozygosity between haplotypes should be most evident in non-conserved intergenic regions and the reduced assembly of *R. vinicolor* in these regions is consistent with a higher rate of heterozygosity within this individual. Our dikaryotic SNP predictions confirm this expectation by placing the majority of SNPs for both genomes within intergenic regions despite the large proportion of both genomes contained within genic regions. A more complete assembly of non-coding regions in *R. vesiculosus* due to reduced heterozygosity is not surprising given that the individual sequenced in this study was drawn from a population that shows signs of inbreeding (Dunham et al. 2013). It is likely that the true genome size of both *R. vinicolor* and *R. vesiculosus* has been underestimated in our assemblies because of the difficulty in reconstructing variable intergenic regions.

SNP analysis found a greater total number of SNPs and a higher rate of non-synonymous SNPs mutations for *R. vinicolor* than for *R. vesiculosus* indicating that this individual of *R. vinicolor* possesses greater heterozygosity in coding regions as well as non-coding regions. In our assembly of the *R. vinicolor* B-locus region we found that two of the contigs containing this region were each terminated by a pair of homologous proteins with a predicted ICMT function. This suggests that contig extension was terminated in this region during the assembly process due to the presence of divergent

alleles between haplotypes. In the genome sequence assembly of the diploid yeast *Candida albicans* a similar pattern of heterozygous regions occurring at the end of assembly contigs was observed and such heterozygous contig junctions were used as evidence to justify the joining of such contigs into scaffolds (Jones et al. 2004). The termination of contig extension during assembly due to encounters with highly heterozygous regions is a probable explanation for the fragmented nature of our *Rhizopogon* genome assemblies. Assembly of the B-locus region in *R. vinicolor* was more fragmented in than in *R. vesiculosus* and this may be due to a greater degree of heterozygosity in these loci for *R. vinicolor*. In cases where assembly contigs are punctuated by heterozygous loci an all-by-all BLASTN (nucleotide) search of assembly contigs has proven to be an effective means of identifying homologous terminal contig regions and closing gaps between these contigs (Jones et al. 2004). BLASTN searches of *R. vinicolor* contigs using the nucleotide sequence of *R. vinicolor* B-locus proteins as query identified two additional contigs of 1000 – 2000 bp in length that contained fragments of P/R and ICMTX genes at their terminus. It is likely that these gene fragments represent incomplete assemblies of divergent alleles from the B-locus region of *R. vinicolor*, which suggests that the mating type region of *R. vinicolor* is highly heterozygous. Because of the difficulty in assembling heterozygous genomic regions it is possible that we have only assembled sequence from a single chromosomal haplotype due to a sequencing or assembly bias. Without homokaryotic reference sequences for the two genome copies present in our assembly it is not possible to differentiate regions that represent a single haplotype from those which fall in genomic regions which naturally lack variation between chromosomal haplotypes.

Conclusions:

In this study we have characterized the genomic structure of the mating systems of the EM fungi *R. vinicolor* and *R vesiculosus.* Our comparison of mating type loci spanned several species from a breadth of Agaricomycetes and found a correlation between bipolar mating systems and reduced genetic diversity at the mating type B-locus. These results indicate that loss of genetic diversity and function at the B-locus may be associated with the transition from tetrapolar to bipolar mating systems in the Boletales. The organization of *R. vinicolor* B-locus was similar to those of tetrapolar Boletales species while the organization of *R. vesiculosus* B-locus shared the greatest similarity with bipolar Boletales species. While this evidence alone does not conclusively demonstrate a bipolar mating system in *R. vesiculosus* or a tetrapolar system in *R. vinicolor*, such a finding would be consistent with observed differences in population structure for these sister species (Dunham et al. 2013, Kretzer et al. 2005).

The *Rhizopogon* genomes presented in this study are the first genomes available for Basidiomycota truffles and represent an important step forward in the genome-enabled study of EM fungi. The dikaryotic nature of these genome assemblies posed some challenges to their assembly but also provided a novel opportunity for the characterization of heterozygosity within an individual. Here we present for the first time a method for the characterization of genome scale heterozygosity by characterizing SNPs and the rate of non-synonymous mutation between haplotypes of a single fungal individual. This technique is applicable to any polyploidy assembly for which gene models are available in .gff3 file format and Perl scripts for performing this analysis are

made available in Appendix 1. The availability of these *Rhizopogon* genomes and the methods demonstrated here will facilitate future studies of mating function and population structure for additional *R. vinicolor* and *R. vesiculosus* isolates. As more Basidiomycota truffle genomes become available we foresee the potential for genome wide association studies that investigate the developmental biology of the hypogeous fruiting habit.

REFERENCES:

Aimi, Tadanori, Rie Yoshida, Mariko Ishikawa, Dapeng Bao, and Yutaka Kitamoto. 2005. Identification and Linkage Mapping of the Genes for the Putative Homeodomain Protein (hox1) and the Putative Pheromone Receptor Protein Homologue (rcb1) in a Bipolar Basidiomycete, *Pholiota nameko*. *Current Genetics* 48 (3): 184–94. doi:10.1007/s00294-005-0012-7.

Ainsworth, A. M., and A. D. M. Rayner. 1990. Mycelial Interactions and Outcrossing in the *Coniophora puteana* Complex. *Mycological Research* 94 (5): 627–34. doi:10.1016/S0953-7562(09)80664-3.

Bakkeren, G., and J. W. Kronstad. 1994. Linkage of Mating-Type Loci Distinguishes Bipolar from Tetrapolar Mating in Basidiomycetous Smut Fungi. *Proceedings of the National Academy of Sciences* 91 (15): 7085–89.

Bakkeren, Guus, Guoqiao Jiang, René L. Warren, Yaron Butterfield, Heesun Shin, Readman Chiu, Rob Linning, et al. 2006. Mating Factor Linkage and Genome Evolution in Basidiomycetous Pathogens of Cereals. *Fungal Genetics and Biology* 43 (9): 655–66. doi:10.1016/j.fgb.2006.04.002.

Beiler KJ, Durall DM, Simard SW, Maxwell SA and Kretzer AM. 2010. Architecture of the wood-wide web: *Rhizopogon spp*. genets link multiple Douglas-fir cohorts *New Phytologist* 185(2) 543-553.

Beiler, Kevin J., Suzanne W. Simard, Valerie LeMay, and Daniel M. Durall. 2012. Vertical Partitioning between Sister Species of *Rhizopogon* Fungi on Mesic and Xeric Sites in an Interior Douglas-Fir Forest. *Molecular Ecology* 21 (24): 6163–74. doi:10.1111/mec.12076.

Bölker, Michael, Martin Urban, and Regine Kahmann. 1992. The a Mating Type Locus of *U. maydis* Specifies Cell Signaling Components. *Cell* 68 (3): 441–50. doi:10.1016/0092-8674(92)90182-C.

Brown, Andrew J., and Lorna A. Casselton. 2001. Mating in Mushrooms: Increasing the Chances but Prolonging the Affair. *Trends in Genetics* 17 (7): 393–400. doi:10.1016/S0168-9525(01)02343-5.

Bruns TD, Fogel R, White TJ, Palmer JD. 1989. Accelerated evolution of a false-truffle from a mushroom ancestor. *Nature* 339:140–142

Caldwell, G. A., F. Naider, and J. M. Becker. 1995. Fungal Lipopeptide Mating Pheromones: A Model System for the Study of Protein Prenylation. *Microbiological Reviews* 59 (3): 406–22.

Cantarel, Brandi L., Ian Korf, Sofia M. C. Robb, Genis Parra, Eric Ross, Barry Moore, Carson Holt, Alejandro Sánchez Alvarado, and Mark Yandell. 2008. MAKER: An Easy-to-Use Annotation Pipeline Designed for Emerging Model Organism Genomes. *Genome Research* 18 (1): 188–96. doi:10.1101/gr.6743907.

Casselton LA, Kües U (2007) *Sex in Fungi: Molecular Determination and Evolutionary Implications*, eds Heitman J, Kronstad JW, Taylor JW, Casselton LA (American Association for Microbiology Press, Washington), pp 283–300.

Casselton, Lorna A., and Natalie S. Olesnicky. 1998. Molecular Genetics of Mating Recognition in Basidiomycete Fungi. *Microbiology and Molecular Biology Reviews* 62 (1): 55–70.

Chen, Peng, Stephanie K. Sapperstein, Jonathan D. Choi, and Susan Michaelis. 1997. Biogenesis of the *Saccharomyces cerevisiae* Mating Pheromone a-Factor. *The Journal of Cell Biology* 136 (2): 251–69. doi:10.1083/jcb.136.2.251.

Dahlberg, Anders, and Jan Stenlid. 1994. Size, Distribution and Biomass of Genets in Populations of *Suillus bovinus* (L.: Fr.) Roussel Revealed by Somatic Incompatibility. *New Phytologist* 128 (2): 225–34. doi:10.1111/j.1469-8137.1994.tb04006.x.

Darling, ACE., Mau B, Blattner FR, and Perna NT. 2004. Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Research* 14 (7): 1394–1403. doi:10.1101/gr.2289704.

Debaud, J. C., and G. Gay. 1987. In Vitro Fruiting Under Controlled Conditions of the Ectomycorrhizal Fungus *Hebeloma cylindrosporum* Associated with *Pinus pinaster*. *New Phytologist* 105 (3): 429–35. doi:10.1111/j.1469-8137.1987.tb00880.x.

De La Bastide, Pall Y., Bradley R. Kropp, and Yves Piché. 1994. Spatial Distribution and Temporal Persistence of Discrete Genotypes of the Ectomycorrhizal Fungus *Laccaria bicolor* (Maire) Orton. *New Phytologist* 127 (3): 547–56. doi:10.1111/j.1469-8137.1994.tb03973.x.

Dieckmann, Ulf, and Michael Doebeli. 1999. On the Origin of Species by Sympatric Speciation. *Nature* 400 (6742): 354–57. doi:10.1038/22521.

Douhan, Greg W., Lucie Vincenot, Hervé Gryta, and Marc-André Selosse. 2011. Popultion Genetics of Ectomycorrhizal Fungi: From Current Knowledge to Emerging Directions. *Fungal Biology* 115 (7): 569–97. doi:10.1016/j.funbio.2011.03.005.

Dunham, Susie M., Alija Bajro Mujic, Joseph W. Spatafora, and Annette M. Kretzer. 2013. Within-Population Genetic Structure Differs between Two Sympatric Sister-Species of Ectomycorrhizal Fungi, *Rhizopogon vinicolor* and *R. vesiculosus*. *Mycologia* 105 (4): 814–26. doi:10.3852/12-265.

Eastwood, Daniel C., Dimitrios Floudas, Manfred Binder, Andrzej Majcherczyk, Patrick Schneider, Andrea Aerts, Fred O. Asiegbu, et al. 2011. The Plant Cell Wall-Decomposing Machinery Underlies the Functional Diversity of Forest Fungi. *Science* 333 (6043): 762–65. doi:10.1126/science.1205411.

Edgar, Robert C. 2004. MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Research* 32 (5): 1792–97. doi:10.1093/nar/gkh340.

Floudas, Dimitrios, Manfred Binder, Robert Riley, Kerrie Barry, Robert A. Blanchette, Bernard Henrissat, Angel T. Martínez, et al. 2012. The Paleozoic Origin of Enzymatic Lignin Decomposition Reconstructed from 31 Fungal Genomes. *Science* 336 (6089): 1715–19. doi:10.1126/science.1221748.

Fowler, Thomas J., Susan M. DeSimone, Michael F. Mitton, Janet Kurjan, and Carlene A.Raper. 1999. Multiple Sex Pheromones and Receptors of a Mushroom-Producing Fungus Elicit Mating in Yeast. *Molecular Biology of the Cell* 10 (8): 2559–72. doi:10.1091/mbc.10.8.2559.

Fries, Nils, and Gregory M. Mueller. 1984. Incompatibility Systems, Cultural Features and Species Circumscriptions in the Ectomycorrhizal Genus *Laccaria* (Agaricales). *Mycologia* 76 (4): 633–42. doi:10.2307/3793220.

Fries N. 1985. Intersterility groups in *Paxillus involutus*. *Mycotaxon*. 24:403-409

Fries, N., and W. Neumann. 1990. Sexual Incompatibility in *Suillus luteus* and *S. granulatus*. *Mycological Research* 94 (1): 64–70. doi:10.1016/S0953-7562(09)81265-3.

Gnerre S, MacCallum I, Przybylski D, Ribeiro F, Burton J, Walker B, Sharpe T, Hall G, Shea T, Sykes S, Berlin A, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences.* 108 (4) 1513-1518.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology* 29(7):644-52. doi: 10.1038/nbt.1883. PubMed PMID: 21572440.

Grigoriev IV, Martinez DA, Salamov AA. 2006. Fungal genomic annotation. In *Applied Mycology and Biotechnology* (Eds. Aurora, DK, Berka, RM, Singh, GB), Elsevier Press, Vol 6 (Bioinformatics), 123-142

Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F, Smirnova T, Nordberg H, Dubchak I, Shabalov I. 2014. MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res*. 42(1):D699-704.

Gugger, Paul F., Shinya Sugita, and Jeannine Cavender-Bares. 2010. Phylogeography of Douglas-Fir Based on Mitochondrial and Chloroplast DNA Sequences: Testing Hypotheses from the Fossil Record. *Molecular Ecology* 19 (9): 1877–97. doi:10.1111/j.1365-294X.2010.04622.x.

Harmsen, L. 1960. Taxonomic and cultural studies on brown spored species of the genus *Merulius*. Friesia 6 (4): 233–77 pp.

Heitman, Joseph, Sheng Sun, and Timothy Y. James. 2013. Evolution of Fungal Sexual Reproduction. *Mycologia* 105 (1): 1–27. doi:10.3852/12-253.

Hibbett, David S., and Michael J. Donoghue. 2001. Analysis of Character Correlations Among Wood Decay Mechanisms, Mating Systems, and Substrate Ranges in Homobasidiomycetes. *Systematic Biology* 50 (2): 215–42. doi:10.1080/10635150121079.

Hibbett, David S. 2006. A Phylogenetic Overview of the Agaricomycotina. *Mycologia* 98 (6): 917–25. doi:10.3852/mycologia.98.6.917.

Hibbett DS, Matheny PB. 2009. The relative ages of ectomycorrhizal mushrooms and their plant hosts estimated using Bayesian relaxed molecular clock analyses. *BMC Biology*. 7:13 doi:10.1186/1741-7007-7-13

Horton, Thomas R. 2006. The Number of Nuclei in Basidiospores of 63 Species of Ectomycorrhizal Homobasidiomycetes. *Mycologia* 98 (2): 233–38. doi:10.3852/mycologia.98.2.233.

Izzo, Antonio D., Marc Meyer, James M. Trappe, Malcolm North, and Thomas D. Bruns. 2005. Hypogeous Ectomycorrhizal Fungal Species on Roots and in Small Mammal Diet in a Mixed-Conifer Forest. *Forest Science* 51(3)243-254

James, Timothy Y., Ursula Kües, Stephen A. Rehner, and Rytas Vilgalys. 2004a. Evolution of the Gene Encoding Mitochondrial Intermediate Peptidase and Its Cosegregation with the A Mating-Type Locus of Mushroom Fungi. *Fungal Genetics and Biology: FG & B 41* (3): 381–90. doi:10.1016/j.fgb.2003.11.008.

James, Timothy Y., Shian-Ren Liou, and Rytas Vilgalys. 2004b. The Genetic Structure and Diversity of the A and B Mating-Type Genes from the Tropical Oyster Mushroom, *Pleurotus djamor*. *Fungal Genetics and Biology:* FG & B 41 (8): 813–25. doi:10.1016/j.fgb.2004.04.005.

James, Timothy Y., Prayook Srivilai, Ursula Kües, and Rytas Vilgalys. 2006. Evolution of the Bipolar Mating System of the Mushroom *Coprinellus disseminatus* From Its Tetrapolar Ancestors Involves Loss of Mating-Type-Specific Pheromone Receptor Function. *Genetics* 172 (3): 1877–91 doi:10.1534/genetics.105.051128.

James TY. 2007. Analysis of mating-type locus organization and synteny in mushroom fungi- beyond model species. In: Heitman J, Kronstad J, Taylor JW, Casselton LA, eds. *Sex in fungi: molecular determination and evolutionary implications*. Washington DC: ASM Press. pp 317–331.

James, Timothy Y., Sheng Sun, Wenjun Li, Joseph Heitman, Hsiao-Che Kuo, Yong-Hwan Lee, Frederick O. Asiegbu, and Åke Olson. 2013. Polyporales Genomes Reveal the Genetic Architecture Underlying Tetrapolar and Bipolar Mating Systems. *Mycologia* 105 (6): 1374–90. doi:10.3852/13-162.

Jones, Ted, Nancy A. Federspiel, Hiroji Chibana, Jan Dungan, Sue Kalman, B. B. Magee, George Newport, et al. 2004. The Diploid Genome Sequence of *Candida albicans*. *Proceedings of the National Academy of Sciences of the United States of America* 101 (19): 7329–34. doi:10.1073/pnas.0401648101.

Koboldt, Daniel C., Ken Chen, Todd Wylie, David E. Larson, Michael D. McLellan, Elaine R. Mardis, George M. Weinstock, Richard K. Wilson, and Li Ding. 2009. VarScan: Variant Detection in Massively Parallel Sequencing of Individual and Pooled Samples. *Bioinformatics* 25 (17): 2283–85 doi:10.1093/bioinformatics/btp373.

Kosted, Paula J., Shirley A. Gerhardt, Cynthia M. Anderson, Andrea Stierle, and John E. Sherwood. 2000. Structural Requirements for Activity of the Pheromones of *Ustilago hordei*. *Fungal Genetics and Biology* 29 (2): 107–17. doi:10.1006/fgbi.2000.1191.

Kretzer AM, Dunham S, Molina R, Spatafora JW. 2003. Microsatellite markers reveal the below ground distribution of genets in two species of *Rhizopogon* forming tuberculate ectomycorrhizas on Douglas fir. *New Phytologist* 161(1) 313-320

Kretzer AM, Dunham S, Molina R, Spatafora JW. 2005. Patterns of vegetative growth and gene flow in Rhizopogon vinicolor and R. vesiculosus (Boletales, Basidiomycota) *Molecular Ecology* 14(8) 2259-2268

Kües U., James T. Y., Heitman J.(Editors), 2011 Mating Type in Basidiomycetes: Unipolar, Bipolar and Tetrapolar Patterns of Sexuality. Springer Verlag, Berlin.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. 2009. Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome. *Genome Biology* 10 (3): R25. doi:10.1186/gb-2009-10-3-r25.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* 25 (16): 2078–79. doi:10.1093/bioinformatics/btp352.

Lynch, Michael, and Allan G. Force. 2000. The Origin of Interspecific Genomic Incompatibility via Gene Duplication. *The American Naturalist* 156 (6): 590–605. doi:10.1086/an.2000.156.issue-6.

Martin, F., A. Aerts, D. Ahrén, A. Brun, E. G. J. Danchin, F. Duchaussoy, J. Gibon, et al. 2008. The Genome of *Laccaria bicolor* Provides Insights into Mycorrhizal Symbiosis. *Nature* 452 (7183): 88–92. doi:10.1038/nature06556.

May, Georgiana, Frank Shaw, Hassan Badrane, and Xavier Vekemans. 1999. The Signature of Balancing Selection: Fungal Mating Compatibility Gene Evolution. *Proceedings of the National Academy of Sciences* 96 (16): 9172–77. doi:10.1073/pnas.96.16.9172.

Michaelis, Susan, and Jemima Barrowman. 2012. Biogenesis of the *Saccharomyces cerevisiae* Pheromone a-Factor, from Yeast Mating to Human Disease. *Microbiology and Molecular Biology Reviews* 76 (3): 626–51. doi:10.1128/MMBR.00010-12.

Molina, Randy, and James M. Trappe. 1994. Biology of the Ectomycorrhizal Genus, *Rhizopogon*: I. Host associations, host-specificity and pure culture syntheses *New Phytologist* 126 (4): 653–75. doi:10.1111/j.1469-8137.1994.tb02961.x.

Niculita-Hirzel H,Hirzel AH (2008) Visualizing the gene order conservation among genomes with ChromoMapper. *International Journal of Computational Intelligence in Bioinformatics and Systems Biology* 1

Niculita-Hirzel, Hélène, Jessy Labbé, Annegret Kohler, François Le Tacon, Francis Martin, Ian R. Sanders, and Ursula Kües. 2008. Gene Organization of the Mating Type Regions in the Ectomycorrhizal Fungus *Laccaria bicolor* Reveals Distinct Evolution between the Two Mating Type Loci. *New Phytologist* 180 (2): 329–42. doi:10.1111/j.1469-8137.2008.02525.x.

Ohm, Robin A., Jan F. de Jong, Luis G. Lugones, Andrea Aerts, Erika Kothe, Jason E. Stajich, Ronald P. de Vries, et al. 2010. Genome Sequence of the Model Mushroom *Schizophyllum commune*. *Nature Biotechnology* 28 (9): 957–63. doi:10.1038/nbt.1643.

Ohm, Robin A., Robert Riley, Asaf Salamov, Byoungnam Min, In-Geol Choi, and Igor V. Grigoriev. 2014. Genomics of Wood-Degrading Fungi. *Fungal Genetics and Biology: FG & B 72* (November): 82–90. doi:10.1016/j.fgb.2014.05.001.

Olesnicky, Natalie S., Andrew J. Brown, Simon J. Dowell, and Lorna A. Casselton. 1999. A Constitutively Active G-Protein-Coupled Receptor Causes Mating Self-Compatibility in the Mushroom Coprinus. *The EMBO Journal* 18 (10): 2756–63. doi:10.1093/emboj/18.10.2756.

O'Shea, Suzanne F., Pushpalata T. Chaure, John R. Halsall, Natalie S. Olesnicky,

Andreas Leibbrandt, Ian F. Connerton, and Lorna A. Casselton. 1998. A Large Pheromone and Receptor Gene Complex Determines Multiple B Mating Type Specificities in*Coprinus cinereus*. *Genetics* 148 (3): 1081–90.

Raper, JR. 1966. *Genetics of Sexuality in Higher Fungi*. Ronald Press Company, New York. viii + 283 pp.

Raper, J. R., and A. S. Flexer, 1971 Mating systems and evolution of the Basidiomycetes, pp. 149–167 in Evolution in the Higher Basidiomycetes, edited by R. H. Petersen. University of Tennessee Press, Knoxville, TN.

Riquelme, Meritxell, Michael P. Challen, Lorna A. Casselton, and Andrew J. Brown. 2005. The Origin of Multiple B Mating Specificities in *Coprinus cinereus*. *Genetics* 170 (3): 1105–19. doi:10.1534/genetics.105.040774.

Sakagami, Youji, Mitsuru Yoshida, Akira Isogai, and Akinoiu Suzuki. 1981. Peptidal Sex Hormones Inducing Conjugation Tube Formation in Compatible Mating-Type Cells of *Tremella mesenterica*. *Science* 212 (4502): 1525–27. doi:10.1126/science.212.4502.1525.

Skrede, Inger, Sundy Maurice, and Håvard Kauserud. 2013. Molecular Characterization of Sexual Diversity in a Population of *Serpula lacrymans*, a Tetrapolar Basidiomycete. *G3: Genes|Genomes|Genetics* 3 (2): 145–52. doi:10.1534/g3.112.003731.

Stajich, Jason E., Sarah K. Wilke, Dag Ahrén, Chun Hang Au, Bruce W. Birren, Mark Borodovsky, Claire Burns, et al. 2010. Insights into Evolution of Multicellular Fungi from the Assembled Chromosomes of the Mushroom *Coprinopsis cinerea* (*Coprinus cinereus*). *Proceedings of the National Academy of Sciences* 107 (26): 11889–94. doi:10.1073/pnas.1003391107.

Stamatakis, A. 2006. RAxML-VI-HPC: Maximum Likelihood-Based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. *Bioinformatics* 22 (21): 2688–90. doi:10.1093/bioinformatics/btl446.

Van Peer, Arend F., Soon-Young Park, Pyung-Gyun Shin, Kab-Yeul Jang, Young-Bok Yoo, Young-Jin Park, Byoung-Moo Lee, Gi-Ho Sung, Timothy Y. James, and Won-Sik Kong. 2011. Comparative Genomics of the Mating-Type Loci of the Mushroom *Flammulina velutipes* Reveals Widespread Synteny and Recent Inversions. *PLoS ONE* 6 (7): e22249. doi:10.1371/journal.pone.0022249.

Wan, Jianing, Yan Li, Norihiro Shimomura, Takeshi Yamaguchi, and Tadanori Aimi. 2012. Characterization of DNA Polymorphisms in *Rhizopogon roseolus* Homeodomain Protein Genes and Their Utilization for Strain Identification. *Mycological Progress* 12 (2): 353–65. doi:10.1007/s11557-012-0840-z.

Wendland, J, L J Vaillancourt, J Hegner, K B Lengeler, K J Laddison, C A Specht, C A Raper, and E Kothe. 1995. The Mating-Type Locus B Alpha 1 of *Schizophyllum commune* Contains a Pheromone Receptor Gene and Putative Pheromone Genes. *The EMBO Journal* 14 (21): 5271–78.

Zerbino DR and Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* 18:821-829

Figure 3.1: Synteny map for A-locus region that shows the highly conserved ~130 Kbp region around the HD proteins of both *Rhizopogon* species and reference genomes. Each gene is represented by an arrow showing the direction of transcription. Gene color follows a heat map from yellow (gene represented in one genome) to red (gene represented in all genomes). HD1 genes are highlighted in blue, H2 in green, and the novel HD genes of C. cinerea are highlighted in light blue. Gene acronyms are in reference to James et al. (2004b). The bottom scale is in nucleotide basepairs.

Figure 3.1

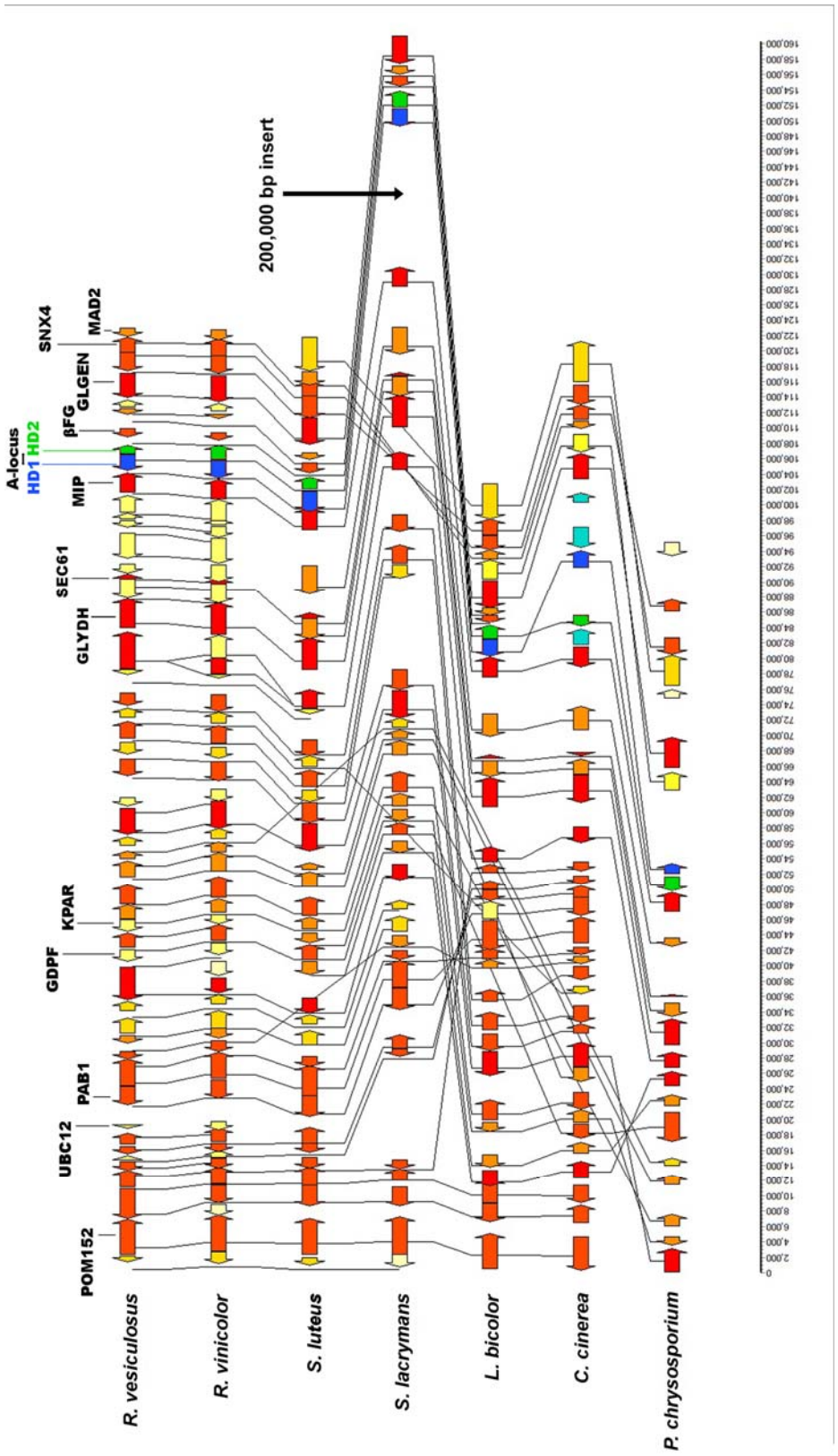

200,000 bp insert
A-locus
HD1 HD2
SNX4
MAD2
GLGEN
βFG
MIP
SEC61
GLYDH
KPAR
GDPF
PAB1
UBC12
POM152
R. vesiculosus
R. vinicolor
S. luteus
S. lacrymans
L. bicolor
C. cinerea
P. chrysosporium

Figure 3.2: Synteny map for A-locus region surrounding the HD proteins of *Rhizopogon* species and reference genomes. Each gene is represented by an arrow showing the direction of transcription. Gene color follows a heat map from yellow (gene represented in one genome) to red (gene represented in all genomes). Contig breaks for *R. vinicolor*, *R. vesiculosus,* and *S. luteus* are shown as vertical black bars. HD1 proteins are marked in blue and HD2 proteins are marked in green. A) Highlighted synteny of the conserved region shown in Figure 1 as well as the area at the far 3' end of the characterized region. B) Highlighted synteny of the region 3' of HD proteins in Boletales and Polyporales genomes which is translocated to the 5' end of HD proteins in Agaricales genomes. Note that this region is inverted in *S. lacrymans* relative to other Boletales genomes.

Figure 3.2:

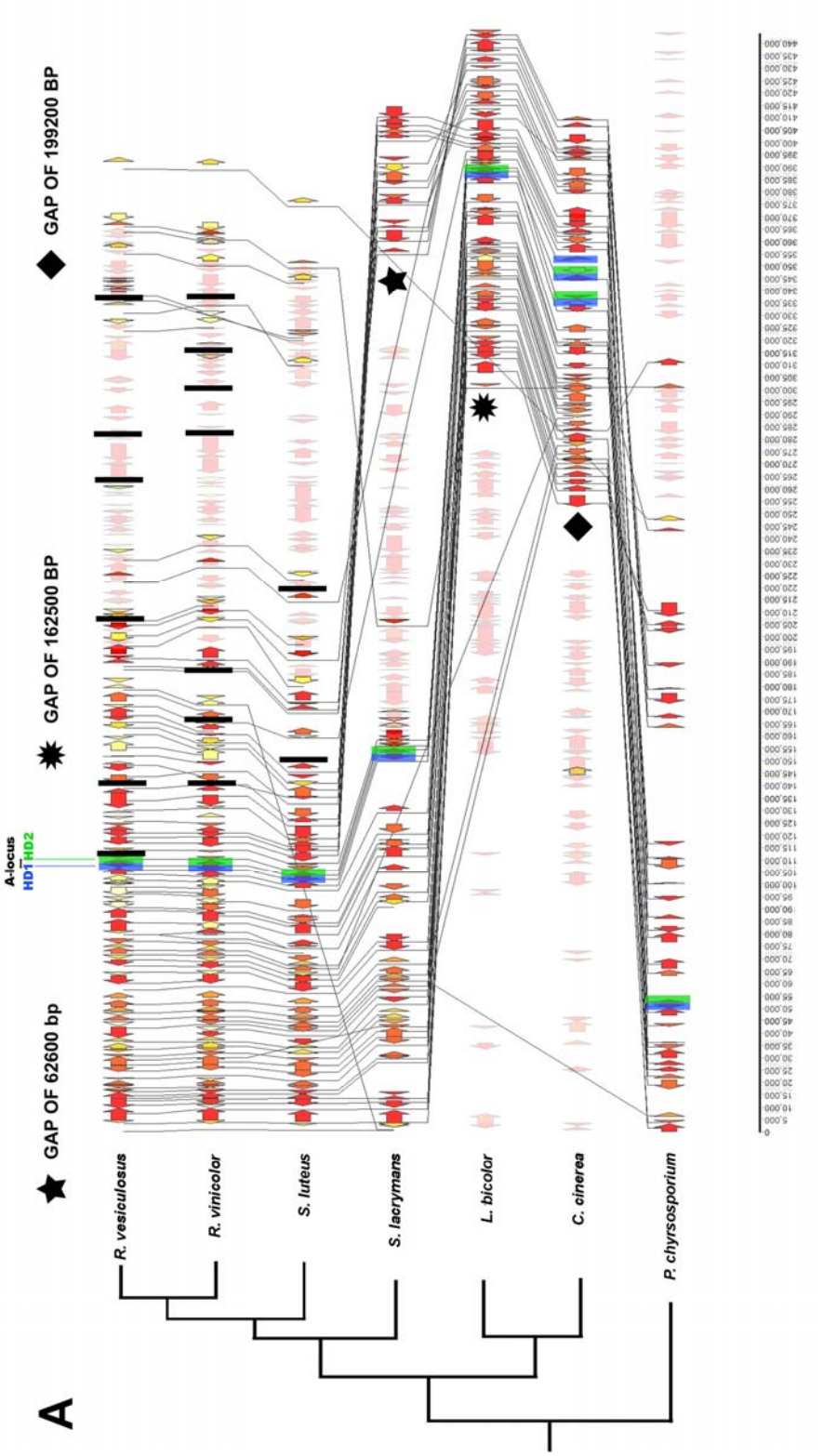

GAP OF 199200 BP
GAP OF 162500 BP
GAP OF 62600 bp
A-locus
HD1
HD2
R. vesiculosus
R. vinicolor
S. luteus
S. lacrymans
L. bicolor
C. cinerea
P. chyrsosporium
A

Figure 3.2: Continued

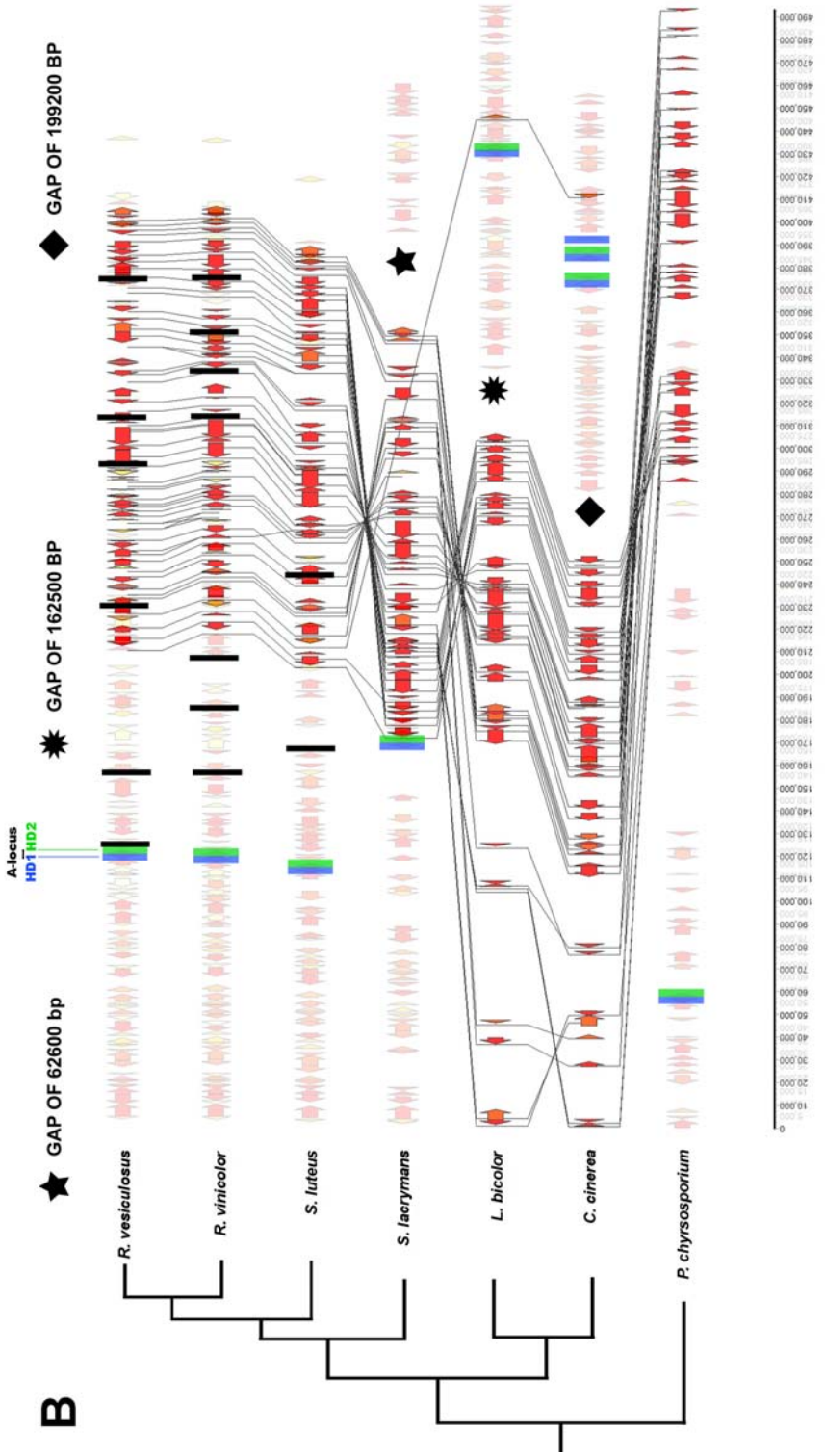


Figure 3.3: Synteny map for B-locus region surrounding the pheromone precursor and pheromone receptor genes identified in *Rhizopogon* genomes and the reference genomes *Suillus luteus, Suillus brevipes, Serpula lacrymans, Paxillus involutus, Coniophora puteana, Laccaria bicolor* and *Coprinopsis cinerea.* Pheromone receptor genes are shown in blue, pheromone precursor genes in green, and isoprenyl cysteine methyltransferase X genes (ICMTX) in light blue. A vertical black bar marks the alignment of two *P. involutus* contigs that are of ambiguous alignment and shown here on a single row for clarity of presentation. "*Paxillus involutus* 7" represents an area containing B-locus genes that has been transposed into a non-homologous genomic region at the center of scaffold 7 in the *Paxillus involutus* assembly. Red bars mark the location of breaks between *Rhizopogon* assembly contigs. Black stars mark *R. vinicolor* ICMT1 and ICMT2 genes that are duplicated at the terminal segment of two assembly contigs. Placement of these genes in *R. vinicolor* marks the boundary between two haplotype contigs. *C. puteana* is marked with a black diamonds at unique ICMTX genes and with parentheses at the location of a 130 kb non-homologous insert. The bottom scale is in nucleotide basepairs.

Figure 3.3

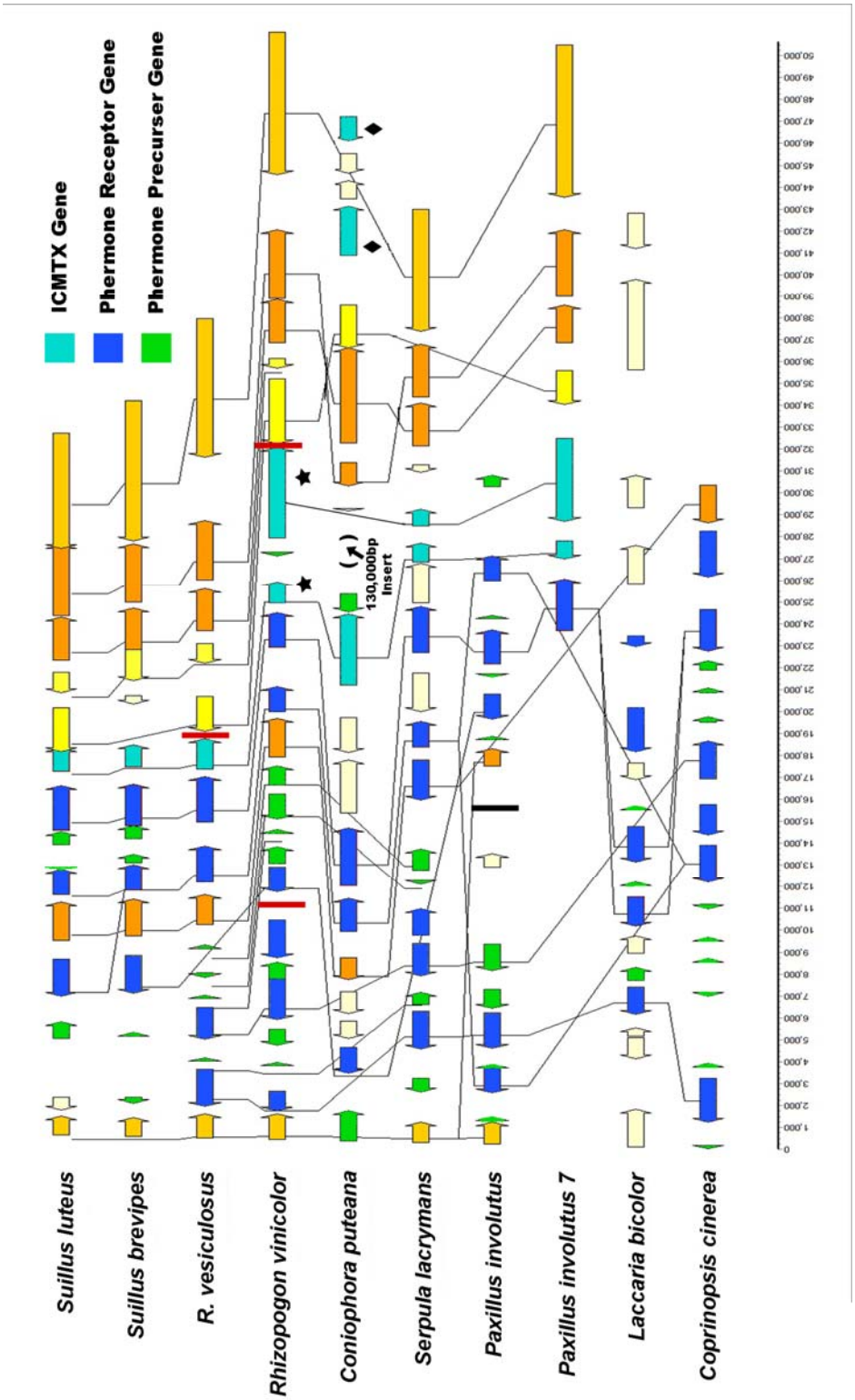
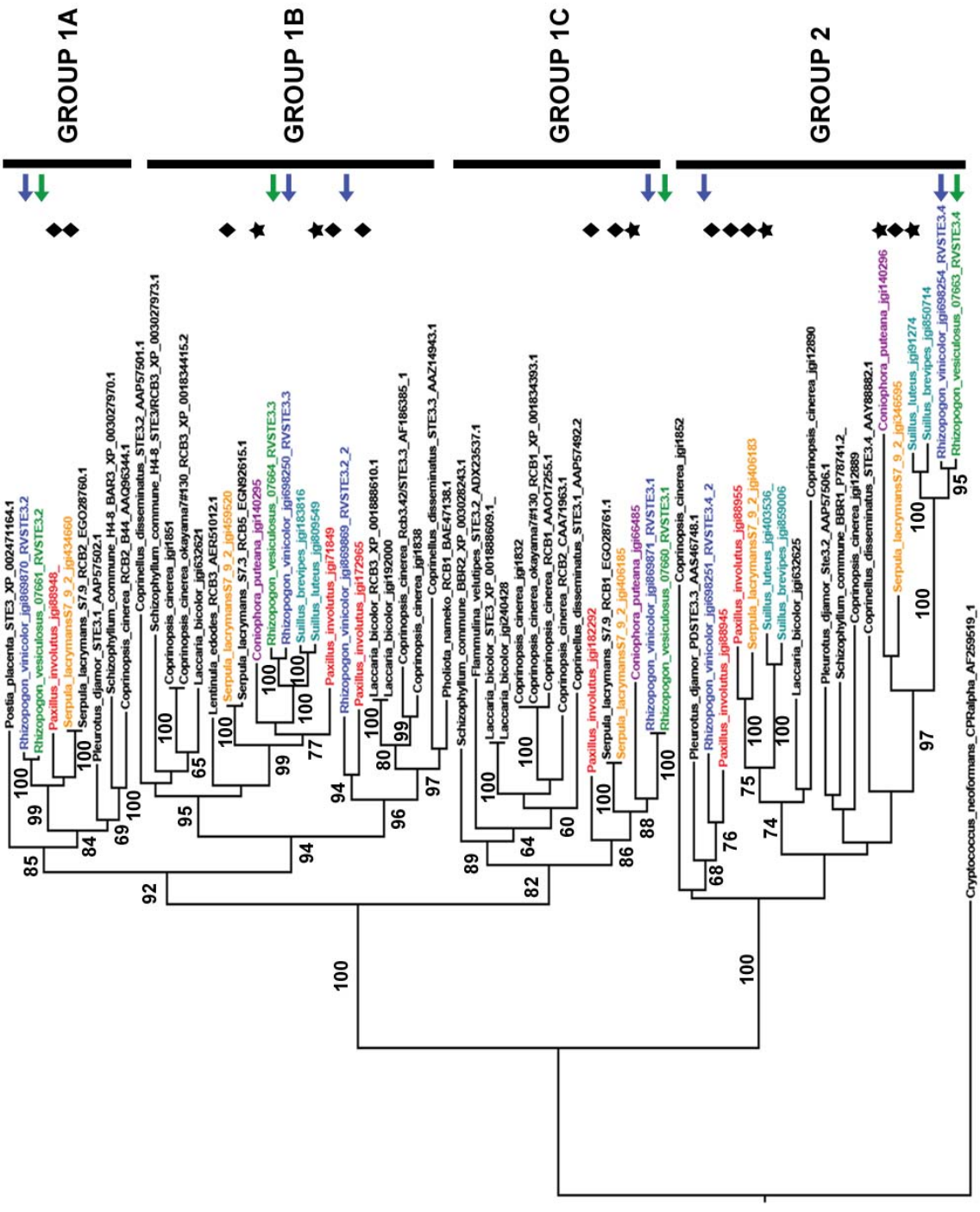

ICMTX Gene
Phermone Receptor Gene
Phermone Precurser Gene
130,000bp Insert
Suillus luteus
Suillus brevipes
R. vesiculosus
Rhizopogon vinicolor
Coniophora puteana
Serpula lacrymans
Paxillus involutus
Paxillus involutus 7
Laccaria bicolor
Coprinopsis cinerea

Figure 3.4: Maximum likelihood phylogram of pheromone receptor genes inferred using RAxML with 1000 bootstrap replicates. Bootstrap support values greater than 60% are shown for all branches. Pheromone receptor genes derived from genomes compared in figure 4 are color coded. *R. vinicolor* genes are marked by dark blue arrows and *R. vesiculosus* genes are marked by green arrows. Major groups of pheromone receptor genes as identified by James et al. (2004b) and Riquelme et al. (2005) are denoted at the right margin. Boletales taxa with known bipolar mating systems are marked with stars and those with known tetrapolar mating systems are marked with diamonds.

Figure 3.4


GROUP 1A
GROUP 1B
GROUP 1C
GROUP 2
Postia_placenta_STE3_XP_002471164.1
Rhizopogon_vinicolor_jgi869870_RVSTE3.2
Rhizopogon_vesiculosus_07661_RVSTE3.2
Paxillus_involutus_jgi88948_
Serpula_lacrymansS7_9_2_jgi434660
Serpula_lacrymans_S7.9_RCB2_EGO28760.1
Pleurotus_djamor_STE3.1_AAP57502.1
Schizophyllum_commune_H4-8_BAR3_XP_003027970.1
Coprinopsis_cinerea_RCB2_B44_AAQ96344.1
Coprinellus_disseminatus_STE3.2_AAP57501.1
Schizophyllum_commune_H4-8_STE3/RCB3_XP_003027973.1
Coprinopsis_cinerea_jgi1851
Coprinopsis_cinerea_okayama7#130_RCB3_XP_001834415.2
Laccaria_bicolor_jgi632621
Lentinula_edodes_RCB3_AER51012.1
Serpula_lacrymansS7_9_2_jgi459520
Serpula_lacrymans_S7.3_RCB5_EGN92615.1
Coniophora_puteana_jgi140295
Rhizopogon_vesiculosus_07664_RVSTE3.3
Rhizopogon_vinicolor_jgi698250_RVSTE3.3
Suillus_brevipes_jgi163816
Suillus_luteus_jgi809549
Paxillus_involutus_jgi71849
Rhizopogon_vinicolor_jgi869869_RVSTE3.2_2
Paxillus_involutus_jgi172965
Laccaria_bicolor_RCB3_XP_001888610.1
Laccaria_bicolor_jgi192000
Coprinopsis_cinerea_Rcb3.42/STE3.3_AF186385_1
Coprinopsis_cinerea_jgi1838
Coprinellus_disseminatus_STE3.3_AAZ14943.1
Pholiota_nameko_RCB1_BAE47138.1
Schizophyllum_commune_BBR2_XP_003028243.1
Flammulina_velutipes_STE3.2_ADX23537.1
Laccaria_bicolor_STE3_XP_001888609.1_
Laccaria_bicolor_jgi240428
Coprinopsis_cinerea_jgi1832
Coprinopsis_cinerea_okayama7#130_RCB1_XP_001834393.1
Coprinopsis_cinerea_RCB1_AAO17255.1
Coprinopsis_cinerea_RCB2_CAA71963.1
Coprinellus_disseminatus_STE3.1_AAP57492.2
Paxillus_involutus_jgi182292
Serpula_lacrymans_S7.9_RCB1_EGO28761.1
Serpula_lacrymansS7_9_2_jgi406185
Coniophora_puteana_jgi66485
Rhizopogon_vinicolor_jgi869871_RVSTE3.1
Rhizopogon_vesiculosus_07660_RVSTE3.1
Coprinopsis_cinerea_jgi1852
Pleurotus_djamor_PDSTE3.3_AAS46748.1
Rhizopogon_vinicolor_jgi698251_RVSTE3.4_2
Paxillus_involutus_jgi88945
Paxillus_involutus_jgi88955
Serpula_lacrymansS7_9_2_jgi406183
Suillus_luteus_jgi403536_
Suillus_brevipes_jgi859006
Laccaria_bicolor_jgi632625
Coprinopsis_cinerea_jgi12890
Pleurotus_djamor_Ste3.2_AAP57506.1
Schizophyllum_commune_BBR1_P78741.2_
Coprinopsis_cinerea_jgi12889
Coprinellus_disseminatus_STE3.4_AAY88882.1
Coniophora_puteana_jgi140296
Serpula_lacrymansS7_9_2_jgi346595
Suillus_luteus_jgi91274
Suillus_brevipes_jgi850714
Rhizopogon_vinicolor_jgi698254_RVSTE3.4
Rhizopogon_vesiculosus_07663_RVSTE3.4
Cryptococcus_neoformans_CPRalpha_AF259519_1
*

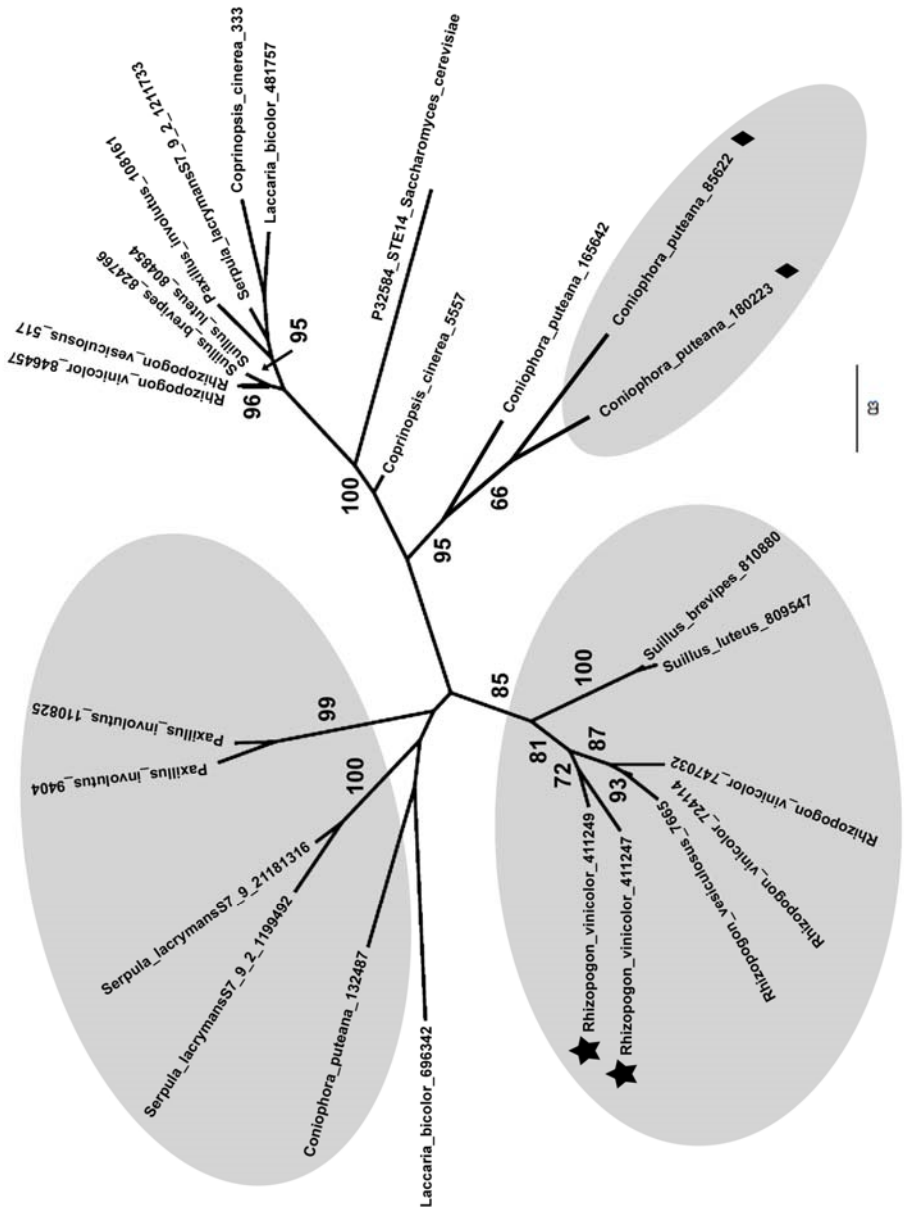


Figure 3.5: Maximum likelihood unrooted phylogram of isoprenyl cysteine methyltransferase genes (ICMT) inferred using RAxML with 500 bootstrap replicates. Bootstrap support values of 60% or greater are shown for all branches. ICMT genes were identified as either having close proximity to the B-locus regions of the examined genomes or as sharing high BLAST sequence identity with the STE14 gene of *Saccharomyces cerevisiae*. Genes identified in proximity to the B-locus regions highlighted by gray circles. The two alleles of the ICMTX genes which were present at a heterozygous break in *R. vinicolor* B-locus contigs are marked by stars. The two unique ICMTX genes in *C. puteana* that are associated with a 130 kb translocation away from the main B-locus gene cluster are marked with diamonds.

Table 3.1: Summary of assembly statistics for the *Rhizopogon vinicolor* and *Rhizopogon vesiculosus* genomes

Species	Minimum Contig Size Retained (bp)	Assembly Length (Mbp)	Percent of Assembly in Genic Regions	Average Read Coverage Depth	Largest Scaffold/Contig (Mbp)	Number of Scaffolds	Number of Contigs	Scaffold/ Contig N50	Raw N50	Number of Gene Models
R. vinicolor	1000	36.1	63.90%	154X	0.56	2310	3018	218	43737	14469
R. vesiculosus	1000	42.2	48.90%	51X	0.45	-	3765	361	30999	13623
R. vesiculosus	300	43.8	47.60%	66X	0.45	-	6700	388	28895	14218

Table 3.2: Predicted amino acid sequence of pheromone precursor genes identified in *Rhizopogon* genomes. The predicted N terminal couplet of mature pheromones and the C terminal –CAAX motif is presented in underlined boldface for each sequence.

Rhizopogon vesiculosus predicted pheromone precursors

Rvphb2.1	MDSFDTFHTLEFPLDAPPDIGVEPESSSSVPLD**ED**SQYWRPGTF**CIIS**
Rvphb3.1	MDMFSSLSSISLEELDIVHTPSSSSSSYSLLAHSDGDSCPIPTEY**EH**SYSASSGWY**CIIA**
Rvphb3.2	MDAFDTILISELFQPDPPLSLDQGSCSSDDEPYLLVDA**DA**KWNYGGY**CVIS**
Rvphb3.3	MDEFVTLPSEPSSFAELSDSGDALQFEYSDALPVDD**DR**LSGYYGSF**CVII**

Rhizopogon vinicolor predicted pheromone precursors

RVphb3.5	MDNFTSVSDIFFPTVSLEGDPSNSGDPNAVGDGEPVLVDA**EY**WTWGSSSQSGF**CVIV**
RVphb2.1	MDSFDTFHIPEFPLNAPPNPGVEPESSSSIPLD**ED**SQYWRPGTF**CIIS**
RVphb2.2	MNNGENTIPFNSGPLCPRSAPLKGQPFSGLIHHTRELITQAQTYCDLTHCDTDYTSNF
	TLRKTLITTIYSMDAFASIAELFSQSISLSEVTNIAPPDDHEMPVEY**ER**NPLSAGGF**CIIS**
RVphb3.3	MDEFITLPSDPSSFAELSHSGNASQFEYIAYSDTLPVDD**DR**LSGYYGSF**CVII**
RVphb2.2	MDNFTSVSDIFFPTVSLEGDPSNSGDPNAVGDGEPVLVDA**EY**WTWGSSSQSGF**CVIV**
RVphb3.3	MDQFITLPSEPSSFAELSHSGDASQFDVEYSDALPVDD**DR**LSGYYGSF**CVII**
RVphb3.1	MDIFSSLSSISLDELDIAHTPSSSSSYSRLAHSDDDACPIPTEY**EH**SYSASSGWY**CIIA**
RVphb3.2	MDAFDTILVSELFQPDPPLSLDQGSCSSDDEPYLLV**DA**DTKWNYGGY**CVIS**
RVphb3.4	MDIFSSLSSISLDELDIAHTPSSSSSYSRLAHSDDDACPIPTEY**EH**SYSASSGWY**CIIA**

Table 3.3: The Boletales genomes examined at the B-locus. The total numbers of pheromone receptor and pheromone precursor genes are listed. B-locus assembly statistics, ecological habit, mating system and reference literature for each genome is also listed.

Species	Pheromone Receptor Genes	Pheromone Precursor Genes	Number of Contigs Spanning B-locus	Mating System	Ecological Habit	Reference
Rhizopogon vesiculosus	4	4	2	-	Ectomycorrhizal	This study
Rhizopogon vinicolor	6	9	3	-	Ectomycorrhizal	This study
Suillus luteus	3	5	2	Bipolar	Ectomycorrhizal	Kohler et al. 2015
Suillus brevipes	3	4	2	-	Ectomycorrhizal	This study
Coniophora puteana	3	3	1	Bipolar	Saprobic	Floudas et al. 2013
Paxillus involutus	5	7	3	Tetrapolar	Ectomycorrhizal	Kohler et al. 2015
Serpula lacrymans S7.9	5	4	1	Tetrapolar	Saprobic	Eastwood et al. 2011

Table 3.4: Summary of SNP mutations identified between the haplotypes of dikaryotic *Rhizopogon* genome assemblies. Percent values for SNPs in columns 3 – 7 denote the percent represented by that column of all SNP mutations identified within a genome.

Species	Minimum contig size of assembly	Total SNPs	Percent of genome in SNPs	SNPs in intergenic regions	SNPs in genic regions	SNPs in Exons	Synonmous Exonic SNPs	Non-Synonymous Exonic SNPs
Rhizopogon vinicolor	1000	392731	1.10%	323677 / 82.4%	69054/ 17.6%	46213 / 11.8%	21219 / 5.4 %	24994 / 6.4%
Rhizopogon vesiculosus	1000	229571	0.54%	203465 / 88.6%	26106 / 11.4%	17020 / 7.4%	8314 / 3.6%	8706 / 3.8%
Rhizopogon vesiculosus	300	251906	0.60%	222588 / 88.4%	29318 / 11.6%	19512 / 7.8%	9298 / 3.7%	10214 / 4.1%

Chapter 4. Out of Western North America: systematics and phylogeography of *Rhizopogon* subgenus *Villosuli* based on genome-scale sequence typing

Alija Bajro Mujic, Joseph W. Spatafora

Target Journal: Molecular Ecology or PNAS

ABSTRACT:

The shared life history traits shared species distributions with hosts of ectomycorrhizal (EM) fungi suggest that comigration and cospeciation have had an influential role in the evolution of the EM symbiosis. However, the generalist nature of many EM fungi and regional host effects upon EM community structure pose difficulties in the study of these phenomena and they have currently been documented from only a limited number of systems. In order to test the hypotheses of comigration and cospeciation between EM symbionts we have conducted a genome-scale phylogeographic analysis of host specific EM fungi in genus *Rhizopogon* that associate with trees in genus *Pseudotsuga.* Four years of multinational, intercontinental sampling efforts throughout the northern Pacific Rim have produced *Rhizopogon* samples from throughout the natural range of all *Pseudotsuga* species. We have used Illumina sequencing technology to produce a genome-scale phylogenetic dataset for *Rhizopogon* species collected throughout our sampling range. The results of phylogenetic analysis strongly support a single evolutionary origin of the EM symbiosis between *Rhizopogon* and *Pseudotsuga* in western North American and phylogeographic trends in the distribution of *Rhizopogon* species show concordance with the population structure and species divergence events of their *Pseudotsuga* hosts. Our results support the hypothesis that the *Rhizopogon-Pseudotsuga* EM symbiosis predates species radiation in *Pseudotsuga* and that these genera have undergone processes of comigration and cospeciation. We detail the process of genome-scale sequence typing (GSST), which can be adapted for any set of taxa for which a large number of orthologous gene clusters can be identified.

INTRODUCTION:

Ectomycorrhizal (EM) fungi are a large component of fungal diversity throughout temperate and boreal forest regions and it is in these regions of greater latitude that the greatest diversity of EM fungi is found (Tedersoo et al. 2012, Tedersoo et al. 2014). This trend is contrary to the pattern of most terrestrial and marine macroorganisms, which display an inverse relationship between latitude and diversity (Hillebrand 2004). A recent global survey of fungal diversity in soil demonstrates that EM fungal diversity at greater latitudes is correlated with the distribution patterns and density of EM plant taxa, most particularly trees of the family Pinaceae, which are widely distributed in higher latitudes of the northern hemisphere (Tedersoo et al. 2014). Many of the plant and fungal symbionts in EM relationships are reliant upon one another for successful establishment (Smith and Read 2008, Horton et al. 1999) and it would be expected that EM fungi should share similar patterns of migration and dispersal with host plants. Some examples of this trend have been observed as in the case of the EM truffle *Tuber melanosporum,* which displays a population structure consistent with postglacial comigration with host trees in Western Europe (Murat et al. 2004). However, signals of comigration and cospeciation in EM symbioses are complicated by the fact that many EM fungal species are host generalists (Horton and Bruns 1998, Horton et al. 1999, Bruns et al. 2002) and that variable host plant community assemblage across biogeographic regions strongly affects EM fungal community assemblage (Ishida et al. 2007, Tedersoo et al. 2008, Bahram et al. 2012) and species divergence (Geml et al. 2008).

Patterns of EM plant and fungal comigration and cospeciation should be most evident in systems where either the fungus or plant possesses high specificity for a particular EM partner. In cases of extreme specificity, such as that displayed by achlorophyllous mycoheterotrophic plants, speciation of the plant has been observed to track that of the fungal mycorrhizal partner (Merckx and Bidartondo 2008) and the range of the parasitic plant can be strictly limited to mirror that of a few fungal taxa (Cullings et al. 1996, Bidartondo and Bruns 2002). Comigration between EM plant host and fungal partner has been well documented between trees of the genus *Alnus* (Alder) and their EM fungal parnters. *Alnus* species are broadly distributed from tropical to boreal climates and are hosts to both EM fungi and Actinomycetes which provide the plant with mineral nutrition and fixed nitrogen, respectively (Yamanaka et al. 2003). *Alnus* species associate with a highly specific set of EM fungal symbionts many of which do not associate with other EM hosts (Molina 1981, Kennedy et al. 2011,Tedersoo et al. 2009, Põlme et al. 2013). The structure of *Alnus* associated EM fungal communities varies along a geographic gradient thought to trace historical migration of Latin American *Alnus* species (Kennedy et al. 2011) and there is some evidence for cospeciation of host specific EM fungal lineages and *Alnus* subgenera in Europe (Rochet et al. 2011). On a global scale *Alnus* EM community structure is most similar between more closely related *Alnus* species regardless of geographic distribution (Põlme et al. 2013) and this suggests complex comigratory patterns between *Alnus* species and their fungal symbionts.

Studies of *Alnus* EM biogeography and cospeciation have greatly expanded evolutionary knowledge of host specific EM mutualisms but these phenomenon have yet to be investigated in a single lineage of closely related EM fungi that share a broad

distribution and with their preferred host. The symbiosis shared between the fungal genus *Rhizopogon* and trees of genus *Pseudotsuga* present an ideal system for such an investigation. *Rhizopogon* and the closely related genus *Suillus* display some of the highest rates of host specificity of all EM fungi (Molina and Trappe 1982, Molina and Trappe 1994) and this shared trait indicates a long evolutionary history of EM host specificity in this group of fungi. Genus *Rhizopogon* forms ectomycorrhizae almost exclusively with trees in family Pinaceae and host specificity for particular host genera or species is further restricted within *Rhizopogon* subgenera (Molina et al. 1994, Grubisha et al. 2002). The work of Grubisha et al. (2002) established the current subgeneric classifications of genus *Rhizopogon* and found that the EM host relationship with *Pseudotsuga menziesii* was restricted to a single clade defining *R.* subgenus *Villosuli.* This finding suggested that the *Rhizopogon-Pseudotsuga* relationship might possess a single evolutionary origin. However, the ecological synapomorphy of the *Rhizopogon-Pseudotsuga* symbiosis in *R.* subgenus *Villosuli* could not be confirmed because the taxonomic sampling of this study was restricted to North American *Rhizopogon* specimens associated only with coastal *P. menziesii var menziesii* and interior *P. menziesii var glauca*. Recent work in Japan has extended the *Rhizopogon-Pseudotsuga* EM symbiosis to include an additional species of *R.* subgenus *Villosuli*, *R. togasawariana* (Mujic et al. 2014), which occurs in EM symbiosis with *P. japonica* and is a common member of the resistant propagule community in the soil of *P. japonica* stands (Murata et al. 2013, Dr. Kazuhide Nara personal communication). Evidence of *R.* subgenus *Villosuli* in association with geographically distant *Pseudotsuga* species provides strong support for the hypothesis that the EM symbiosis between these genera evolved once in

association with a common ancestor of *Pseudotsuga* species and *R.* subgenus *Villosuli*. However, additional sampling of the remaining *Pseudotsuga* species is required to demonstrate the complete specificity of this relationship and to support comigration between these genera.

The natural range of *Pseudotsuga* extends around the northern Pacific Rim from central Mexico to southern China and northern Vietnam (Farjon 2010). The taxonomy of *Pseudotsuga* in Asia and in Mexico has been debated and modern treatments of the genus recognize between four (Farjon 2010) and eight (Hermann 1982) species within the genus. Here we will discuss *Pseudotsuga* using the species concepts of Gernandt and Liston (1999) and refer to five species of Taiwanese (*P. wilsoniana*), Chinese (*P. sinensis*), Japanese (*P. japonica*), and western North American (*P. macrocarpa*, *P. menziesii*) distribution. Molecular phylogenies of the genus all support Asian species and North American species as sister lineages (Strauss et al. 1990, Gernandt and Liston 1999, Wei et al. 2010), but these phylogenies alone do not conclusively determine the geographic origin of the genus. However, the older fossil records of *Pseudotsuga* in North America (Hermann et al. 1985, Yabe 2011) as well as similar RAPD profiles between North American *Pseudotsuga* species with North American larch, *Larix occidentalis*, and *P. japonica* suggest a North American origin for the genus with subsequent step-wise species radiation into East Asia (Strauss et al. 1990).

Molecular phylogenies of the internal transcribed spacer (ITS) region of *Rhizopogon* species support a subgeneric structure consisting of five distinct *Rhizopogon* subgenera: *Amylopogon, Rhizopogon, Roseoli, Versicolores,* and *Villosuli* (Molina et al. 1999, Grubisha et al. 2002). Phylogenetic analysis of the ITS and ribosomal large

subunit (LSU) region have also been used to clarify some of the species concepts within subgenus *Villosuli* (Bubriski and Kennedy 2014). However, support for the systematic relationships between subgenera and between species within R. subgenus *Villosuli* are lacking from these analyses. The goals of this study are to investigate the intrageneric systematics of *Rhizopogon* and to test the hypothesis of comigration between *Pseudotsuga* and *R.* subgenus *Villosuli*. We hypothesize that the *Rhizopogon-Pseudotsuga* symbiosis has evolved only once in a common ancestor of *R.* subgenus *Villosuli* and that members of *R.* subgenus *Villosuli* have undergone a process of comigration and cospeciation with their *Pseudotsuga* hosts. To test these hypotheses we have conducted a multicontinental sampling of *Rhizopogon* across the range of all *Pseudotsuga* species and conducted a genome scale phylogeographic analysis of *Rhizopogon* species recovered in this sampling. Our results provide strong evidence for comigration and cospeciation between an EM host and fungal symbiont and conclusively resolve systematic relationships in genus *Rhizopogon.*

MATERIALS AND METHODS:

Field Collection and Expedition localities:

Field collections of *Rhizopogon* basidiomata and EM root tips were made over the span of four years in stands of *Pseudotsuga* species in Japan, Taiwan, China, Mexico and the western United States. The season for collection in each site was chosen to coincide either with the rainy season or within one month of the end of the last rains of the season. A detailed list of collection localities can be found in Table 4.1. *Pseudotsuga* populations

at the limits of dispersal for the genus are fragmented irregularly upon the landscape (Farjon 2010). Because of the relative rarity of some *Pseudotsuga* species the localities of field sites were determined during the season preceding field expeditions with the assistance of state and federal forest managers as well as local botanists. Historical collection data from herbarium vouchers of *Pseudotsuga* species was also instrumental in locating the exact location of many field sites. Collections of *Rhizopogon* basidiomata were made by gently raking the surface litter from soil in the vicinity of *Pseudotsuga* trees with further raking of the upper 10 cm of soil to uncover basidiomata. Each basidiome was catalogued and a small section (1 mm^3) of clean tissue was taken from the interior gleba tissue for in-vitro culture and a second sample (5 mm^3) was taken for storage in 2% CTAB buffer until DNA extraction could be performed. All basidiomata were dried at 37 C after processing and within 24 hours of collection. EM root tips were collected from *Pseudotsuga* forests in order to positively associate *Rhizopogon* species with their *Pseudotsuga* hosts. Roots were collected from within the drip line of *Pseudotsuga* trees both by raking and by the use of a digging trowel. Fine root samples of approximately 10 cm in length and no greater than 5 mm diameter were collected. Roots were stored in plastic bags along with a small amount of forest litter, moistened, and stored at 4 C for no longer than 2 weeks until processing. Root samples were washed in deionized water and individual root tips were dissected from fine roots under a dissecting microscope. Each root tip was stored in 2% CTAB buffer until DNA extraction could be performed.

Tissue culture, tissue harvest, and DNA extraction:

Rhizopogon tissue cultures were derived directly from dikaryotic basidiomata using sterile technique in field laboratory conditions. Tissue from the gleba of each basidiome was transferred to modified Melin-Norkrans (MMN) agar media supplemented with broad spectrum antibiotics (Streptomycin 5 pmm, Ampicillin 5 pmm) and incubated at 20°C until a diameter of approximately 2 cm was achieved. Small wedges (1 mm^3) of actively growing mycelium were then transferred to MMN media topped with sterile cellophane membrane (BioRad, Hercules, CA) and allowed to grow at 20°C to a diameter of approximately 2 cm. Mycelial tissue was harvested under sterile conditions by first excising and discarding the agar wedge used as inoculum and then peeling the mycelium from the cellophane membrane. Harvested mycelium was flash frozen in liquid nitrogen, lyophilized, and stored at -80°C until DNA extraction could be performed. DNA extraction followed a modified CTAB/Phenol/Chloroform extraction protocol as described in chapter 3 of this dissertation with the following modifications. Lyophylized tissue was chilled in liquid nitrogen and ground in a sterile 1.5 ml microcentrifuge tube using an autoclaved micropestle. Approximately 50 µl of finely ground tissue was taken into each DNA extraction and following the RNAse treatment and final precipitation DNA pellets were rehydrated in molecular grade water to a final volume of 30 – 50 µl. In order to produce DNA extraction of sufficient mass and quality for Illumina sequencing we then combined 3 – 5 extracts of each individual into a single pool and brought the final volume of combined extracts to 250 µl in molecular grade water. Each combined extract was mixed 1:1 with Guanidine Hydrochloride (GHCL) buffer and subjected to an additional phenol/chloroform purification and ethanol precipitation following the

protocol of Falcão et al. (2007). Final extracts were quantified using a Qubit flourometer (Life Technologies, Grand Island, NY) and quality was accessed by visualizing extracted DNA under UV transillumination upon 2% agarose electrophoresis gels stained with ethidium bromide. DNA was extracted from CTAB preserved *Rhizopogon* basidiomata and EM root tips by the same protocol but without the additional GHCL buffer purification step.

PCR and Sanger sequencing:

We sequenced the internal transcribed spacer region (ITS) of all *Rhizopogon* basidiomata, cultures, and EM root tips in order to positively identify each sample and to include ITS sequence data in phylogenetic datasets. For EM root tips we also sequenced the matK chloroplast gene if there was any uncertainty concerning the identity of the host tree. ITS and matK sequences were identified to generic level by BLAST comparison to the GENBANK database available at the website of the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov). PCR reactions were conducted under various thermocycler conditions using either Phire Hot Start II DNA Polymerase (Thermo Fisher Scientific, Waltham, MA) or a combination of Taq Polymerase (Genescript, Piscataway, NJ) and PCR Optimization Buffer E (Epicentre, Madison, WI). All PCR reactions were performed in 25 µl volumes with each primer at 8.0 µM concentration. Other reagents were added to make 25 µl following polymerase manufacturer protocols. PCR reactions targeting the matK region used the primer pair matK_EquisetumF (RBG Kew, 5'-ATACCCCATTTTATTCATCC -3') and matK_2F_R (matK primer Plant Working Group, 5'- CGTACTTTTATGTTTACAGGCT -3') and

those targeting the ITS region used the forward primer ITS1F and either ITS4 or ITS4B as the reverse primer (White et al. 1990, Gardes and Bruns 1993).

Illumina library preparation, sequencing and assembly:

Illumina sequencing libraries were prepared and sequenced for 25 *Rhizopogon* cultures representing the range of geographic distribution and taxonomic diversity encountered for *R.* subgenus *Villosuli* during field collections. In addition, a single representative culture from each of the other four *Rhizopogon* subgenera (*Rhizopogon, Versicolores, Roseoli*, and *Amylopogon*) to allow testing of systematic hypotheses across all of genus *Rhizopogon*. Illumina libraries were constructed using 1 – 5 μg of high molecular weight DNA as starting material. In cases where 1 μg of DNA was not available libraries were constructed using at least 0.5 μg of DNA with half reaction volumes at all steps of the protocol. Initial shearing of genomic DNA was performed by sonication at 4°C using a Biorupter Pico device (Diagenode, Denville, NJ) with 10 – 12 cycles of 30 seconds shearing followed by 30 seconds of rest. Libraries were then prepared using the NEBnext DNA Mastermix Kit for Illumina and indexed for multiplexed sequencing using the NEBnext dual indices kit for Illumina (New England Biolabs, Ipswich, MA) following manufacturer protocols. All sample concentration, cleanup, and size-selection steps of Illumina library construction were performed using a homemade mix of Carboxyl-modified Sera-Mag Magnetic Speed-beads (Thermo Fisher Scientific, Waltham, MA) in a PEG/NaCl buffer following the protocol of Rohland and Reich (2012). Low coverage Illumina sequencing was performed with 9 – 10 samples multiplexed per lane on an Illumina HiSeq 2000 at the Oregon State

University Center for Genomic Research and Biocomputing (CGRB) (http://cgrb.oregonstate.edu). Raw Illumina read data was quality filtered using the fastx tool kit (http://hannonlab.cshl.edu/fastx_toolkit/) and custom Perl scripts. De novo assembly of quality-filtered reads was performed using VELVET 1.19 (Zerbino and Birney 2008).

Identification of conserved protein coding regions:

In order to identify protein coding regions of phylogenetic utility in genus *Rhizopogon* we performed Markov clustering of protein models from the *R. vinicolor* AM-OR11-026 and *R. vesiculosus* AM-OR11-056 genome assemblies described in chapter 3 as well as two reference proteomes of *Suillus luteus* and *S. brevipes* available from the United States Department of Energy Joint Genome Institute (DOE-JGI) MycoCosm (Grigoriev et al. 2014). *Suillus* is a genus of mushroom forming fungi closely related to *Rhizopogon* (Bruns et al. 1989) and the set of conserved genes in *Rhizopogon* is expected to subsume the set of conserved genes between these genera. Markov clustering of protein models into protein families was performed in FastOrtho (Gillespie et al. 2011). We then used custom Perl script (see Appendix 1 for code) and BioPerl modules (Stajich et al. 2002) to identify gene families of single copy protein coding genes by screening the output the FastOrtho 4 taxon statement for clusters that contained a single protein model for each species. Our script used protein coordinates in gff3 format (http://www.sequenceontology.org/gff3.shtml) from the DOE-JGI "best filtered models" for *S. luteus, S. brevipes,* and *R. vinicolor* and MAKER (Cantarel et al. 2008) generated protein models for *R. vesiculosus*. Coordinates and strand information for each protein

was used to mine the nucleotide sequence data from the draft genome assemblies of *R. vinicolor, R. vesiculosus*, and *S. luteus*. Our script filtered protein models for use in phylogenomic analysis by retaining only those proteins that possessed a total exonic length of at least 960 bp with all exons exceeding a length of 120 bp each following the algorithm described by Weitemier et al. (2014).

Phylogenetic dataset construction:

All single copy protein coding genes identified by FastOrtho and retained by our phylogenetic utility filter were included into alignments for use in phylogenomic analysis. Nucleotide sequences of gene orthologs were identified in low coverage *Rhizopogon* genomes by BLASTN queries of protein orthologs identified from *R. vinicolor* with an e-value threshold of 1e-5. We designed Perl script (see Appendix 1 for code) to mine sequence data from draft *Rhizopogon* assemblies using the coordinates of the top scoring BLAST hit to each *R. vinicolor* protein model. Our script determines the strand and the maximum and minimum coordinates of all high scoring positions (HSP) within the best hit; it then returns the corresponding sequence data spanning all HSPs from the draft assembly. Sequence data from each assembly and from the *R. vinicolor* reference are collated into FASTA format files containing the nucleotide sequence of each ortholog for all samples. An additional Perl script (see APPENDIX 1 for code) was then used to insert sequence orthologs from *R. vesiculosus* and *S. luteus* into these FASTA files. Each ortholog FASTA file was aligned using MUSCLE (Edgar 2004) and then trimmed of uninformative alignment positions using the "gappyout" heuristic implemented in TrimAL (Capella-Gutiérrez et al. 2009). Final alignments were

concatenated into a single super alignment using custom Perl script (see APPENDIX 1 for code).

Samples collected during field expeditions in Japan and China were not successfully cultured. We successfully retrieved *Rhizopogon* basidiomata from Japan and EM root tips from China but tissue from these samples were not appropriate for low coverage Illumina sequencing because of the high proportion of DNA present from contaminating fungi and bacteria. In lieu of Illumina sequencing we instead collected sequence data from Japanese basidiomata and Chinese EM root tips using a PCR based approach. We identified 10 protein models of 1000 – 1100 bp length from the set of protein coding models identified by our phylogenetic quality filter script. Protein models were chosen from the potential set of appropriate lengths based upon the quality of JGI annotation and the number of orthologous proteins identified through BLAST searches. PCR primers for these regions were designed using the online primer design tool Primer3 (http://biotools.umassmed.edu/bioapps/primer3_www.cgi) (Rozen and Skaletsky 1999). In addition we also collected PCR sequence data for the ribosomal large subunit (LSU) and translation elongation factor 1-alpha (TEF) using the primer pairs ITS1F – LR5 (Gardes and Bruns 1993, Vilgalys and Hester 1990) and 983F - 2218R (Rehner and Buckley 2005), respectively. PCR conditions and reagents for the 10 protein coding regions, LSU, and TEF were the same as those described for ITS and matK.

Phylogenetic and Phylogeographic analyses:

We conducted phylogenetic analysis of all sequence alignment datasets using the maximum likelihood algorithm implemented in RAxML 7.04 (Stamatakis 2006). The phylogeny inferred for ITS sequence dataset used the GTRGAMMA substitution model and 1000 bootstrap replicates. Individual alignments for single copy protein coding genes as well as the entire concatenated dataset were analyzed using the GTRGAMMA substitution model and 100 bootstrap replicates. In order to investigate gene tree – species tree incongruence we analyzed the set of all individual gene trees produced for single copy protein coding genes using the maximum-pseudolikelihood coalescent model in MP-EST (Liu et al. 2010) as implemented on the STRAW webserver (Shaw et al. 2013). The geographic affiliations of the common ancestors of *R.* subgenus *Villosuli* were investigated by performing ancestral state reconstruction (ASR) upon the final super-alignment phylogeny using the maximum likelihood criterion in Mesquite 3.02 (Maddison and Maddison 2015).

RESULTS:

Field collections and selection of specimens for Illumina sequencing:

We successfully collected *Rhizopogon* basidiomata from *Pseudotsuga* forests in Japan, Taiwan, the Mexican states of Hidalgo, Tlaxcala, and Puebla and the American states of California, Oregon, Utah, and Arizona. Across these collection localities a total of 143 *Rhizopogon* basidiomata belonging to *R.* subgenus *Villosuli* and 8 basidiomata of other *Rhizopogon* subgenera were collected. The EM host relationship of *Rhizopogon* species was confirmed by the proximity of available host trees or by sequencing the ITS

and matK genes of EM root tips found in proximity to basidiomata. Basidiomata of *R.* subgenus *Villosuli* were found in association with *P. wilsoniana* in Taiwan, *P. japonica* in Japan, *P. macrocarpa* in southern California, *P. menziesii* var. *menziesii* in Oregon and California, and *P. menziesii* var. *glauca* in the American Southwest (ASW) and Mexico. Three undescribed species of *Rhizopogon* were identified from these collections. One new species was recovered in association with Japanese *P. japonica* trees and has been described as *R. togasawariana* (Mujic et al. 2014). The second new species was found in association with *P. macrocarpa* in southern California and the third species was found in association with *P. wilsoniana* in Taiwan and *P. sinensis* var. *sinensis* in China (EM root tips only). Species descriptions for the two undescribed species are in preparation.

Of all basidiomata and EM root tips collected a total of 81 cultures were successfully isolated, 73 of them belonging to *R.* subgenus *Villosuli*. All cultures were verified as *Rhizopogon* species through macroscopic and microscopic morphological characters as well as BLAST comparison of ITS sequence data to the NCBI Genbank database. ITS data from all cultures, basidiomata, and EM root tips were aligned by hand into the ITS dataset of Mujic et al. (2014) to preserve previous alignment characteristics. Phylogenetic analysis was performed in RAxML. The ITS phylogeny inferred by RAxML (Figure 4.1) shows strong support for all *Pseudotsuga* associated *Rhizopogon* species as members of *R.* subgenus *Villosuli*. Within *R.* subgenus *Villosuli* some species show ranges restricted to particular species of *Pseudotsuga* or to only parts of a species range. The new species found in association with *P. macrocarpa* grouped as distinct member of *R.* subgenus *Villosuli* section *Villosuli* (sensu Grubisha et al. 2002) species complex in north America and appears to occur only in association with *P. macrocarpa*.

Members of the Asian clade show distinct separation from North American species as well as separation between Japanese and Chinese/Taiwanese clades. However, there is poor support for separation of Taiwanese species of *Rhizopogon* from those detected in China. In North America, *R. vesiculosus* was found growing only in populations of coastal *P. menziesii* var. *menziesii* in California and Oregon. In comparison its sister species *R. vinicolor* was found in association with both coastal *P. menziesii* var. *menziesii* and interior *P. menziesii* var. *glauca* as far south as northern Arizona. The other major North American clade containing individuals of *R. rudus* was discovered only from the ASW and Mexico in association with *P. menziesii* var. *glauca*. The results of the ITS phylogenetic analysis were used to guide Illumina sampling of 25 cultures of *R.* subgenus *Villosuli* as exemplars of both species and geographic clades. A single culture from each of the remaining four subgenera of *Rhizopogon* was obtained and these samples were included in Illumina sequencing. A list of all *Rhizopogon* cultures included in Illumina sequencing along with collection localities, culture information, and VELVET genome assembly statistics can be found in Table 4.1.

Identification of conserved protein coding regions and phylogenetic dataset construction:

FastOrtho analysis identified 5138 gene families of single copy protein coding genes shared amongst the genomes of *R. vinicolor, R. vesiculosus, S. luteus,* and *S. brevipes*. Of these, 989 genes passed our phylogenetic utility filter and were carried into further analysis. In addition to these 989 genes we added a representative LSU sequence from *R. villosulus* (Genbank accession #: AF071464) as well as the TEF sequence from

R. vinicolor (JGI protein # 766829) to our probe set for BLAST query of VELVET assemblies. TEF was not a protein included in the set of 5138 single copy protein coding genes and to identify this protein in *R. vinicolor* we used a TEF sequence from Japanese *R. togasawariana* as a BLAST query against the *R. vinicolor* genome. The LSU sequence for *R. vinicolor* was not included in the probe set because BLAST queries of ribosomal DNA (ITS, LSU, and the ribosomal small subunit (SSU)) against the *R. vinicolor* genome did not yield convincing sequence hits at any e-value threshold. The LSU sequence of *R. villosulus* (Genbank # AF071464) was removed from the analysis after it was used as a BLAST query to identify the LSU region in low coverage *Rhizopogon* assemblies.

A total of 991 loci were used as BLAST queries against each of the 29 draft genome assemblies. Of these 991 loci 24 were not found in one or more draft assemblies and only 30 total protein sequences were missing across all 29 assemblies. We designed primer pairs for 10 of the 989 protein coding genes identified for phylogenomic analysis and of these pairs 6 were successful at amplifying the correct region from Japanese *R. togasawariana* basidiomata. We were also successful in amplifying several of these proteins from Taiwanese and Chinese EM root tips. The sequence of novel primer pairs used in PCR reactions and sequence data derived from all Asian material using PCR is listed in Table 4.2. Sequence data from two *R. togasawariana* basidiomata as well as one Chinese and one Taiwanese EM root tip were manually inserted into the sequence datasets before alignment and trimming. The ITS dataset used to produce the ITS phylogram in Figure 4.1 was culled to just those sequences which were included in the Illumina sequencing (Table 4.1) and to the 4 east Asian samples amplified with PCR

(Table 4.2). This culled ITS alignment was included with the 991 BLAST identified loci and concatenated to a total alignment length of 1,969,478 bp.

Phylogenomic and Phylogeographic analyses:

Maximum likelihood phylogenetic analysis of the 1.97 mbp concatenated dataset achieved 100% bootstrap support for all clades of *Rhizopogon* except for the Taiwanese and Chinese clade of *R.* subgenus *Villosuli* (Figure 4.2). The species tree topology inferred from individual gene trees by MP-EST was in close agreement with the topology inferred from the concatenated dataset inferred by RAxML and small modifications to the structure of terminal clades in the MP-EST topology are shown in Figure 4.2. This analysis strongly supports *Rhizopogon* associated with all species of *Pseudotsuga* as members of *R.* subgenus *Villosuli* and there are several systematic relationships that are evident here for the first time. East Asian members *R.* subgenus *Villosuli* share a most recent common ancestor with *R.* subgenus *Villosuli* section *Villosuli* and Japanese *R. togasawariana* is supported as a distinct species from a second species with shared distribution in Taiwan and China. The amount of data collected from Chinese EM root tips of *P. sinensis* was limited and the placement of the sole Chinese specimen within the Taiwanese/Chinese clade is uncertain. *R.* subgenus *Villosuli* section *Villosuli* and *R.* subgenus *Villosuli* section *Vinicolores* are restricted to a North American distribution. The new species found in association with *P. macrocarpa* in southern California is conclusively supported as sister taxon to *R. hawkerae,* whereas this relationship was ambiguous in the ITS phylogram of Figure 4.1. *R. rudus* is supported as sister taxon to *R.*

villosulus and the previously recognized sister relationship between *R. vinicolor* and *R. vesiculosus* (Kretzer et al. 2003a) is upheld. Relationships between *Rhizopogon* subgenera are also resolved with confidence for the first time. *R.* subgenus *Roseoli* is seen here to be sister to *R.* subgenus *Villosuli* and these two subgenera share a sister relationship with the remaining three *Rhizopogon* subgenera. Our findings reinforce the conclusions of previous phylogenetic analysis (Molina et al. 1999, Grubisha et al. 2002) by placing *R.* subgenus *Versicolores* and *R.* subgenus *Rhizopogon* in a close phylogenetic relationship to one another and we observe here, for the first time, the sister relationship of *R.* subgenus *Amylopogon* to these subgenera.

There are strong phylogeographic trends within the species level clades of the concatenated dataset phylogeny, which indicate geographic disjuncts between and within taxa (Figure 4.3). As was previously observed in the ITS phylogeny of Figure 4.1, *R. vesiculosus* was only found from collections along the coastal distribution of *P. menziesii* var. *menziesii* and *R. vinicolor* was found throughout the distribution of *P. menziesii* in the United States. *R. vinicolor* shows a strong geographic clustering of interior ASW populations growing in association with *P. menziesii* var. *glauca* and coastal populations growing in association with *P. menziesii* var. *menziesii* (Figure 4.3). *R.* subgenus *Villosuli* section *Villosuli* shows similar interior and coastal disjuncts in species distribution. *R. hawkerae* and *R. parksii* were sampled only in coastal populations of *P. menziesii* var. *menziesii* in Oregon and the sister taxon to *R. hawkerae* was found only in southern California in association with *P. macrocarpa*. *R. rudus* was found only in interior populations of *P. menziesii* var. *glauca* and the members of this clade show a distinct north to south phylogenetic grade from northern Arizona into Central Mexico (Figure

4.3). ASR analysis in Mesquite reveal that the geographic disjunction observed here indicates a probable geographic origin of *R.* subgenus *Villosuli* in coastal western North America (Figure 4.2) with at least two migrations by *R. vinicolor* and *R. rudus* into the intermountain region of western North America.

DISCUSSION:

Phylogeography of Rhizopogon subgenus Villosuli and genus Pseudotsuga:

We have performed a genome-scale phylogeographic analysis of *Rhizopogon* specimens recovered from across the range of all *Pseudotsuga* species. The results of this analysis reveal comigrational geographic patterns of *Rhizopogon* species in EM association with *Pseudotsuga* and support a single evolutionary origin of this relationship. Both the ITS phylogeny (Figure 4.1) and concatenated genomic phylogeny (Figure 4.2) support Asian species as a distinct clade of *R.* subgenus *Villosuli* that occur with the Asian *Pseudotsuga* species *P. japonica, P. sinensis* and *P. wilsoniana*. The concatenated phylogeny indicates that Asian and North American members of *R.* subgenus *Villosuli* section *Villosuli* shared a common ancestor that diverged from *R.* subgenus *Villosuli* section *Vinicolores* prior to divergence of *R.* subgenus *Villosuli* and Asia species. ASR of *Rhizopogon* geography (Figure 4.2) indicates a high probability that this divergence event took place in coastal western North America and that the common ancestor of all *R.* subgenus *Villosuli* most likely evolved its EM host relationship with *Pseudotsuga* in coastal western North America. The oldest *Pseudotsuga* fossils known from Asia are from Japanese deposits dating to the early Miocene (~23 mya, Yabe 2011

and references therein) and this date is corroborated by molecular clock estimates, which predict an age of 20.26 ± 5.84 Mya for the divergence of Asian and North American *Pseudotsuga* species (Wei et al. 2010). The fossil record of *Pseudotsuga* in North America is far more extensive than those found in Asian deposits and the oldest *Pseudotsuga* fossils known are from deposits in Idaho, Nevada, and British Columbia dating to the early Eocene (~50-58 mya, Hermann 1985). The greater age of the *Rhizopogon-Pseudotsuga* EM relationship in North America inferred by ASR implies that migration of *Pseudotsuga* species into Asia is likely to have occurred after development of EM symbiosis with *Rhizopogon*. It is conceivable that *Rhizopogon* EM symbiosis was acquired by *Pseudotsuga* species upon arrival in North America after migration from Asia and that *Rhizopogon* species then migrated along host corridors into Asia. However, given the combination of fossil evidence and the phylogeographic analysis presented here the most probable migration scenario is the step-wise radiation model of *Pseudotsuga* species from North America into Asia as proposed by Strauss et al. (1990) and comigration of *R.* subgenus *Villosuli* along the same course.

Phylogeography of Rhizopogon subgenus Villosuli and Pseudotsuga in North America:

The geographic structure of *R.* subgenus *Villosuli* clades (Figure 4.3) is strong evidence for comigration and codivergence of *P. menziesii* varieties and *R.* subgenus *Villosuli* species in North America. *P. menziesii* is divided into two varieties, *P. menziesii* var. *menziesii* along the coast and *P. menziesii* var. *glauca* in the interior range (Faron 2010). The population structure of *P. menziesii* varieties is well characterized and

morphological, molecular, and chemical analyses divide *P. menziesii* varieties into coastal (BC, WA, OR, CA, P. *menziesii* var. *menziesii*), northern interior (BC, MT, WY, ID *P. menziesii* var. *glauca*), southern interior (CO, UT, AZ, NM *P. menziesii* var. *glauca*), and Mexican populations (*P. menziesii* var. *glauca*) (Strauss et al. 1990, Aagaard et al. 1995, Gugger et al. 2010, Gugger et al. 2011, Adams et al. 2012, Adams et al. 2013). Divergence of coastal and interior *P. menziesii* varieties is the result of vicariance associated with the orogeny of the Cascade and Sierra Nevada mountain ranges and the xerification of the Great Basin Desert approximately 3 mya (Gugger et al. 2010). The phylogeographic structure between interior (UT, AZ) and coastal (OR, CA) *R. vinicolor* (Figure 4.3) is concordant with the geographic division of coastal and interior *P. menziesii* varieties and it is likely that *R. vinicolor* has undergone the same vicariance event experienced by its host. Though we sampled more than 65 total *R. vinicolor* basidiomata across multiple sites in Oregon, California, Utah, Arizona and Mexico we failed to sample its sister species *R. vesiculosus* outside of Oregon and California. These two species share a sympatric range in the PNW and often occur at similar frequencies upon the same site (Kretzer et al. 2003b, Kretzer et al. 2005, Beiler et al. 2010, Beiler et al. 2012, Dunham et al. 2013). The lack of *R. vesiculosus* in association with interior *P. menziesii* var. *glauca* may indicate that *R. vesiculosus* diverged from *R. vesiculosus* independent of the vicariance event, which separated varieties of its host.

In *R.* subgenus *Villosuli* section *Villosuli* a geographic disjunct is seen between species rather than within species. *R. hawkerae, R. parksii* and a new species found in EM association with *P. macrocarpa* are restricted to the coast while *R. villosulus* and *R. rudus* are restricted to the ASW. This biogeographic pattern is due in part to insufficient

geographic sampling of section *Villosuli* as *R. villosulus* and *R. hawkerae* most certainly occur in northern interior stands of *P. menziesii* var. *glauca* and northern coastal stands of *P. menziesii* var. *menziesii* because sequenced type and non-type specimens of these species were collected in Idaho, Washington, and Oregon (Smith and Zeller 1966, Bubriski and Kennedy 2014, this study). However, all historical records of *R. parksii* are known from Washington, Oregon, and California (Smith and Zeller 1966, Bubriski and Kennedy 2014, this study) and it is likely that this species occurs only in association with coastal *P. menziesii* var. *menziesii.* Similarly, we only recovered *R. rudus* from from the ASW and Mexico. This species was described from type specimens occuring in Idaho (Smith and Zeller 1966) and the sequence data used to identify this species in the ITS phylogeny was derived from the *R. rudus* holotype. Our sampling and all confirmed type material indicate that *R. rudus* is restricted to a host association with interior *P. menziesii* var. *glauca*. Most interestingly, *R. rudus* collections made in the ASW and Mexico display a north to south phylogeographic grade (Figure 4.3). Following the divergence of interior and coastal *P. menziesii* varieties further population division of each variety occurred during Pleistocene glacial cycles and the modern population structure of both varieties is the result of recolonization from several glacial refugia (Gugger et al. 2010). *P. menziesii* var. *glauca* expanded into Mexico from such glacial refugia during a southward Pleistocene migration (Gugger et al. 2011). The topological structure of *R. rudus* in the concatenated phylogeny (Figure 4.3) shows strong concordance with the southward migration of its host, providing evidence for comigration of *R. rudus* and *P. menziesii* var. *glauca* into Mexico.

Evolution and ecology of host specific EM associations in Rhizopogon:

EM host association with *Pinus* hosts is proposed as a plesiomorphic character of the genera *Rhizopogon* and *Suillus* (Grubisha et al. 2002). Our results indicate that deviations from this plesiomorphic state have occurred at least three times in genus *Rhizopogon*. These transitions are observed in *R. subsalmonius* (*R.* subgenus *Versicolores*) which occurs with *Abies* (fir) EM hosts, *R*. subgenus *Amylopogon,* which occurs with various Pinaceae and Ericaceae hosts, and *R*. subgenus *Villosuli* which occurs with *Pseudotsuga spp.* (Molina et al. 1994, Massicotte et al. 1994, present study). It should be noted, however, that host association is considered here as corresponding to observed records of basidiome production in nature. Many species of *Suillus* and *Rhizopogon* are observed in nature fruiting only with their preferred EM host tree but pure culture synthesis EM trials (Molina and Trappe 1982, as reviewed in Molina and Trappe 1994), as well as spore (Massicotte et al. 1994) inoculation trials, demonstrate that many species are capable of forming EM root tips with various Pinaceae hosts. It may be that host association with the preferred host is a requirement for the production of basidiomata or simply that laboratory conditions allow for EM formation with hosts that could not occur in nature. Regardless, asexual reproduction has never been observed in genus *Rhizopogon* and dispersal is dependent upon production of sexual basidiomata and the presence of appropriate EM host trees. The spatial model of mutualism evolution proposed by Yamamura et al. (2004) predicts that mutualisms are most likely to evolve in organisms with a limited dispersal distance. Lack of spore production in the absence of an EM host and short range dispersal through mammalian mycophagy are consistent with

this model and are likely factors in the maintenance of host specific EM mutualism in *Rhizopogon*.

There are approximately 2500 species of EM fungi that associate with *P. menziesii* throughout its range (Molina et al. 1994) and *R.* subgenus *Villosuli* is only a part of this large fungal community. Given the large community of EM fungi, which are compatible with *P. menziesii,* it follows that *R.* subgenus *Villosuli* species should be the more dependent partner in EM symbiosis and that *Pseudotsuga* is likely to be the lead partner in comigration and cospeciation. However, these are several lines of evidence that are not consistent with this expectation. Species of *R.* subgenus *Villosuli* are a major component of the EM community colonizing *P. menziesii* at all successional stages (Molina et al. 1999, Smith et al. 2002, Twieg et al. 2007) and they are especially abundant upon the roots of young trees following disturbance (Luoma et al. 2006, Twieg et al. 2007). Spores of *R.* subgenus *Villosuli* are dominant members of the resistant spore bank community which colonize recruiting seedlings of *P. menziesii* (Luoma et al. 2006) and *P. japonica* (Murata et al. 2013) even when EM associations of *Rhizopogon* species are infrequent or absent upon the roots of mature trees in the same stand. Out-planted seedlings of *P. menziesii* inoculated with *R. vinicolor* and *R. hawkerae* (=*R. colossus*) have significantly greater rates of establishment and survival than uninoculated controls even though all seedlings are colonized after transplantation by indigenous EM fungi (Castellano and Trappe 1985). These findings all suggest that members of *R.* subgenus *Villosuli* are important for the early stages of forest establishment and may provide disproportionate benefits to their *Pseudotsuga* hosts compared with other EM fungi. Thus, a certain degree of mutual codependence may have driven the comigrational

pattern observed between *Pseudotsuga* species and *R.* subgenus *Villosuli* rather than fungal dependence upon the host alone.

Conclusions:

We have conducted a genome scale phylogeographic analysis of *R.* subgenus *Villosuli* and demonstrate a single evolutionary origin of the *Rhizopogon-Pseudotsuga* EM symbiosis involving the North American ancestors of these genera. Systematic relationships within *R.* subgenus *Villosuli* show strong phylogeographic concordance with the geographic population structure of *P. menziesii* in North America, which suggests close EM association and comigration of *Pseudotsuga* and *Rhizopogon* species. The common ancestors of North American *R.* subgenus *Villosuli* section *Vinicolores* and section *Villosuli* lineages diverged prior to the divergence of Asian species from section *Villosuli* and ASR of geographic distribution across *R.* subgenus *Villosuli* places the common ancestor of the subgenus in the coastal region of Western North America. These findings together support a geographic origin of genus *Pseudotsuga* in western North America and subsequent comigration and species radiation of *Pseudotsuga* and associated *Rhizopogon* symbionts into Asia. Molecular markers developed in this study have been successfully applied in phylogenetic analysis to determine intrageneric relationships of genus *Rhizopogon*. Deeper sampling and analysis of these markers from other *Rhizopogon* subgenera will greatly expand evolutionary knowledge of this ecologically significant EM genus.

In this study we demonstrate a novel technique, genome-scale sequence typing, which utilizes freely available software packages and custom BioPerl script for the

extraction of homologous protein coding DNA sequence data from low coverage de novo genome assemblies. Sequence recovery from de novo assemblies allows for generation of larger datasets than in-vitro sequence probe methods and avoids the potential sequence bias introduced by reference guided assembly methods. This method has proven effective for inference of phylogenetic relationships at both intrageneric and intraspecific scales between recently diverged non-model fungal taxa. This method can be used to develop hundreds to thousands of informative molecular markers given only a limited set of high coverage annotated reference genomes. Through the efforts of the 1000 fungal genomes project of the DOE-JGI we have seen the rapid expansion and availability of fungal genomic sequence data (Grigoriev et al. 2014). The application of our technique to the fungal genomes available as reference through this resource has great potential to expand evolutionary knowledge across Kingdom Fungi.

REFERENCES:

Aagaard, JE, Vollmer SS, Sorensen FC, and Strauss SH. 1995. Mitochondrial DNA Products among RAPD Profiles Are Frequent and Strongly Differentiated between Races of Douglas-Fir. *Molecular Ecology* 4 (4): 441–46. doi:10.1111/j.1365-294X.1995.tb00237.x.

Adams RP, Vargas-Hernández JJ, Elizondo MSG, Hunter G. Fairhall TA, Thornburg D, Callahan F. 2013. Taxonomy of Douglas fir (*Pseudotsuga menziesii*) infraspecific taxa: vars. *menziesii, glauca* and *oaxacana*: nrDNA, cpDNA sequences and leaf essential oils. *Phytologia* 95(1):94-102

Adams RP, Vargas-Hernández JJ, Elizondo MSG, Hunter G. Fairhall TA, Thornburg D, Callahan F. 2012. Geographic variation in leaf essential oils of Douglas fir (*Pseudotsuga menziesii*). *Phytologia* 94(2):199-218.

Bahram, Mohammad, Sergei Põlme, Urmas Kõljalg, Shahin Zarre, and Leho Tedersoo. 2012. Regional and Local Patterns of Ectomycorrhizal Fungal Diversity and Community Structure along an Altitudinal Gradient in the Hyrcanian Forests of Northern Iran. *New Phytologist* 193 (2): 465–73. doi:10.1111/j.1469-8137.2011.03927.x.

Beiler KJ, Durall DM, Simard SW, Maxwell SA and Kretzer AM. 2010. Architecture of the wood-wide web: *Rhizopogon spp*. genets link multiple Douglas-fir cohorts *New Phytologist* 185(2) 543-553.

Beiler, Kevin J., Suzanne W. Simard, Valerie LeMay, and Daniel M. Durall. 2012. Vertical Partitioning between Sister Species of *Rhizopogon* Fungi on Mesic and Xeric Sites in an Interior Douglas-Fir Forest. *Molecular Ecology* 21 (24): 6163–74. doi:10.1111/mec.12076.

Bidartondo, M. I., and T. D. Bruns. 2002. Fine-Level Mycorrhizal Specificity in the Monotropoideae (Ericaceae): Specificity for Fungal Species Groups. *Molecular Ecology* 11 (3): 557–69. doi:10.1046/j.0962-1083.2001.01443.x.

Bruns, Thomas D., Robert Fogel, Thomas J. White, and Jeffrey D. Palmer. 1989. Accelerated Evolution of a False-Truffle from a Mushroom Ancestor. *Nature* 339 (6220): 140–42. doi:10.1038/339140a0.

Bruns, Thomas D., Martin I. Bidartondo, and D. Lee Taylor. 2002. Host Specificity in Ectomycorrhizal Communities: What Do the Exceptions Tell Us? *Integrative and Comparative Biology* 42 (2): 352–59. doi:10.1093/icb/42.2.352.

Bubriski, Ryan, and Peter Kennedy. 2014. A Molecular and Morphological Analysis of the Genus *Rhizopogon* Subgenus *Villosuli* Section *Villosuli* as a Preface to Ecological Monitoring. *Mycologia* 106 (2): 353–61. doi:10.3852/106.2.353.

Cantarel, Brandi L., Ian Korf, Sofia M. C. Robb, Genis Parra, Eric Ross, Barry Moore, Carson Holt, Alejandro Sánchez Alvarado, and Mark Yandell. 2008. MAKER: An Easy-to-Use Annotation Pipeline Designed for Emerging Model Organism Genomes. *Genome Research* 18 (1): 188–96. doi:10.1101/gr.6743907.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses. *Bioinformatics* 25 (15): 1972–73. doi:10.1093/bioinformatics/btp348.

Castellano, Michael A., and James M. Trappe. 1985. Ectomycorrhizal Formation and

Plantation Performance of Douglas-Fir Nursery Stock Inoculated with *Rhizopogon* Spores. *Canadian Journal of Forest Research* 15 (4): 613–17. doi:10.1139/x85-100.

Cullings KW, Szaro TM , Bruns TD. 1996. Evolution of extreme specialization within a lineage of ectomycorrhizal epiparasites. *Nature* 379:63-66

Debreczy Z. and Racz I. 1995. New species and varieties of conifers from Mexico. *Phytologia* 78:217–243.

Dunham, Susie M., Alija Bajro Mujic, Joseph W. Spatafora, and Annette M. Kretzer. 2013. Within-Population Genetic Structure Differs between Two Sympatric Sister-Species of Ectomycorrhizal Fungi, *Rhizopogon vinicolor* and *R. vesiculosus*. *Mycologia* 105 (4): 814–26. doi:10.3852/12-265.

Edgar, Robert C. 2004. MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput. *Nucleic Acids Research* 32 (5): 1792–97. doi:10.1093/nar/gkh340.

Falcão, Vanessa Dos Reis, Angela Pedroso Tonon, Mariana Cabral Oliveira, and Pio Colepicolo. 2007. RNA Isolation Method for Polysaccharide Rich Algae: Agar Producing *Gracilaria tenuistipitata* (Rhodophyta). *Journal of Applied Phycology* 20 (1): 9–12. doi:10.1007/s10811-007-9174-7.

Farjon A. 2010. *A handbook of the world's conifers*. Brill Leiden-Boston.

Gardes M, Bruns TD. 1993. ITS primers with enhanced specificity for basidiomycetes — application to the identification of mycorrhizae and rusts. *Molecular Ecology* 2: 113–118.

Geml, József, Rodham E. Tulloss, Gary A. Laursen, Nina A. Sazanova, and D. L. Taylor. 2008. Evidence for Strong Inter- and Intracontinental Phylogeographic Structure in *Amanita muscaria*, a Wind-Dispersed Ectomycorrhizal Basidiomycete. *Molecular Phylogenetics and Evolution* 48 (2): 694–701. doi:10.1016/j.ympev.2008.04.029.

Gernandt, David S., and Aaron Liston. 1999. Internal Transcribed Spacer Region Evolution in *Larix* and *Pseudotsuga* (Pinaceae). *American Journal of Botany* 86 (5): 711–23.

Gillespie, Joseph J., Alice R. Wattam, Stephen A. Cammer, Joseph L. Gabbard, Maulik P. Shukla, Oral Dalay, Timothy Driscoll, et al. 2011. PATRIC: The Comprehensive Bacterial Bioinformatics Resource with a Focus on Human Pathogenic Species. *Infection and Immunity* 79 (11): 4286–98. doi:10.1128/IAI.00207-11.

Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F, Smirnova T, Nordberg H, Dubchak I, Shabalov I. 2014. MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Research* 42(1):D699-704.

Grubisha, Lisa C., James M. Trappe, Randy Molina, and Joseph W. Spatafora. 2002. Biology of the Ectomycorrhizal Genus *Rhizopogon*. VI. Re-Examination of Infrageneric Relationships Inferred from Phylogenetic Analyses of ITS Sequences. *Mycologia* 94 (4): 607–19.

Gugger, Paul F., Shinya Sugita, and Jeannine Cavender-Bares. 2010. Phylogeography of

Douglas-Fir Based on Mitochondrial and Chloroplast DNA Sequences: Testing Hypotheses from the Fossil Record. *Molecular Ecology* 19 (9): 1877–97. doi:10.1111/j.1365-294X.2010.04622.x.

Gugger, Paul F., Antonio González-Rodríguez, Hernando Rodríguez-Correa, Shinya Sugita, and Jeannine Cavender-Bares. 2011. Southward Pleistocene Migration of Douglas-Fir into Mexico: Phylogeography, Ecological Niche Modeling, and Conservation of 'rear Edge' Populations. *New Phytologist* 189 (4): 1185–99. doi:10.1111/j.1469-8137.2010.03559.x.

Hermann RK 1982. The genus *Pseudotsuga*: historical records and nomenclature. Forest Research Laboratory, Special Publication 2a. Oregon State University, Corvallis, OR.

Hermann RK. 1985. The Genus *Pseudotsuga*: Ancestral History and Past Distribution, Forest Research Laboratory, Special Publication 2b. Oregon State University, Corvallis, OR.

Hillebrand, Helmut. 2004. On the Generality of the Latitudinal Diversity Gradient. *The American Naturalist* 163 (2): 192–211. doi:10.1086/381004.

Horton, Thomas R., and Thomas D. Bruns. 1998. Multiple-Host Fungi Are the Most Frequent and Abundant Ectomycorrhizal Types in a Mixed Stand of Douglas Fir (*Pseudotsuga menziesii*) and Bishop Pine (*Pinus muricata*). *New Phytologist* 139 (2): 331–39. doi:10.1046/j.1469-8137.1998.00185.x.

Horton, Thomas R, Thomas D Bruns, and V Thomas Parker. 1999. Ectomycorrhizal Fungi Associated with *Arctostaphylos* Contribute to *Pseudotsuga menziesii* Establishment. *Canadian Journal of Botany* 77 (1): 93–102. doi:10.1139/b98-208.

Ishida, Takahide A., Kazuhide Nara, and Taizo Hogetsu. 2007. Host Effects on Ectomycorrhizal Fungal Communities: Insight from Eight Host Species in Mixed Conifer–broadleaf Forests. *New Phytologist* 174 (2): 430–40. doi:10.1111/j.1469-8137.2007.02016.x.

Kennedy, Peter G., Roberto Garibay-Orijel, Logan M. Higgins, and Rodolfo Angeles-Arguiz. 2011. Ectomycorrhizal Fungi in Mexican *Alnus* Forests Support the Host Co-Migration Hypothesis and Continental-Scale Patterns in Phylogeography. *Mycorrhiza* 21 (6): 559–68. doi:10.1007/s00572-011-0366-2.

Kretzer, Annette M., Daniel L. Luoma, Randy Molina, and Joseph W. Spatafora. 2003a. Taxonomy of the *Rhizopogon vinicolor* Species Complex Based on Analysis of ITS Sequences and Microsatellite Loci. *Mycologia* 95 (3): 480–87.

Kretzer AM, Dunham S, Molina R, Spatafora JW. 2003b. Microsatellite markers reveal the below ground distribution of genets in two species of *Rhizopogon* forming tuberculate ectomycorrhizas on Douglas fir. *New Phytologist* 161(1) 313-320

Liu, Liang, Lili Yu, and Scott V. Edwards. 2010. A Maximum Pseudo-Likelihood Approach for Estimating Species Trees under the Coalescent Model. *BMC Evolutionary Biology* 10 (1): 302. doi:10.1186/1471-2148-10-302.

Luoma, Daniel L, Christopher A Stockdale, Randy Molina, and Joyce L Eberhart. 2006. The Spatial Influence of *Pseudotsuga menziesii* Retention Trees on Ectomycorrhiza Diversity. *Canadian Journal of Forest Research* 36 (10): 2561–73. doi:10.1139/x06-143.

Maddison, W. P. and D.R. Maddison. 2015. Mesquite: a modular system for evolutionary

analysis. Version 3.02 http://mesquiteproject.org
Massicotte, Hugues B., Randy Molina, Daniel L. Luoma, and Jane E. Smith. 1994. Biology of the Ectomycorrhizal Genus, *Rhizopogon*. *New Phytologist* 126 (4): 677–90. doi:10.1111/j.1469-8137.1994.tb02962.x.
Merckx, Vincent, and Martin I. Bidartondo. 2008. Breakdown and Delayed Cospeciation in the Arbuscular Mycorrhizal Mutualism. *Proceedings of the Royal Society of London B: Biological Sciences* 275 (1638): 1029–35. doi:10.1098/rspb.2007.1622.
Molina, Randy. 1981. Ectomycorrhizal Specificity in the Genus *Alnus*. *Canadian Journal of Botany* 59 (3): 325–34. doi:10.1139/b81-045.
Molina, Randy, and James M. Trappe. 1982. Patterns of Ectomycorrhizal Host Specificity and Potential among Pacific Northwest Conifers and Fungi. *Forest Science* 28(3):423-458
Molina, Randy, and James M. Trappe. 1994. Biology of the Ectomycorrhizal Genus, *Rhizopogon*: I. Host associations, host-specificity and pure culture syntheses *New Phytologist* 126 (4): 653–75. doi:10.1111/j.1469-8137.1994.tb02961.x.
Molina, R., J. M. Trappe, L. C. Grubisha, and J. W. Spatafora. 1999. *Rhizopogon*. In *Ectomycorrhizal Fungi Key Genera in Profile*, edited by Dr John W. G. Cairney and Dr Susan M. Chambers (Eds.), 129–61. Springer Berlin Heidelberg.
Mujic, Alija B., Kentaro Hosaka, and Joseph W. Spatafora. 2014. *Rhizopogon togasawariana* Sp. Nov., the First Report of *Rhizopogon* Associated with an Asian Species of *Pseudotsuga*. *Mycologia* 106 (1): 105–12. doi:10.3852/13-055.
Murat, Claude, Jesús Díez, Patricia Luis, Christine Delaruelle, Chantal Dupré, Gérard Chevalier, Paola Bonfante, and Francis Martin. 2004. Polymorphism at the Ribosomal DNA ITS and Its Relation to Postglacial Re-Colonization Routes of the Perigord Truffle *Tuber melanosporum*. *New Phytologist* 164 (2): 401–11. doi:10.1111/j.1469-8137.2004.01189.x.
Murata, Masao, Akihiko Kinoshita, and Kazuhide Nara. 2013. Revisiting the Host Effect on Ectomycorrhizal Fungal Communities: Implications from Host–fungal Associations in Relict *Pseudotsuga japonica* Forests. *Mycorrhiza* 23 (8): 641–53. doi:10.1007/s00572-013-0504-0.
Põlme, Sergei, Mohammad Bahram, Takashi Yamanaka, Kazuhide Nara, Yu Cheng Dai, Tine Grebenc, Hojka Kraigher, et al. 2013. Biogeography of Ectomycorrhizal Fungi Associated with Alders (*Alnus* Spp.) in Relation to Biotic and Abiotic Variables at the Global Scale. *New Phytologist* 198 (4): 1239–49. doi:10.1111/nph.12170.
Rehner, Stephen A., and Ellen Buckley. 2005. A *Beauveria* Phylogeny Inferred from Nuclear ITS and EF1-A Sequences: Evidence for Cryptic Diversification and Links to *Cordyceps* Teleomorphs. *Mycologia* 97 (1): 84–98. doi:10.3852/mycologia.97.1.84.
Rochet, Juliette, Pierre-Arthur Moreau, Sophie Manzi, and Monique Gardes. 2011. Comparative Phylogenies and Host Specialization in the Alder Ectomycorrhizal Fungi *Alnicola*, *Alpova* and *Lactarius* (Basidiomycota) in Europe. *BMC Evolutionary Biology* 11 (1): 40. doi:10.1186/1471-2148-11-40.
Rohland, Nadin, and David Reich. 2012. Cost-Effective, High-Throughput DNA

Sequencing Libraries for Multiplexed Target Capture. *Genome Research* 22 (5): 939–46. doi:10.1101/gr.128124.111.

Rozen S, Skaletsky H. 1999. Primer3 on the WWW for General Users and for Biologist Programmers - Springer. In . *Bioinformatics Methods and Protocols: Methods in Molecular Biology* 132:365-386. Misener S, Krawetz, SA (eds). Humana Press.

Shaw, Timothy I., Zheng Ruan, Travis C. Glenn, and Liang Liu. 2013. STRAW: Species TRee Analysis Web Server. *Nucleic Acids Research* 41 (W1): W238–41. doi:10.1093/nar/gkt377.

Smith, AH., and Zeller SM. 1966. Preliminary Account of the North American Species of *Rhizopogon*. *Memoirs of the New York Botanical Garden* 14(2):177 pgs

Stajich, Jason E., David Block, Kris Boulez, Steven E. Brenner, Stephen A. Chervitz, Chris Dagdigian, Georg Fuellen, et al. 2002. The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Research* 12 (10): 1611–18. doi:10.1101/gr.361602.

Smith JE, Moina R, Huso MMP, Luoma DL, McKay D, Castellano MA, Lebel T, Valachuvic Y. 2002. Species richness, abundance, and composition of hypogeous and epigeous ectomycorrhizal fungal sporocarps in young, rotation-age, and old-growth stands of Douglas-fir (*Pseudotsuga menziesii*) in the Cascade range of Oregon, U.S.A. *Canadian Journal of Botany* 80:186-204

Stamatakis, A. 2006. RAxML-VI-HPC: Maximum Likelihood-Based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. *Bioinformatics* 22 (21): 2688–90. doi:10.1093/bioinformatics/btl446.

Strauss, Steven H., Allan H. Doerksen, and Joyce R. Byrne. 1990. Evolutionary Relationships of Douglas-Fir and Its Relatives (genus *Pseudotsuga*) from DNA Restriction Fragment Analysis. *Canadian Journal of Botany* 68 (7): 1502–10. doi:10.1139/b90-191.

Tedersoo, Leho, Teele Jairus, Bryony M. Horton, Kessy Abarenkov, Triin Suvi, Irja Saar, and Urmas Kõljalg. 2008. Strong Host Preference of Ectomycorrhizal Fungi in a Tasmanian Wet Sclerophyll Forest as Revealed by DNA Barcoding and Taxon-Specific Primers. *New Phytologist* 180 (2): 479–90. doi:10.1111/j.1469-8137.2008.02561.x.

Tedersoo, Leho, Triin Suvi, Teele Jairus, Ivika Ostonen, and Sergei Põlme. 2009. Revisiting Ectomycorrhizal Fungi of the Genus *Alnus*: Differential Host Specificity, Diversity and Determinants of the Fungal Community. *New Phytologist* 182 (3): 727–35. doi:10.1111/j.1469-8137.2009.02792.x.

Tedersoo, Leho, Mohammad Bahram, Märt Toots, Abdala G. Diédhiou, Terry W. Henkel, Rasmus Kjøller, Melissa H. Morris, et al. 2012. Towards Global Patterns in the Diversity and Community Structure of Ectomycorrhizal Fungi. *Molecular Ecology* 21 (17): 4160–70. doi:10.1111/j.1365-294X.2012.05602.x.

Tedersoo, Leho, Mohammad Bahram, Sergei Põlme, Urmas Kõljalg, Nourou S. Yorou, Ravi Wijesundera, Luis Villarreal Ruiz, et al. 2014. Global Diversity and Geography of Soil Fungi. *Science* 346 (6213): 1256688. doi:10.1126/science.1256688.

Twieg, Brendan D., Daniel M. Durall, and Suzanne W. Simard. 2007. Ectomycorrhizal Fungal Succession in Mixed Temperate Forests. *New Phytologist* 176 (2): 437–47. doi:10.1111/j.1469-8137.2007.02173.x.

Vilgalys, R., and M. Hester. 1990. Rapid Genetic Identification and Mapping of Enzymatically Amplified Ribosomal DNA from Several *Cryptococcus* Species. *Journal of Bacteriology* 172 (8): 4238–46.

Wei XX, Yang ZU, Li Y, Wang XQ. 2010. Molecular phylogeny and biogeography of *Pseudotsuga* (Pinaceae): Insights into the floristic relationship between Taiwan and its adjacent areas. *Molecular Phylogenetics and Evolution.* 55(3):776-785.

Weitemier, Kevin, Shannon C. K. Straub, Richard C. Cronn, Mark Fishbein, Roswitha Schmickl, Angela McDonnell, and Aaron Liston. 2014. Hyb-Seq: Combining Target Enrichment and Genome Skimming for Plant phylogenomics. *Applications in Plant Sciences* 2 (9). doi:10.3732/apps.1400042.

White TJ, Bruns TD, Lee SW, Taylor JW. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis MA, Gelfand DH, Sninsky JJ, White TJ, (Eds.), *PCR Protocols--a Guide to Methods and Applications*. Academic Press, San Diego, CA. Pp 315–322.

Yabe, Atsushi. 2011. *Pseudotsuga tanaii* Huzioka from the Earliest Miocene Shichiku Flora of Northeast Japan: Systematics and Ecological Conditions. *Paleontological Research* 15 (1): 1–11. doi:10.2517/1342-8144-15.1.001.

Yamamura, Norio, Masahiko Higashi, Narayan Behera, and Joe Yuichiro Wakano. 2004. Evolution of Mutualism through Spatial Effects. *Journal of Theoretical Biology* 226 (4): 421–28. doi:10.1016/j.jtbi.2003.09.016.

Yamanaka, Takashi, Ching-Y. Li, Bernard T. Bormann, and Hiroaki Okabe. 2003. Tripartite Associations in an Alder: Effects of *Frankia* and *Alpova diplophloeus* on the Growth, Nitrogen Fixation and Mineral Acquisition of *Alnus tenuifolia*. *Plant and Soil* 254 (1): 179–86. doi:10.1023/A:1024938712822.

Zerbino, Daniel R., and Ewan Birney. 2008. Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs. *Genome Research* 18 (5): 821–29. doi:10.1101/gr.074492.107.

Figure 4.1: ITS phylogram inferred in RAxML using the GTRGAMMA substitution model and 1000 bootstrap replicates. Bootstrap support values 70% or greater are shown for major clades. Color coding of *R.* subgenus *Villosuli* taxa sequenced in this study denotes host affiliation. Taxon names drawn from the dataset of Mujic et al. (2014) are unaltered to allow reference to the original manuscript. Species concepts of labeled clades in *R.* subgenus *Villosuli* section *Villosuli* are based upon Bubriski and Kennedy (2014). A) Genus *Rhizopogon* with major clades of *R.* subgenus *Villosuli* drawn as cartoons. B) *R.* subgenus *Villosuli* section *Villosuli* and Asian species. C) *R.* subgenus *Villosuli* section *Vinicolores* with *R. parksii* (section *Villosuli*).

Figure 4.1

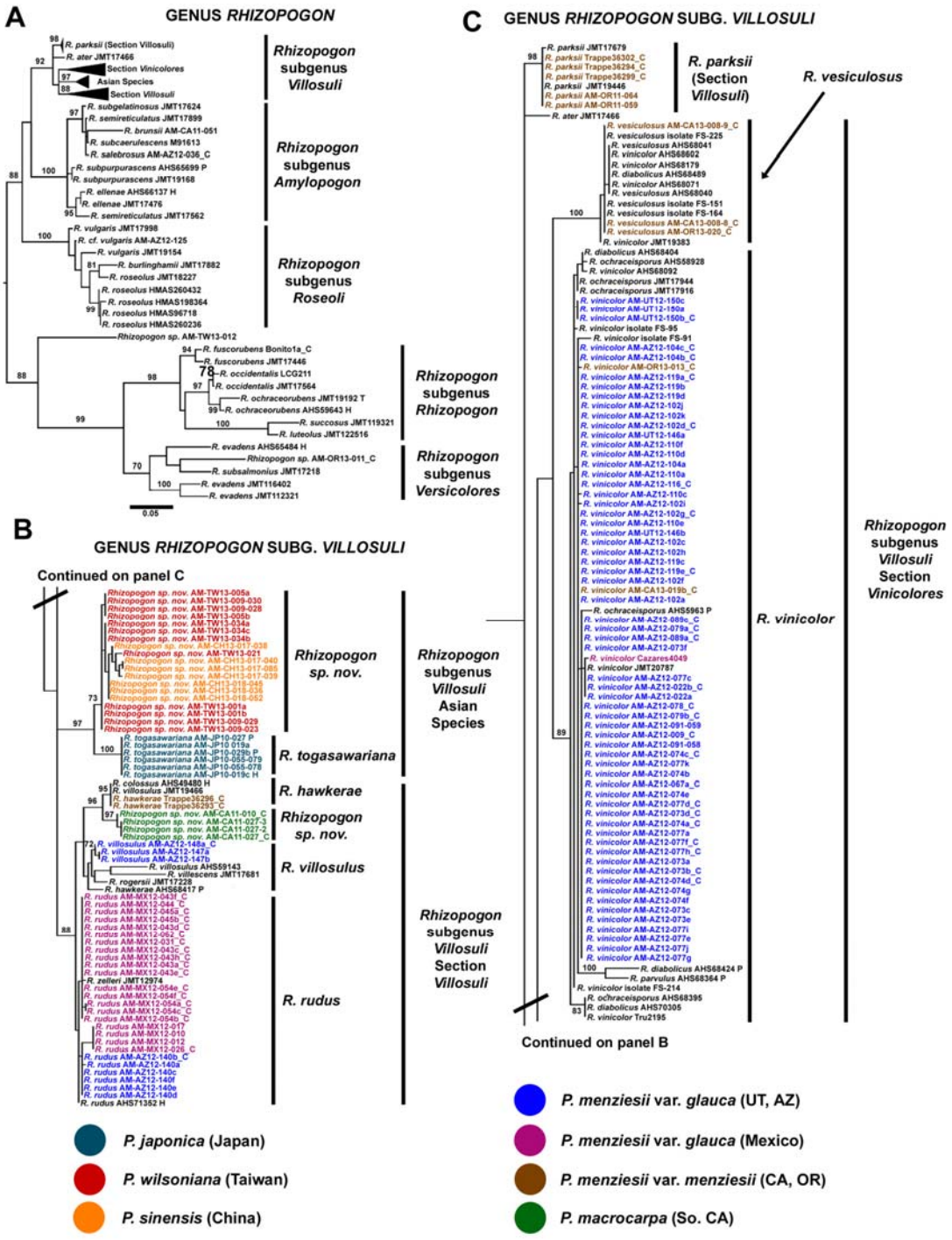

A
GENUS RHIZOPOGON
Rhizopogon subgenus Villosuli
Rhizopogon subgenus Amylopogon
Rhizopogon subgenus Roseoli
Rhizopogon subgenus Rhizopogon
Rhizopogon subgenus Versicolores
B
GENUS RHIZOPOGON SUBG. VILLOSULI
Continued on panel C
Rhizopogon sp. nov.
R. togasawariana
R. hawkerae
Rhizopogon sp. nov.
R. villosulus
R. rudus
Rhizopogon subgenus Villosuli Asian Species
Rhizopogon subgenus Villosuli Section Villosuli
C
GENUS RHIZOPOGON SUBG. VILLOSULI
R. parksii (Section Villosuli)
R. vesiculosus
R. vinicolor
Rhizopogon subgenus Villosuli Section Vinicolores
Continued on panel B
P. japonica (Japan)
P. wilsoniana (Taiwan)
P. sinensis (China)
P. menziesii var. glauca (UT, AZ)
P. menziesii var. glauca (Mexico)
P. menziesii var. menziesii (CA, OR)
P. macrocarpa (So. CA)

Figure 4.2: Phylogram inferred in RAxML using the concatenated 1.97 mbp concatenated dataset with the GTRGAMMA substitution model and 100 bootstrap replicates. Topological rearrangements of terminal taxa inferred by MP-EST are shown by light red lines. The percent likelihood for ancestral geographic range as predicted by ASR are noted for each major internode of *R.* subgenus *Villosuli.* Bootstrap support values of all clades are 100% except where noted in Asian species of *R.* subgenus *Villosuli.* Species concepts of *R.* subgenus *Villosuli* section *Villosuli* are based upon Bubriski and Kennedy (2014). Abbreviations: AS = Asia, ASW = American Southwest, CWNA =coastal western North America, MX = Mexico.

Figure 4.2

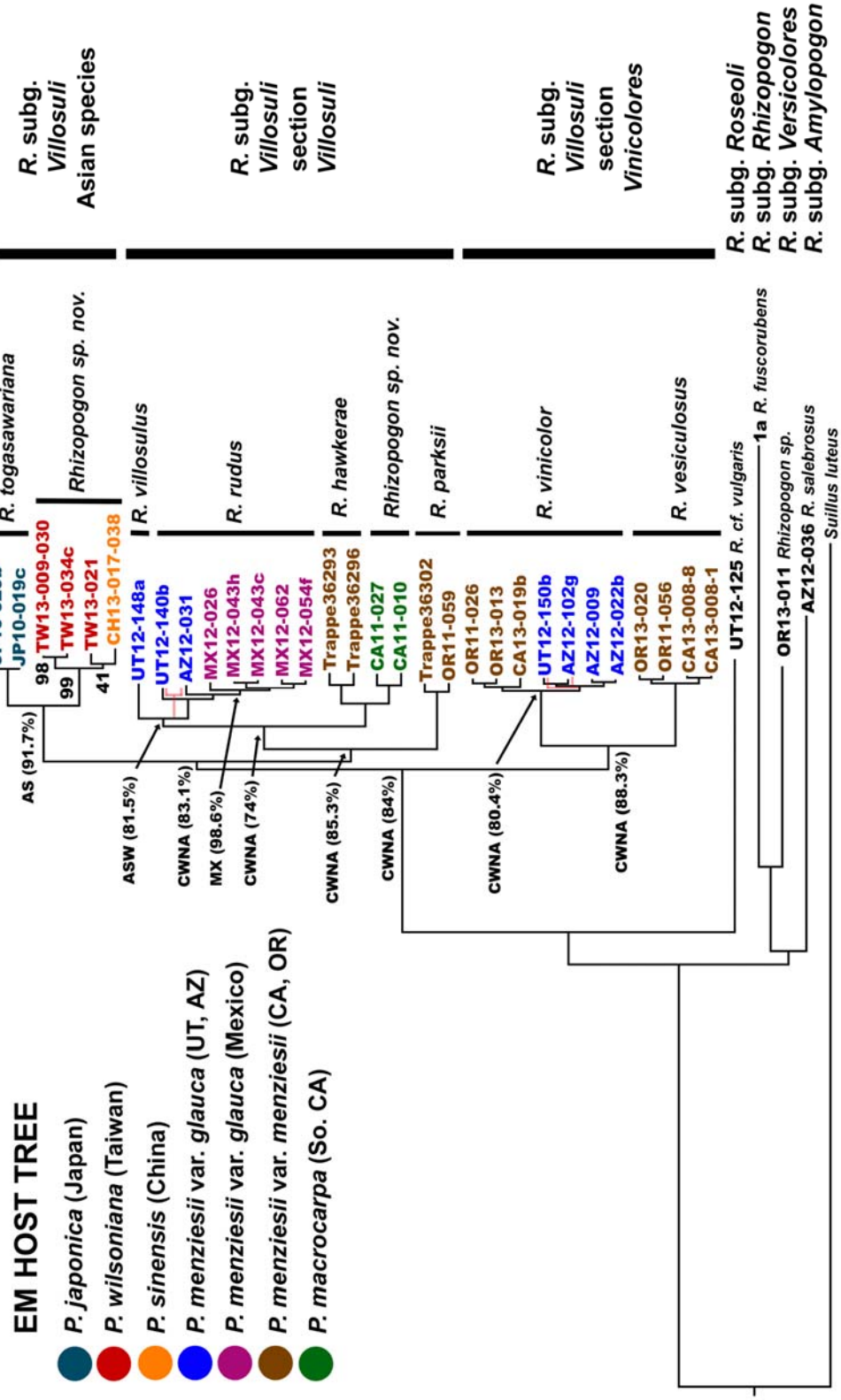

EM HOST TREE
P. japonica (Japan)
P. wilsoniana (Taiwan)
P. sinensis (China)
P. menziesii var. glauca (UT, AZ)
P. menziesii var. glauca (Mexico)
P. menziesii var. menziesii (CA, OR)
P. macrocarpa (So. CA)
R. subg. Villosuli Asian species
R. subg. Villosuli section Villosuli
R. subg. Villosuli section Vinicolores
R. subg. Roseoli
R. subg. Rhizopogon
R. subg. Versicolores
R. subg. Amylopogon
R. togasawariana
Rhizopogon sp. nov.
R. villosulus
R. rudus
R. hawkerae
Rhizopogon sp. nov.
R. parksii
R. vinicolor
R. vesiculosus
R. cf. vulgaris
1a R. fuscorubens
Rhizopogon sp.
R. salebrosus
Suillus luteus
JP10-029b
JP10-019c
TW13-009-030
TW13-034c
TW13-021
CH13-017-038
UT12-148a
UT12-140b
AZ12-031
MX12-026
MX12-043h
MX12-043c
MX12-062
MX12-054f
Trappe36293
Trappe36296
CA11-027
CA11-010
Trappe36302
OR11-059
OR11-026
OR13-013
CA13-019b
UT12-150b
AZ12-102g
AZ12-009
AZ12-022b
OR13-020
OR11-056
CA13-008-8
CA13-008-1
UT12-125
OR13-011
AZ12-036
98
99
41
AS (91.7%)
ASW (81.5%)
CWNA (83.1%)
MX (98.6%)
CWNA (74%)
CWNA (85.3%)
CWNA (84%)
CWNA (80.4%)
CWNA (88.3%)

Figure 4.3: Phylogeographic patterns observed for the species level clades of *R.* subgenus *Villosuli.* Red markers denote collection locality of each sample. Taxon names in each phylogram are colored to denote host association as in Figure 2. A) Phylogeography of *Rhizopogon vinicolor* in the western United States. B) Phylogeography of *Rhizopogon rudus* in the ASW and Mexico

Figure 4.3

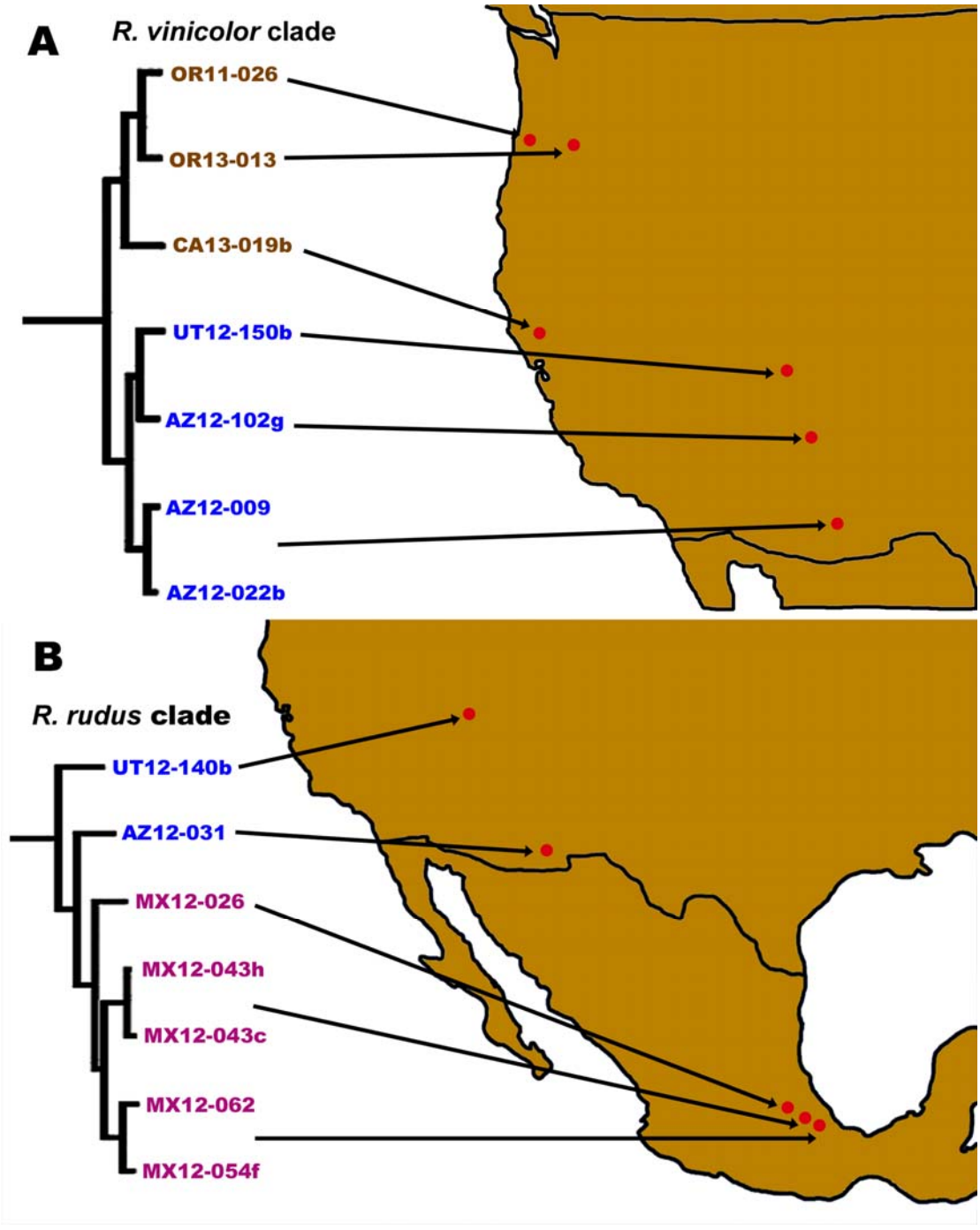

A
R. vinicolor clade
OR11-026
OR13-013
CA13-019b
UT12-150b
AZ12-102g
AZ12-009
AZ12-022b
B
R. rudus clade
UT12-140b
AZ12-031
MX12-026
MX12-043h
MX12-043c
MX12-062
MX12-054f

Table 4.1: A list of all *Rhizopogon* cultures selected for low coverage Illumina sequencing showing collection information and VELVET assembly statistics for all samples. Raw N50 = At least 50% of all assembled sequence is contained within a contig of this size or larger. Contig N50 = The number of contigs containing at least 50% of all assembled sequence.

Collector's Number	Species	*Rhizopogon* Subgenus	Country	State / Prefecture	Collection Locality	Latitude / Longitude	Altitude	Assembly Length	Assembly Contigs	Assembly Coverage	Raw N50	Contig N50	Largest Contig
Trappe36293	*R. hawkerae*	*Villosuli*	United States	Oregon	Willamette NF, off highway 46, Fox Creek CG	44 46 57.6 N, 122 2 59.76 W	597 meters	36933405	16566	21.1	5986	1411	118008
Trappe36296	*R. hawkerae*	*Villosuli*	United States	Oregon	Mt. Hood NF, along road 46.9 about 4 miles from highway 46	44 52 14.64 N, 121 49 26.4 W	1146 meters	36997820	16394	19.29	6426	1335	143158
AZ12-031	*R. rudus*	*Villosuli*	United States	Arizona	Chiricahua mountains, Coronado NF, Herb Martyr CG	31 52 20.9 N, 109 14 8.5 W	1790 meters	37041257	9599	28.04	17387	511	166113
UT12-140b	*R. rudus*	*Villosuli*	United States	Utah	Dixie NF, Site # 4	37 31 46.6 W, 112 45 15.1 W	2847 meters	36401066	16337	21.47	6298	1371	109564
MX12-026	*R. rudus*	*Villosuli*	Mexico	Hidalgo	El Chico NP, oak conifer forest	20 10 48.1 N, 98 43 58.2 W	2937 meters	37111067	8466	26.34	19709	487	212162
MX12-043c	*R. rudus*	*Villosuli*	Mexico	Tlaxcala	Near Villa Real, above highway.	19 31 7.8 N, 97 54 5.1 W	2987 meters	37108367	8026	42.74	21432	426	228706
MX12-043h	*R. rudus*	*Villosuli*	Mexico	Tlaxcala	Near Villa Real, above highway.	19 31 7.8 N, 97 54 5.1 W	2987 meters	35493029	12962	28.95	9179	923	106221
MX12-054f	*R. rudus*	*Villosuli*	Mexico	Puebla	Above village of Cuauhtemoc, barbeque site	19 12 13.4 N, 97 18 46 W	3009 meters	34904978	12830	22.64	8116	1027	84302
MX12-062	*R. rudus*	*Villosuli*	Mexico	Puebla	SE of Guadalupe Victoria, Near forestry hut, on forest road	19 8 4.3 N, 97 18 34.7 W	3187 meters	33199645	21719	19.46	2754	3240	25953
OR11-059	*R. parksii*	*Villosuli*	United States	Oregon	Peavy Arboretum	44 40 6.05 N, 123 14 23.05 W	188 meters	32348656	20075	18.95	2886	2837	46208
Trappe36302	*R. parksii*	*Villosuli*	United States	Oregon	Willamette NF, roadside on highway 46	44 46 59.22 N, 122 1 37.74 W	648 meters	35807735	15420	19.41	6692	1240	93538
CA11-010	*R. sp. nov*	*Villosuli*	United States	California	Mt. Palomar, South Grade Road	33 18 17.4 N, 116 52 30.4 W	1323 meters	34501171	16972	17.29	4775	1751	56562
CA11-027	*R. sp. nov*	*Villosuli*	United States	California	San Jacinto Mts., San Bernardino NF, HW 74	33 41 0.6 N, 116 41 42 W	1444 meters	36456402	7647	26.19	22889	385	312656
TW13-021	*R. sp. nov*	*Villosuli*	Taiwan	Nantou	Mt Bilu Trail, Taroko NP	24 11 16.9 N, 121 18 42.5 E	2550 meters	39405955	18001	21.41	5223	1750	58049
TW13-034c	*R. sp. nov*	*Villosuli*	Taiwan	Nantou	Mt Bilu Trail, Taroko NP	24 11 15.3 N, 121 18 46.9 E	2588 meters	40165079	22913	21.57	3626	2493	62832
UT12-148a	*R. villosulus*	*Villosuli*	United States	Utah	Dixie NF, Site # 9	37 31 50.9 N, 112 45 22.4 W	2842 meters	33280808	17976	18.53	3623	2336	39969
CA13-008-1	*R. vesiculosus*	*Villosuli*	United States	California	UCSC, Marshall field	37 1 24.9 N, 122 5 2.6 W	375 meters	38562603	12895	20.23	9244	933	157051
CA13-008-8	*R. vesiculosus*	*Villosuli*	United States	California	UCSC, Marshall field	37 1 24.9 N, 122 5 2.6 W	375 meters	39635651	12775	22.58	10493	819	179723
OR13-020	*R. vesiculosus*	*Villosuli*	United States	Oregon	Carter lake area, south of Florence	43 51 55.6 N, 124 5 35.7 W	31 meters	40141744	10692	21.45	13370	697	154638
CA13-019b	*R. vinicolor*	*Villosuli*	United States	Oregon	HW 29 north of Calistoga	38 39 50.2 N, 122 35 54.7 W	565 meters	35024968	9196	27.98	12323	660	174525
OR13-013	*R. vinicolor*	*Villosuli*	United States	Oregon	Metolius river trail, Deschutes NF, north of Lower Bridge	44 33 48.5 N, 121 37 9.7 W	847 meters	35455041	13806	17.63	1695	1695	5610
AZ12-009	*R. vinicolor*	*Villosuli*	United States	Arizona	Mt Lemmon, Coronado NF, Bearwallow road, Site #1	32 25 13.3 N, 110 43 52 W	2459 meters	36139051	7416	28.96	21121	409	232707
AZ12-022b	*R. vinicolor*	*Villosuli*	United States	Arizona	Mt Lemmon, Coronado NF, Incinerator Ridge Road	32 24 36.5 N, 110 42 34.2 W	2462 meters	35210942	9640	23.72	10950	841	110712
AZ12-102g	*R. vinicolor*	*Villosuli*	United States	Arizona	Snowbowl Road, Cononino NF, Site #1	35 19 23.3 N, 111 42 54.4 W	2789 meters	35159755	10229	23.79	10732	781	204277
UT12-150b	*R. vinicolor*	*Villosuli*	United States	Utah	Dixie NF, Site # 10	37 31 53 N, 112 45 22.5 W	2838 meters	34805438	9743	27.85	11957	715	126198
UT12-125	*R. cf. vulgaris*	*Roseoli*	United States	Utah	Dixie NF, Site # 3	37 31 43.2 N, 112 45 13.7 W	2841 meters	29523219	7160	40.67	16053	480	128491
Bonito1a	*R. fuscorubens*	*Rhizopogon*	United States	North Carolina	Duke University Forest, Greenlot			37674896	28620	20.75	2159	4606	32854
OR13-011	*R. sp.*	*Versicolores*	United States	Oregon	Metolius river trail, Deschutes NF, north of Lower Bridge CG	44 33 48.5 N, 121 37 9.7 W	847 meters	28115827	10429	38.85	6494	1046	79052
AZ12-036	*R. salebrosus*	*Amylopogon*	United States	Arizona	Chiricahua mountains, Coronado NF, Herb Martyr CG	31 52 20.1 N, 109 14 3.3 W	1770 meters	39527889	18010	18.86	5710	1634	104370

Table 4.2: A list of all PCR sequence data collected for Asian EM root tips and basidiomata. The sequences of PCR primers designed in this study are named based upon the JGI protein ID assigned to the protein ortholog in the *R. vinicolor* AM-OR11-026 reference genome. TW13-009-030 is the same collection locality as TW13-034c listed in Table 4.1. Collection locality for JP10-019c and JP10-029b are given in Mujic et al. 2014. The collection locality for CH13-017-038 is as follows: China, Jiangxi Province, Sanqinshan National Park, above "Wind Gate" near the Taoist Temple.

SAMPLE	SAMPLE TYPE	ITS	LSU	TEF	691883	697028	706455	765311	781237	789397
CH13-017-038	Chinese *Rhizopogon sp. nov.* EM root tip	X	X	X		X			X	X
TW13-009-030	Taiwanese *Rhizopogon sp. nov.* EM root tip	X	X		X	X	X	X	X	X
JP10-019c	Japanese *R. togasawariana* basidiome - HOLOTYPE	X	X	X	X	X	X	X	X	X
JP10-029b	Japanese *R. togasawariana* basidiome - PARATYPE	X	X	X						X

KEY TO FUNGAL PROTEIN IDs OF CUSTOM PRIMER PAIRS

PROTEIN ID	JGI ANNOTATION	FORWARD PRIMER	REVERSE PRIMER
691883	palmitoyl-protein thioesterase	TCGCAGGTCTCATCAACATC	CAGTTTGTGCTCAACGATCC
697028	sphingosine hydrolase	TTTTGCCATCGGTGAAACCG	TCTAGTCATGCGCGTACCAA
706455	Unknown protein	CGCGCAAATTACATCAAGAA	GGACATGCAAAAGCTACTTGG
765311	peptide alpha-n-acetyltransferase	CAGCGGACTTGGAACAGAGT	GTGCAACATGTCTCCGATTG
781237	unknown protein	CCACTTACCCATCGAGCTTC	TTCTGTGCCAAAAAGCAGGT
789397	dihydroorotate oxidase	TCACCTTCCATGGTGAACTG	CGCGTTTCCAGACAAAATCT

Chapter 5. Conclusion

This dissertation expands our understanding of comigration and cospeciation in ectomycorrhial (EM) symbionts and provides insight into fungal microevolutionary models constituent to these processes. We have performed the first experimental test of the hypothesis that resource heterogeneity affects competitive outcomes and vertical resource partitioning in ectomycorrhizal fungi. We demonstrate that the spatial heterogeneity of resources affects competitive outcome of EM interactions and identify the first known positive correlation between competitive success and the rate of reproductive success in an EM fungus. Deeper investigation of mating-type genetics in *R. vinicolor* and *R. vesiculosus* reveal that differential mating type systems of these fungi may reinforce the positive correlation between competitive success and mating in *R. vinicolor* and that fungi of the order Boletales possess structural differences in their mating type loci which correlate with bipolar and tetrapolar mating systems. Broadening our scope to the macroevolutionary level we have tested the systematic hypotheses of the genus *Rhizopogon* and provide support for the hypothesis that host specific EM fungi undergo comigratory patterns of dispersal with their EM hosts. The results of our investigations contribute to the literature of fungal mating type genetics, ecology, and evolutionary biology.

Through the course of our investigations many additional lines of inquiry became apparent and are worthy of consideration in future studies. In summary we recommend additional container plant studies that further investigate competition between individuals of *R. vinicolor* and *R. vesiculosus* to resolve the relationship between reproductive success and competitive success in these *Rhizopogon* species. Studies that characterize the mating genetics of additional strains of both *R. vinicolor* and *R. vesiculosus* would

build upon competition studies by investigating the potential for outcrossing in these species. By expanding this work to a continental geographic scale we could gain valuable insight into the historical migration and population divergence of *R. vinicolor* and *R. vesiculosus*. Finally, additional exploration of fragmented *Pseudostuga* populations for *Rhizopogon* species is needed to characterize EM community composition of these sensitive ecosystems. The following text will discuss these recommendations in detail.

The container plant experiments presented in chapter 2 of this dissertation were designed to investigate the roles of fungal competition and substrate stratification in driving patterns of vertical resource partitioning between *R. vinicolor* and *R. vesiculosus.* In the initial experimental design we had not intended these experiments to yield information on genotypic diversity and thus the genetic diversity of parent fungal genotypes was not extensively tested before experimental setup took place. As a result, the genotypic diversity of *R. vesiculosus* parent genotypes was too low to place confidence in the relationship between mating and competitive success in *R. vesiculosus.* Additional experiments with a robust genotypic sampling design would yield higher resolution of this relationship in both *R. vinicolor* and *R. vesiculosus*. The initial development of microsatellite markers for *R. vinicolor* and *R. vesiculosus* was performed by probing genomic DNA with short repeat oligonucleotides (Kretzer et al. 2000). While this technique yielded many high quality microsatellite markers for both *R. vinicolor* and *R. vesiculosus* the breadth of markers available through this method has largely been exhausted. The availability of genome sequences for both *R. vinicolor* and *R. vesiculosus* will allow for the rapid development of many additional microsatellite markers through in-silico detection. By selecting parent genotypes, which are polymorphic across existing

and newly developed microsatellite markers, for container plant experiments the statistical power of population analyses in the F1 generation could be greatly improved.

Future container plant experiments could be performed not only with the consideration of increased genotypic diversity but also in consideration of EM community succession. The current container plant experiments sampled only a single time point of community succession and it is possible that our sampling missed community transitions informative to the competitive strategies of *R. vinicolor* and *R. vesiculosus.* A chronological series of root tip sampling or differential temporal inoculations of the two species could be used to investigate the role of priority effects in structuring *Rhizopogon* communities and allow for the observation of community shifts during primary colonization of host seedlings (Kennedy and Bruns 2005, Kennedy et al. 2009). Additional replicates could be produced for destructive sampling at multiple time points (e.g. 3 months, 6 months, and 9 months) or container replicates could be inoculated with the spores of different species at different timepoints. Inclusion of more polymorphic microsatellite markers and implementation of a chronoseries sampling or inoculation design in future experiments would provide deeper resolution of competitive interactions between these species and provide insight into the relationship between mating success and competitive success in *R. vesiculosus*.

The genomes produced for *R. vinicolor* AM-OR11-026 and *R. vesiculosus* AM-OR11-056 were sequenced with the intention of producing high coverage dikaryotic genomes for the characterization of fungal mating-type loci. The genomes we have produced were well suited to this end and we have characterized the entire A and B mating type locus for these fungi and several reference Boletales genomes. However, our

work has characterized only a single individual of each species that was derived from populations existing at the same geographic locality. Future work could expand upon the results presented here by characterizing mating locus organization and gene content for several more isolates from across the range of both *R. vinicolor* and *R. vesiculosus.* This work could be advanced using both genomic as well as PCR based approaches. The genetic diversity of pheromone receptor gene alleles is expected to be very high for outcrossing (tetrapolar) basidiomycetes (Van Peer et al. 2011, James et al. 2013) and the genetic diversity of alleles at these loci can be used to infer the mating-type system of Agaricomycetes which are not suitable for traditional in-vitro mating type determination (James et al. 2013). Species-specific PCR primers could be designed for the B-locus pheromone receptor genes using the genome strains of *R. vinicolor* and *R. vesiculosus* as reference. The mating type systems for both species could be inferred using this PCR based approach by characterizing the genetic diversity of pheromone receptor genes for multiple strains. Additional high coverage genomes of *R. vinicolor* and *R. vesiculosus* from the geographic range of both species would provide valuable insight into the organization of mating type loci of both species and provide insight into mating-type evolution in EM fungi with limited dispersal. Additional genomes from monokaryotic strains would be particularly valuable and help to disambiguate haplotypic sequence data. R*hizopogon* spores do not readily germinate in culture to produce monokaryons but specialized culture techniques have been developed using sterile host roots to induce spore germination in *R. roseolus* (Kawai et al. 2008). A protocol for the isolation of monokaryotic strains of *R. vinicolor* and *R. vesiculosus* could potentially be developed through similar methods.

Current sampling efforts have recovered a much greater diversity of *R.* subgenus *Villosuli* species in North America than in Asia. This finding is consistent with an origin of the *Rhizopogon-Pseudotsuga* EM symbiosis in western North America as the relationship would have had a longer evolutionary history in this geographic region and a greater potential for population and species level divergence events. However, our sampling efforts represent only a small timespan within a single season at most localities and it is possible that additional *Rhizopogon* species display fruiting phenologies that fall outside of the range our sampling. It is also unclear in our preliminary sampling efforts if we have captured the complete diversity of *R.* subgenus *Villosuli* species at the range limits of *Pseudotsuga* hosts. The fragmented populations of *Pseudotsuga* species in Asia (Murata et al. 2013), southern California (Lombardo et al. 2009), and Mexico (Gugger et al. 2011) are relict distributions of species that once possessed a broader geographic range. In the case of Mexican *P. menziesii* var. *glauca* these isolated populations are in a state of notable decline from the lack of gene flow between populations and subsequent inbreeding depression (Vargas-Hernández et al. 2004, Mápula-larreta et al. 2007, Velasco-García et al. 2007). The potential for development of endemic *Rhizopogon* species could be high given isolated host populations of sufficient size. However, local extinctions of *Rhizopogon* species during the fragmentation of relict host populations seems equally likely given the small size and low genetic diversity of Mexican host populations. Additional full season sampling for basidiomata in *Pseudotsuga* forests outside of the Pacific Northwest could yield additional *R.* subgenus *Villosuli* species. *Rhizopogon* species are encountered as one of the most abundant EM species colonizing seedling roots in plant bioassays using *Pseudotsuga* forest soil from both *P. japonica*

(Murata et al. 2013) and *P. menziesii* stands (Luoma et al. 2006). Therefore, container plant bioassays utilizing soils from *Pseudotsuga* forests as inoculum could prove an effective screening technique for detection of unknown *Rhizopogon* species.

The nature of the divergence event which separated the sister species *R. vinicolor* and *R. vesiculosus* is of particular interest for future study. We sampled basidiomata extensively throughout stands of *P. menziesii* var. *menziesii* of coastal western North Amarica and *P. menziesii* var. *glauca* in the intermountain west. Though we recovered dozens of *R. vinicolor* basidiomata from sites in the intermountain west we failed to recover *R. vesiculosus* from any of these interior stands. These two species share a similar fruiting phenology and occur at similar frequencies in the Pacific Northwest (Kretzer et al. 2005, Beiler et al. 2010, Beiler et al. 2012, Dunham et al. 2013) and the apparent absence of *R. vesiculosus* from the intermountain west sampling suggests that *R. vesiculosus* may be restricted to a coastal distribution. However, the stands of *P. menziesii* var. *glauca* sampled in the intermountain west consisted mostly of mature trees and recruitment levels in xeric interior stands of *P. menziesii* var. *glauca* are naturally lower compared to the more mesic sites where coastal *P. menziesii* var. *menziesii* occurs (Hamman and Wang 2006, Klenner et al. 2008). It is possible that *R. vesiculosus* is specialized upon disturbed or young host stands and the older interior stands sampled had natural attrition from populations of early successional *R. vesiculosus*. There is anecdotal evidence to support this hypothesis as EM communities sampled from mature *P. menziesii* var. *glauca* stands in interior British Columbia contain only *R. vinicolor* where young stands from nearby sites contain a mixed community of both *R. vesiculosus* and *R vinicolor* (Dr. Daniel Durall, personal communication). Container plant bioassay

experiments using soil inoculum from across the intermountain range of *P. menziesii* var. *glauca* could be used to test this hypothesis. A deeper level community sample from selected sites throughout the range of *P. mensziesii* could also be used reveal the distribution and regions of geographic origin for these two *Rhizopogon* species. The population structure and post-glacial migration history of *P. menziesii* has been characterized using coalescent analysis of mitochondrial and chloroplast DNA (Gugger et al. 2010, Gugger et al. 2011). With a deeper sampling of several populations across the range of both *R. vinicolor* and *R. vesiculosus* a similar coalescent approach could be applied to these species. The set of protein coding loci identified in our phylogeographic barcoding analysis of *R.* subgenus *Villosuli* would be well suited to this end. The results of this analysis could provide clues as to the geographic origins of these species and determine, with greater resolution, the comigrational history of both taxa with *P. menziesii*.

REFERENCES:

Beiler KJ, Durall DM, Simard SW, Maxwell SA and Kretzer AM. 2010. Architecture of the wood-wide web*: Rhizopogon spp*. genets link multiple Douglas-fir cohorts *New Phytologist* 185(2) 543-553.

Beiler, Kevin J., Suzanne W. Simard, Valerie LeMay, and Daniel M. Durall. 2012. Vertical Partitioning between Sister Species of *Rhizopogon* Fungi on Mesic and Xeric Sites in an Interior Douglas-Fir Forest. *Molecular Ecology* 21 (24): 6163–74. doi:10.1111/mec.12076.

Dunham, Susie M., Alija Bajro Mujic, Joseph W. Spatafora, and Annette M. Kretzer. 2013. Within-Population Genetic Structure Differs between Two Sympatric Sister-Species of Ectomycorrhizal Fungi, *Rhizopogon vinicolor* and *R. vesiculosus*. *Mycologia* 105 (4): 814–26. doi:10.3852/12-265.

Gugger, Paul F., Shinya Sugita, and Jeannine Cavender-Bares. 2010. Phylogeography of Douglas-Fir Based on Mitochondrial and Chloroplast DNA Sequences: Testing Hypotheses from the Fossil Record. *Molecular Ecology* 19 (9): 1877–97. doi:10.1111/j.1365-294X.2010.04622.x.

Gugger, Paul F., Antonio González-Rodríguez, Hernando Rodríguez-Correa, Shinya Sugita, and Jeannine Cavender-Bares. 2011. Southward Pleistocene Migration of Douglas-Fir into Mexico: Phylogeography, Ecological Niche Modeling, and Conservation of 'rear Edge' Populations. *New Phytologist* 189 (4): 1185–99. doi:10.1111/j.1469-8137.2010.03559.x.

Hamman A, Wang T. 2006. Potential effects of climate change on ecosystem and tree species distribution in British Columbia. *Ecology* **87:** 2773–2786.

James, Timothy Y., Sheng Sun, Wenjun Li, Joseph Heitman, Hsiao-Che Kuo, Yong-Hwan Lee, Frederick O. Asiegbu, and Åke Olson. 2013. Polyporales Genomes Reveal the Genetic Architecture Underlying Tetrapolar and Bipolar Mating Systems. *Mycologia* 105 (6): 1374–90. doi:10.3852/13-162.

Kawai, Masataka, Mina Yamahara, and Akira Ohta. 2008. Bipolar Incompatibility System of an Ectomycorrhizal Basidiomycete, *Rhizopogon rubescens*. *Mycorrhiza* 18 (4): 205–10. doi:10.1007/s00572-008-0167-4.

Kennedy, Peter G., and Thomas D. Bruns. 2005. Priority Effects Determine the Outcome of Ectomycorrhizal Competition between Two *Rhizopogon* Species Colonizing *Pinus muricata* Seedlings. *The New Phytologist* 166 (2): 631–38. doi:10.1111/j.1469-8137.2005.01355.x.

Kennedy, Peter G., Kabir G. Peay, and Thomas D. Bruns. 2009. Root Tip Competition among Ectomycorrhizal Fungi: Are Priority Effects a Rule or an Exception? *Ecology* 90 (8): 2098–2107. doi:10.1890/08-1291.1.

Klenner W, Walton R, Arsenault A, Kramseter L. 2008. Dry forests in the southern interior of British Columbia: historic disturbances and implications for restoration and management. *Forest Ecology & Management* **206:** 1711–1722.

Kretzer, Annette M., Randy Molina, and Joseph W. Spatafora. 2000. Microsatellite Markers for the Ectomycorrhizal Basidiomycete *Rhizopogon vinicolor*. *Molecular Ecology* 9 (8): 1190–91. doi:10.1046/j.1365-294x.2000.00954-12.x.

Kretzer AM, Dunham S, Molina R, Spatafora JW. 2005. Patterns of vegetative growth and gene flow in *Rhizopogon vinicolor* and *R. vesiculosus* (Boletales, Basidiomycota) *Molecular Ecology* 14(8) 2259-2268

Lombardo KJ, Swetnam TW, Baisan CH, Borchert MI. 2009. Using bigcone Douglas-fir fire scars and tree rings to reconstruct interior chaparral fire history. *Fire Ecology* 5(3): 35–56

Luoma, Daniel L, Christopher A Stockdale, Randy Molina, and Joyce L Eberhart. 2006. The Spatial Influence of *Pseudotsuga menziesii* Retention Trees on Ectomycorrhiza Diversity. *Canadian Journal of Forest Research* 36 (10): 2561–73. doi:10.1139/x06-143.

Mápula-larreta, M., J. López-Upton, J. J. Vargas-Hernández, and A. Hernández-Livera. 2007. Reproductive Indicators in Natural Populations of Douglas-Fir in Mexico. *Biodiversity and Conservation* 16 (3): 727–42. doi:10.1007/s10531-005-5821-y.

Murata, Masao, Akihiko Kinoshita, and Kazuhide Nara. 2013. Revisiting the Host Effect on Ectomycorrhizal Fungal Communities: Implications from Host–fungal Associations in Relict *Pseudotsuga japonica* Forests. *Mycorrhiza* 23 (8): 641–53. doi:10.1007/s00572-013-0504-0.

Van Peer, Arend F., Soon-Young Park, Pyung-Gyun Shin, Kab-Yeul Jang, Young-Bok Yoo, Young-Jin Park, Byoung-Moo Lee, Gi-Ho Sung, Timothy Y. James, and Won-Sik Kong. 2011. Comparative Genomics of the Mating-Type Loci of the Mushroom *Flammulina velutipes* Reveals Widespread Synteny and Recent Inversions. *PLoS ONE* 6 (7): e22249. doi:10.1371/journal.pone.0022249.

Vargas-Hernández JJ, López-Upton J, Reyes-Hernández VJ, Domínguez-Alvarez A, Mápula-Larreta M. 2004. Natural populations of Douglas fir in Mexico: current status and needs for conservation. In: Beaulieu J, ed. *Silviculture and the conservation of genetic resources for sustainable forest management: proceedings of the symposium of the North American forest commission, forest genetic resources and silviculture working groups, and the International Union of Forest Research Organizations (IUFRO). Quebec City, QC, Canada, 26–36.*

Velasco-García, MV, J López-Upton, G Angeles-Pérez, J Vargas-Hernández, and V Guerra-de la Cruz. 2007. *Pseudotsuga menziesii* Seed Dispersion in Populations of Central México. *Agrosciencia* 41 (1): 121–31.

Appendix 1: Perl code written to facilitate analyses in this dissertation

```
#!/usr/bin/perl -w

#Name blast_sequence_extractor.pl - Extracts blast hit sequences from a
directory of .blast output files in .xml format.
#               See usage() function at the end of program for details and
usage instructions.

use strict;
use Getopt::Std;
use File::Basename;
use Bio::SeqIO;
use Bio::SearchIO;
use Bio::DB::Fasta;
use List::Util qw(min max);

sub reverse_complement($);
sub trim($);
sub usage;


#declare the global variables
#hashes
my %Ref_sequence_length;
my %Blast_sequences;
my %Blast_seq_count;
my %Blast_seq_count_cumulative;
my %Ref_sequence;

#my %Protein_length;
#my %Protein_SNPcount;

#arrays
my @Assembly_list;
my @Blast_list;

my @Protein_holder;
my @Protein_current;
my @Exon_holder;
my @Exon_current;
my @Exons_toPrint;
my @target_list;

#strings
my $current_strand;
my $Temp_contig;
#my $Current_contig;
my $Last_contig;
my $Current_protein;
my $next_line;
my $Fail_string;
my $active_seq;
my $temp_header;
my $active_protein_seq;
my $temp_protein_header;


#integers
my $Ref_seq_count=0;
my $Print_seq_count=0;
my $Sample_count=0;
```

```
my $Reverse_flag=0;
my $Fail_flag_120=0;
my $Fail_flag_960=0;
my $Failed_proteins=0;
my $Passed_proteins=0;
my $Exons_totalLength=0;
my $Passed_exons_totallength=0;

#declare the options string
my $opt_string='ha:b:r:s:p:PVe:';
my $prg = basename $0;
my %opt;
my $output_header;
my $list_file;
my $gff_file;
my $reference_file;
my $contig_file;
my $Blast_directory;
my $Assembly_directory;
my $f_sfx;
my $reference_prefix;
my $Price_flag=0;
my $Velvet_flag=0;
my $Sig_thresh;

####process the options and set global variables for options
getopts("$opt_string", \%opt) or usage();
$opt{h} and usage();
usage() unless ($opt{r} && $opt{b} && $opt{a} && $opt{s} && $opt{p});
usage() unless ($opt{P} || $opt{V});
if($opt{P} && $opt{V}){ print "You must specify either the -P option or the -V
option, but not both\n" and usage();}
print "Could not open file: $opt{r}\n" and die(usage()) unless (-e $opt{r});
print "Could not read directory: $opt{b}\n" and die(usage()) unless (-e $opt{b}
);
print "Could not read directory: $opt{a}\n" and die(usage()) unless (-e $opt{a}
);
print "Bad file suffix provided: $opt{s}\n" and die(usage()) unless (length
($opt{s}));
print "Bad Reference ID provided: $opt{p}\n" and die(usage()) unless (length
($opt{p}));
if($opt{e}){print "Bad significance threshold provided: $opt{e}\n" and die
(usage()) unless (length($opt{e}));}

#set the input files
$reference_file=$opt{r};
$Blast_directory=$opt{b};
$Assembly_directory=$opt{a};
$f_sfx=$opt{s};
$reference_prefix=$opt{p};
$Velvet_flag= $opt{V} ? 1 : 0;
$Price_flag= $opt{P} ? 1 : 0;
$Sig_thresh= $opt{e} ? $opt{e} : 1e-5;

# open up BLAST and ASSEMBLY directories and reference file
opendir(BLASTDIR, $Blast_directory) or die "can't open directory:
$Blast_directory $!\n";
opendir(ASMDIR, $Assembly_directory) or die "can't open directory:
$Assembly_directory $!\n";
```

```
#open(REF, $reference_file) or die "can't open reference file: $reference_file
\n";

# if input directory does not end with '/', add it.
#This is for the purposes of printing files later on, no need for the opening
of the directory.
if ($Blast_directory !~ /\/$/){
        $Blast_directory .= "/";
}
if ($Assembly_directory !~ /\/$/){
        $Assembly_directory .= "/";
}
######inform the user of program settings
print STDERR "Reference sequence file set to : $reference_file\n";
print STDERR "Extracting all top sequence hits from $f_sfx files in directory:
$Blast_directory\n";
print STDERR "Using sequence data from the assemblies stored in the directory:
$Assembly_directory\n";

#declare the sequence handle and container stuff
my $sequence_handle=Bio::SeqIO->new( -file=>$reference_file, -format=>"Fasta");
my $db = Bio::DB::Fasta->new($reference_file);
my @active_header;

#populate the Reference sequence hash with all of the reference sequences
while(my $seq_obj=$sequence_handle->next_seq){
        $temp_header=$db->header($seq_obj->id);
        chomp($temp_header);
        @active_header=split(/\./, $temp_header);
        $active_seq= $seq_obj->seq;
        #insert the sequence into a new hash
        $Ref_sequence{$active_header[1]}=$active_seq;
        $Ref_sequence_length{$active_header[1]}=length($active_seq);
        #Increment the total number of sequences in reference and initialize
the count for this
        #sequence in the seq count hash.
        $Ref_seq_count++;
        $Blast_seq_count{$active_header[1]}=0;
}#end while

#populate the Assembly list array
foreach my $assembly_file (readdir (ASMDIR)){
        next if ($assembly_file =~ /^\./); # skip if this file is default '.'
or '..' directory
        next if ($assembly_file !~ /.fasta$/); # die if fasta files do not end
does not end with the right file suffix
        push (@Assembly_list, basename $assembly_file);
}

#populate the BLAST list array
foreach my $blast_file (readdir (BLASTDIR)){
        next if ($blast_file =~ /^\./); # skip if this file is default '.' or
'..' directory
        next if ($blast_file !~ /$f_sfx$/); # skip if this file does not end
with the right file suffix
        push (@Blast_list, basename $blast_file);

}

#now sort the arrays
```

```
@Assembly_list = sort @Assembly_list;
@Blast_list = sort @Blast_list;

#check to see that arrays are the same length
if (@Assembly_list != @Blast_list){print "Assembly and blast directories
contain different numbers of files\n" and die();}

$Sample_count=(@Blast_list);

##OK, we got this far, let's start creating some directories to house our files
mkdir "Blast-hits_by-Sample" or die("Could not make sample file output
directory: $! \n");
mkdir "Blast-hits_by-Protein" or die("Could not make protein file output
directory: $! \n");

################################BEGIN blast file processing
#loop over each file in the blast directory and extract all of the sequences
while(@Blast_list){

        #Open the current blast hash file
        my $current_blast_file= shift(@Blast_list);

        #Populate the Current assembly hash with the assembly cooresponding to
the current blast file
        #first clear out the current assembly hash.
        my %current_assembly=();
        my $current_assembly_file= shift(@Assembly_list);

        #first test to see that these files are coming from the same Sample
        my @blast_file_array = split(/_/,$current_blast_file);
        my @assembly_file_array = split(/_/,$current_assembly_file);
        if($blast_file_array[0] ne $assembly_file_array[0]){print "The files
$current_blast_file and $current_assembly_file have disimilar Sample ID's.
Exiting with error\n" and die();}

        #OK, now we can open the assembly file and populate the assembly hash
        #declare the sequence handle and container stuff
        my $assembly_handle=Bio::SeqIO->new( -file=>$Assembly_directory.
$current_assembly_file, -format=>"Fasta");
        my $assembly_db = Bio::DB::Fasta->new($Assembly_directory.
$current_assembly_file);
        my @assembly_header;

        #populate the Reference sequence hash with all of the reference
sequences
        while(my $seq_obj=$assembly_handle->next_seq){
                my $contig_header=$assembly_db->header($seq_obj->id);
                chomp($contig_header);
                if($Price_flag){
                        @assembly_header=split(/ /, $contig_header);
                        @assembly_header=split(/_/, $assembly_header[0]);
                }
                if($Velvet_flag){
                        @assembly_header=split(/_/, $contig_header);
                }

                my $active_seq= $seq_obj->seq;
                #insert the sequence into a new hash
                $current_assembly{$assembly_header[1]}=$active_seq;
        }#end while
```

```
        #create an output file handle for blast sequence information in this
blast file
        my $sample_handle=Bio::SeqIO->new( -file=>">Blast-hits_by-Sample/
$blast_file_array[0]_BEST-hits.fasta", -format=>"Fasta");
        my $seq_obj_out;

        my $current_query;
        #Now, initialize the blast SearchIO file and get ready to loop over
the results of the blast file
        my $blast_handle=Bio::SearchIO->new( -file=>$Blast_directory.
$current_blast_file, -format=>"blastxml");
        #print $Blast_directory.$current_blast_file ."\n" and die();
        while(my $result = $blast_handle->next_result){

                my $current_query = $result->query_name;
                my @name_holder = split(/\./, $current_query);
                #we want only the top hit
                my $hit;
                ##skip this result if there are no hits
                if($result->num_hits > 0){$hit = $result->next_hit;}else
{next;}
                ##set a threshold for hit significance from the command line
if one was specified
                ##skip if this hit is below the threshhold
                if($hit->significance > $Sig_thresh){next;}

                #OK, we have a hit and it passed muster. let's proceed
                my $hit_name = $hit->name;
                chomp($hit_name);
                my @hit_header;
                if($Price_flag){
                        @hit_header=split(/ /, $hit_name);
                        @hit_header=split(/_/, $hit_header[0]);
                }
                if($Velvet_flag){
                        @hit_header=split(/_/, $hit_name);
                }
                #print "$Blast_directory$blast_file_array[0]_BEST-hits.fasta
\n" and die();

                #if there is only one HSP, set the singleton flag
                my $HSP_single = ($hit->num_hsps == 1) ? 1 : 0;
                #set a few values for the sake of storing start and end of
protein region in case
                #we have multiple HSPs. Also make a flag to denote if this is
the first HSP we are viewing
                my $HSP_min_val=0;
                my $HSP_max_val=0;
                my $HSP_first=1;
                my $HSP_strand=1;

                while(my $hsp = $hit->next_hsp){
                        my $current_seq = $hsp->hit_string;
                        #is this the first HSP? if so, set the min and max
values of the hit for checking later
                        if($HSP_first){
                                $HSP_min_val=min($hsp->range('hit'));
                                $HSP_max_val=max($hsp->range('hit'));
                                $HSP_first=0;
```

```
                                #set the strand of this region
                                if(($hsp->strand('hit') == -1) && ($hsp->strand
('query') == 1)){$HSP_strand = -1;}
                                if(($hsp->strand('hit') == 1) && ($hsp->strand
('query') == -1)){$HSP_strand = -1;}
                                if(($hsp->strand('hit') == -1) && ($hsp->strand
('query') == -1)){$HSP_strand = -1;}
                        }

                        #if there is just a single HSP in this hit, then
insert the sequence and be done with it
                        if($HSP_single){

                                #is the query or hit reversed? if so, reverse
complement the hit string
                                if($HSP_strand == -1)
{$current_seq=reverse_complement($current_seq);}

                                if($Blast_sequences{$name_holder[1]}){
                                        #protein has hit info already, add
this one to the list
                                        $Blast_sequences{$name_holder[1]} .=
"\t$blast_file_array[0] $current_seq";
                                }else{
                                        #protein has no no info, create this
entry as the first
                                        $Blast_sequences{$name_holder[1]} =
"$blast_file_array[0] $current_seq";
                                }
                                #now let's set the header, reverse complement
if neccessary and then print out
                                #the sequence data to a file
                                my $temp_protein_header = "$blast_file_array
[0].$name_holder[1]";

                                $seq_obj_out= Bio::Seq->new(-
seq=>"$current_seq", -display_id=>"$temp_protein_header", -alphabet=>"dna");
                                $sample_handle->write_seq($seq_obj_out);
                                #increment the counter for this protein to
show that we printed a sequence for this protein
                                $Blast_seq_count{$name_holder[1]}+=1;

                        }
                        #exit the loop if we have just one HSP
                        last if($HSP_single);

                        #OK, we have more than one HSP, let's get coordinates
of the entire hit and grab
                        #the sequence from the assembly sequences on
record
                        #check to see if the current HSP is below or above the
range of the current
                        #min and max values. We want to print the entire range
                        if($HSP_min_val > min($hsp->range('hit')))
{$HSP_min_val=min($hsp->range('hit'));}
                        if($HSP_max_val < max($hsp->range('hit')))
{$HSP_max_val=max($hsp->range('hit'));}
                }

                #if multiple HSP information was analyzed, enter the sequence
```

```
now
                if(!$HSP_single){
                        my $current_contig=$hit_header[1];
                        #print "$current_contig \n" and die();
                        #grab just the protein_sequence for the target protein
being examined
                        ##############reverse complement if neccessary, and
then print to file

                        my $active_protein_seq=substr($current_assembly
{$current_contig}, ($HSP_min_val - 1), ($HSP_max_val - $HSP_min_val + 1));
                        #is the query reversed? if so, reverse complement the
hit string
                        if($HSP_strand == -1)
{$active_protein_seq=reverse_complement($active_protein_seq);}

                        if($Blast_sequences{$name_holder[1]}){
                                #protein has hit info already, add this one to
the list
                                $Blast_sequences{$name_holder[1]} .= "\t
$blast_file_array[0] $active_protein_seq";
                        }else{
                                #protein has no no info, create this entry as
the first
                                $Blast_sequences{$name_holder[1]} =
"$blast_file_array[0] $active_protein_seq";
                        }

                        #print the complete protein sequence to the sample file
                        my $temp_protein_header .= "$blast_file_array[0].
$name_holder[1]";
                        $seq_obj_out= Bio::Seq->new(-
seq=>"$active_protein_seq", -display_id=>"$temp_protein_header", -
alphabet=>"dna");
                        $sample_handle->write_seq($seq_obj_out);
                        #increment the counter for this protein to show that
we printed a sequence for this protein
                        $Blast_seq_count{$name_holder[1]}
+=1;
                }
        }#end result processing loop
}#end blast file processing loop

####OK, now that we've gathered all of the sequences from all of the files,
let's print a file for each alignment into
###the folder BLAST-hits_by-Protein/
while(my ($key, $value) = each %Blast_sequences){
        my $protein_handle=Bio::SeqIO->new( -file=>">Blast-hits_by-Protein/".
$key."_BEST-hits.fasta", -format=>"Fasta");

        #first, let's print the reference sequence if the $Ref_sequence hash
contains that protein
        if(exists($Ref_sequence{$key})){
                my $seq_obj_out=Bio::Seq->new(-seq=>"$Ref_sequence{$key}", -
display_id=>$reference_prefix.".$key", -alphabet=>"dna");
                $protein_handle->write_seq($seq_obj_out);
        }
        #now, let's loop over all the sequences stored for this protein and
print them to the cooresponding file
        my @protein_hits = split(/\t/,$value);
```

```
        while(@protein_hits){
                my $hit_list=shift(@protein_hits);
                my @seq_info = split(/ /, $hit_list);
                my $seq_obj_out=Bio::Seq->new(-seq=>"$seq_info[1]", -
display_id=>"$seq_info[0]".".$key", -alphabet=>"dna");
                $protein_handle->write_seq($seq_obj_out);
                $Print_seq_count++;
        }
}

#OK, we've printed all the material that needs to be printed. Let's exit with
some info and print one last file
#containing a list of proteins for which there is missing data.
my $Missing_proteins=0;
my @Missing_data;

while(my ($key, $value) = each %Blast_seq_count){
        my $missing_samples= $Sample_count - $value;
        if($missing_samples > 0){
                $Missing_proteins++;
                push(@Missing_data, "$key\t$missing_samples")
        }
}

print STDERR "Printed $Print_seq_count sequences for $Sample_count Samples\n";

if($Missing_proteins>0){
        open(INFO, ">>Proteins_missing-data.txt") or die("Could not open
output file Proteins_missing-data.txt: $!");
        while(@Missing_data){
                print INFO shift(@Missing_data)."\n";
        }
        print STDERR "There were $Missing_proteins proteins that were lacking
data for 1 or more samples.\n";
        print STDERR "Information on missing proteins in the format \"protein
<tab> number-of-missing-samples\"\n";
        print STDERR "was printed to the file \"Proteins_missing-data.txt\"
\n";
}

##################Begin PERL subroutine definitions##################
#####################################################################

#subroutine to reverse complement a string of DNA sequence
sub reverse_complement($) {
        my $dna = shift;

        # reverse the DNA sequence
        my $revcomp = reverse($dna);

        # complement the reversed DNA sequence
        $revcomp =~ tr/NACGTacgt/NTGCAtgca/;
        return $revcomp;
}

# Perl trim function to remove whitespace from the start and end of the string
sub trim($)
{
        my $string = shift;
        $string =~ s/^\s+//;
```

```
        $string =~ s/\s+$//;
        return $string;
}

# PERL usage function for this program
sub usage
{
        print STDERR << "EOF";

        Name $prg - Extracts blast hit sequences from a directory of .blast
output files in .xml format.
                        You must provide two directory paths, one directory
full of .fasta assembly files
                        and a second directory full of .blast searches against
those assemblies.
                        The number of files in each directory and all Sample
IDs must be identical.
                        The program expects .blast taxon input files in
format:
                        "Sample-ID_type-of-blast.blast".
                        The program expects .fasta assembly input files in
format:
                        "Sample-ID_type-of-assembly.fasta".
                        The placement of of "_" is significant as the program
will split file names
                        basedupon this character.
                        Fasta files MUST end in .fasta.
                        Expected protein header format of original file
queried:
                        scaffold_id.protein_id.protein_start.protein_end.strand

                        The program will create two directories:
                        Blast-hits_by-Sample/ -> contains .fasta output files
each containing all the sequences
                                                 for all proteins extracted
from a particular taxon. File
                                                 output format:
                                                 Taxon-ID_proteins.fasta
                        Blast-hits_by-Protein/ -> contains .fasta output files
each containing sequences
                                                 of a single protein from each
taxon file. File output format:
                                                 Protein-ID.fasta

        usage:  $prg (-l input) (-g JGI_input) (-c fasta_input) [-a header]

        -h              :       print this help message
        -V              :       Specifies that input assemblies were
constructed with Velvet, mutually exclusive with -P.
        -P              :       Specifies that input assemblies were
constructed with PRICE, mutually exclusive with -V.
        -a directory    :       A directory full of .fasta files with
assemblies queried in .blast files.
                                Sets input file.
        -b directory    :       A directory full of .blast files in xml format,
sets input files
        -r fasta_input  :       Fasta format file defining nucleotide
sequences of reference queries
                                used in blast analysis, sets inputfile
        -s file_suffix  :       Sets the suffix of files to process (should
```

```
usually be ".blast")
        -p Reference_ID :       Sets the Sample ID to use for naming the
reference sequences in output files
        -e threshold    :       Sets the e-value threshold for the upper limit
of hits to be included in output,
                                (Default: 1e-5)

        ex: $prg -d my_directory/ -r reference_sequences.fasta -s .blast -p
OR11-026

EOF
        exit;

}
```

```
#!/usr/bin/perl -w

#Name cluster_ref_extractor_V1.pl - Extracts proteins (listed by JGI protein
ID) for
#               the target taxon that cluster with particularproteins
specified from
#               a list in the reference taxa.
#       See usage() function at the end of program for details and usage
instructions.

#use warnings;
use strict;
use Getopt::Std;
use File::Basename;
use Bio::SeqIO;
use Bio::DB::Fasta;


sub reverse_complement($);
sub trim($);
sub usage;


#declare the global variables
#hashes
my %Contig_length;
my %Protein_map_ID;
my %Protein_info;
my %Reference_list;
my %Contig_sequence;
my %target_list;
#my %Protein_length;
#my %Protein_SNPcount;
my %Protein_exons;

#arrays
my @Protein_holder;
my @Protein_current;
my @Exon_holder;
my @Exon_current;
my @Exons_toPrint;
my @target_list;

#strings
my $current_strand;
my $Temp_contig;
my $Current_contig;
my $Last_contig;
my $Current_protein;
my $next_line;
my $Fail_string;
my $active_seq;
my $temp_header;
my $active_protein_seq;
my $temp_protein_header;


#integers
my $jgi_flag=0;
my $maker_flag=0;
```

```
my $Reverse_flag=0;
my $Fail_flag_120=0;
my $Fail_flag_960=0;
my $Failed_proteins=0;
my $Passed_proteins=0;
my $Exons_totalLength=0;
my $Passed_exons_totallength=0;

#declare the options string
my $opt_string='hl:g:c:a:b:r:m:';
my $prg = basename $0;
my %opt;
my $reference_header;
my $target_header;
my $list_file;
my $reference_file;
my $gff_file;
my $contig_file;

####process the options and set global variables for options
getopts("$opt_string", \%opt) or usage();
$opt{h} and usage();
usage() unless ($opt{l} && $opt{c} && $opt{a} && $opt{b} && $opt{r});
usage() if (($opt{g} && $opt{m}) || (!$opt{g} && !$opt{m}));

$reference_header = $opt{a};
$target_header = $opt{b};

######inform the user of program settings
print STDERR "Reference taxon set to: $reference_header\n";
print STDERR "Target taxon set to: $target_header\n";
$opt{l} and print STDERR "Extracting single copy protein coding genes for
target taxon from cluster list specified in input file: $opt{l}which match
proteins from the list specified in: $opt{r}\n";
$opt{g} and print STDERR "Using protein coordinates from input file: $opt{g}
\n";
$opt{c} and print STDERR "and nucleotide sequences from input file: $opt{c}\n";

$jgi_flag= $opt{g} ? 1 : 0;
$maker_flag = $opt{m} ? 1 :0;

#set the input files
$list_file = $opt{l};
$reference_file = $opt{r};
if($jgi_flag){$gff_file=$opt{g};}
if($maker_flag){$gff_file=$opt{m};}
$contig_file = $opt{c};

#declare the sequence handle and container stuff
my $sequence_handle=Bio::SeqIO->new( -file=>$contig_file, -format=>"Fasta");
my $db = Bio::DB::Fasta->new($contig_file);
my @active_header;


#check the indices here, this is not yet determined to work for JGI headers
#populate the Contig sequence hash with all of the contig sequences
while(my $seq_obj=$sequence_handle->next_seq){
        $temp_header=$db->header($seq_obj->id);
        chomp($temp_header);
        @active_header=split(/_/, $temp_header);
```

```
        $active_seq= $seq_obj->seq;
        #insert the sequence into a new hash
        $Contig_sequence{$active_header[1]}=$active_seq;
        $Contig_length{$active_header[1]}=length($active_seq);
}#end while


open(PROT_FILE, $gff_file) or die("could not open protein coordinates file
$gff_file: $!\n");
#open(CONTIG_FILE, "$ARGV[@ARGV-1]") or die("could not open contig file $ARGV
[@ARGV-1]: $!\n");
open(TARGET_FILE, $list_file) or die("could not open protein cluster
definition file $list_file: $!\n");
open(REFERENCE_FILE, $reference_file) or die("could not open protein reference
definition file $reference_file: $!\n");

#populate the reference array
while(<REFERENCE_FILE>){
        chomp;
        $Reference_list{$_}=" ";
}


###############first loop over the cluster definition file and get all those
protein id from the taxon defined by -a
############## which belong to clusters that are single copy protein coding
genes
#########here we will be adding entries to a new hash that maps the proteins
back to

while(<TARGET_FILE>){
        chomp;
        #scan over each line and grab those lines that contain clusters with
single copy protein coding genes (4 genes, 4 taxa)
        #$1 =cluster name $2= protein_id from taxon specified by -a option $3
= protein_id from taxon specified by-b option
        if(m/ORTHOMCL([0-9]+) \(4 genes,4 taxa\):.*jgi\|$reference_header\|
([0-9]+)\|.*jgi\|$target_header\|([0-9]+)\|/){
                if(exists($Reference_list{$2})){
                        $target_list{$3}=$2;
                }
        }

        #scan over each line and grab those lines that contain clusters with
single copy protein coding genes (4 genes, 4 taxa)
        #$1 =cluster name $3= protein_id from taxon specified by -a option $2
= protein_id from taxon specified by-b option
        if(m/ORTHOMCL([0-9]+) \(4 genes,4 taxa\):.*jgi\|$target_header\|([0-9]
+)\|.*jgi\|$reference_header\|([0-9]+)\|.*/){
                if(exists($Reference_list{$3})){
                        $target_list{$2}=$3;
                }
        }

        #scan over each line and grab those lines that contain clusters with
single copy protein coding genes (4 genes, 4 taxa)
        #$1 =cluster name $2= protein_id from taxon specified by -a option $3
= protein_id from taxon specified by-b option
        if(m/ORTHOMCL([0-9]+) \(4 genes,4 taxa\):.*jgi\|$reference_header\|
([0-9]+)\|.*$target_header\.([0-9]+)/){
```

```
                if(exists($Reference_list{$2})){
                        $target_list{$3}=$2;
                }
        }

        #scan over each line and grab those lines that contain clusters with
single copy protein coding genes (4 genes, 4 taxa)
        #$1 =cluster name $3= protein_id from taxon specified by -a option $2
= protein_id from taxon specified by-b option
        if(m/ORTHOMCL([0-9]+) \(4 genes,4 taxa\):.*$target_header\.([0-9]
+).*jgi\|$reference_header\|([0-9]+)\|/){
                if(exists($Reference_list{$3})){
                        $target_list{$2}=$3;
                }
        }
}#end cluster processing loop


###loop over the protein input file and , line by line, and populate the
protein global hashes
while(<PROT_FILE>){
        chomp;
        #OK, first check to see if this line is defining an mRNA. If so, map
the mRNA id to the protein ID that it represents
        #in this block $1=Parent_scaffold_ID $2=mRNA_start $3=mRNA_end
$4=strand $5=mRNA_ID $6=Protein_ID
        if($jgi_flag && m/scaffold_([0-9]+)\tprediction\tmRNA\t([0-9]+)\t([0-9]
+)\t.*\t([+-])\t.*\tID=mRNA_([0-9]+);Name=jgi\.p\|$target_header\|([0-9]+);/){
                #print "$1 $2 $3 $4 $5 $6\n" and die();
                if($Protein_map_ID{$5}){
                        die("mRNA map ID $5 is duplicated in the .gff3 file: $!
\n");
                }else{
                        #OK, let's catalog the information for this protein
                        $Protein_map_ID{$5}=$6;
                        $Protein_info{$6}= "$1 $2 $3 $4"
                }
        }

        if($maker_flag && m/NODE_([0-9]+).*\tmaker\tgene\t([0-9]+)\t([0-9]+)
\t.*\t([+-])\t.*\tID=Rves056\.([0-9]+);/){

                $Protein_info{$5}= "$1 $2 $3 $4"

        }



}#end PROTFILE PROCESSING while



###############begin TARGET processing algorithm
@Protein_holder= ();
@Protein_current= ();
@Exon_holder= ();
$active_protein_seq="";
$temp_protein_header="";
$active_seq="";
$temp_header="";
```

```
#declare the file output handles

my $Passprotein_outfile= $target_header . "_reference-proteins.fasta";


my $protein_handle_Passout=Bio::SeqIO->new( -file=>">$Passprotein_outfile", -
format=>"Fasta");

my $seq_obj_out;


while(my ($key,$value) = each %target_list){
        ############With the information from the cluster files we extracted
earlier, let's screen all proteins
        #here key is equal to the protein id we want to extract from the
target, $value is the protein id of the
        #homolog from the reference


        #print "$Current_protein \n" and die();
        #print "$Protein_info{$Current_protein} \n" and die();


        ########################first load the current protein information, wait
to print the protein until we know if it passed
        @Protein_holder=split(" ", $Protein_info{$key});
        #print @Protein_holder;
        #print "\n" and die();
        ###############determine if reverse complement needed
        if($Protein_holder[3] eq "+"){
                $Reverse_flag=0;
                $current_strand="forward";
        }else{
                $Reverse_flag=1;
                $current_strand= "reverse";
        }
        $Current_contig=$Protein_holder[0];
        #print "$Current_contig \n" and die();
        #grab just the protein_sequence for the target protein being examined
        ###############reverse complement if neccessary, and then print to
file
        $temp_protein_header= $target_header. "." . $value . "." . $key ."_in
$target_header.". $Protein_holder[1] . "." . $Protein_holder[2] . "." .
$current_strand;
        $active_protein_seq=substr($Contig_sequence{$Current_contig},
($Protein_holder[1] - 1), ($Protein_holder[2] - $Protein_holder[1]+1));
        if($Reverse_flag){
                $active_protein_seq=reverse_complement
($active_protein_seq);
        }


        ############print the entire protein to the appropriate file: filtered
proteins if it passed, rejected proteins if it failed

        $seq_obj_out= Bio::Seq->new(-seq=>"$active_protein_seq", -
display_id=>"$temp_protein_header", -alphabet=>"dna");
        $Passed_proteins++;
        $protein_handle_Passout->write_seq($seq_obj_out);
```

```
}#end TARGET processing while

print STDERR "Printed $Passed_proteins filtered proteins to
$Passprotein_outfile\n";

close(PROT_FILE);
close(TARGET_FILE);

#################Begin PERL subroutine definitions#################
###################################################################


#subroutine to reverse complement a string of DNA sequence
sub reverse_complement($) {
        my $dna = shift;

        # reverse the DNA sequence
        my $revcomp = reverse($dna);

        # complement the reversed DNA sequence
        $revcomp =~ tr/NACGTacgt/NTGCAtgca/;
        return $revcomp;
}




# Perl trim function to remove whitespace from the start and end of the string
sub trim($)
{
        my $string = shift;
        $string =~ s/^\s+//;
        $string =~ s/\s+$//;
        return $string;
}

# PERL usage function for this program, SNP_density_calc.pl
sub usage
{
        print STDERR << "EOF";

        Name $prg - Extracts proteins (listed by JGI protein ID) for the
target taxon that cluster
                        with particular proteins specified from a list in the
reference taxa. Takes a cluster
                        input file containing a newline delimited list of
orthologous proteins clusters.
                        Only those proteins belonging to single copy protein
coding clusters, with all
                        exons greater than 120bp and a total exon length of
greater than 960 bp will be
                        extracted. One file is printed, containing the
sequence of all matching proteins
                        from the complementary taxa
                        protein header format: TaxonID.proteinID-
inREF.proteinID-inTARGET.protein_start.protein_end.strand

        usage:  $prg (-l input) (-g JGI_input) (-c fasta_input) [-a header]
```

```
        -h              :       print this help message
        -l input        :       Cluster file that contains a newline delimited
list of protein
                                clusters. This file is the ".end" file
generated as output of
                                the FastOrtho program, sets inputfile
        -r input        :       A newline delimited list of protein IDs from
the reference taxon
                                that are of interest. Homologs from target
taxon should be extracted.,
                                sets inputfile.
        -g JGI_input    :       JGI format .gff3 input file with exon
coordinates of target taxon,
                                sets inputfile
        -m Maker_input  :       Maker format .gff3 input file with exon
coordinates of target taxon,
                                sets inputfile
        -c fasta_input  :       Fasta format file defining nucleotide sequence
of assembly contigs
                                for the target taxon, sets inputfile
        -a header       :       The reference taxon abbreviation to use for
searching cluster file
                                and naming output files (Note: this
abbreviation must exactly match
                                the jgi taxon from the second field of jgi
protein headers)
        -b header       :       The target taxon abbreviation to use for
searching cluster file
                                and naming output files (Note: this
abbreviation must exactly match
                                the jgi taxon from the second field of jgi
protein headers)

        ex: $prg -l FastOrthoOutput.end -g JGI_predictions.gff3 -c contigs.fa -
a MyTaxonAbbreviation

EOF
        exit;

}
```

```
#!/usr/bin/perl -w

#Name Exon_extractor_V3.pl - Extracts proteins (listed by JGI protein ID) for
the target taxon from a cluster
#               input file containing a newline delimited list of orthologous
proteins clusters.
#       See usage() function at the end of program for details and usage
instructions.

#use warnings;
use strict;
use Getopt::Std;
use File::Basename;
use Bio::SeqIO;
use Bio::DB::Fasta;


sub reverse_complement($);
sub trim($);
sub usage;


#declare the global variables
#hashes
my %Contig_length;
my %Protein_map_ID;
my %Protein_info;
my %Contig_sequence;
#my %Protein_length;
#my %Protein_SNPcount;
my %Protein_exons;

#arrays
my @Protein_holder;
my @Protein_current;
my @Exon_holder;
my @Exon_current;
my @Exons_toPrint;
my @target_list;

#strings
my $current_strand;
my $Temp_contig;
my $Current_contig;
my $Last_contig;
my $Current_protein;
my $next_line;
my $Fail_string;
my $active_seq;
my $temp_header;
my $active_protein_seq;
my $temp_protein_header;


#integers
my $Reverse_flag=0;
my $Fail_flag_120=0;
my $Fail_flag_960=0;
my $Failed_proteins=0;
my $Passed_proteins=0;
```

```
my $Exons_totalLength=0;
my $Passed_exons_totallength=0;

#declare the options string
my $opt_string='hl:g:c:a:';
my $prg = basename $0;
my %opt;
my $output_header;
my $list_file;
my $gff_file;
my $contig_file;

####process the options and set global variables for options
getopts("$opt_string", \%opt) or usage();
$opt{h} and usage();
usage() unless ($opt{l} && $opt{g} && $opt{c} && $opt{a});
$output_header = $opt{a};


######inform the user of program settings
print STDERR "Target taxon set to: $output_header\n";
$opt{l} and print STDERR "Extracting single copy protein coding genes for
target taxon from cluster list specified in input file: $opt{l}\n";
$opt{g} and print STDERR "Using exon coordinates from input file: $opt{g}\n";
$opt{c} and print STDERR "and nucleotide sequences from input file: $opt{c}\n";

#set the input files
$list_file = $opt{l};
$gff_file = $opt{g};
$contig_file = $opt{c};

#declare the sequence handle and container stuff
my $sequence_handle=Bio::SeqIO->new( -file=>$contig_file, -format=>"Fasta");
my $db = Bio::DB::Fasta->new($contig_file);
my @active_header;


#check the indices here, this is not yet determined to work for JGI headers
#populate the Contig sequence hash with all of the contig sequences
while(my $seq_obj=$sequence_handle->next_seq){
        $temp_header=$db->header($seq_obj->id);
        chomp($temp_header);
        @active_header=split(/_/, $temp_header);
        $active_seq= $seq_obj->seq;
        #insert the sequence into a new hash
        $Contig_sequence{$active_header[1]}=$active_seq;
        $Contig_length{$active_header[1]}=length($active_seq);
}#end while


open(PROT_FILE, $gff_file) or die("could not open protein coordinates file
$gff_file: $!\n");
#open(CONTIG_FILE, "$ARGV[@ARGV-1]") or die("could not open contig file $ARGV
[@ARGV-1]: $!\n");
open(TARGET_FILE, $list_file) or die("could not open protein cluster
definition file $list_file: $!\n");


###############first loop over the cluster definition file and get all those
protein id from the taxon defined by -a
```

```
############## which belong to clusters that are single copy protein coding
genes
while(<TARGET_FILE>){
        chomp;
        #scan over each line and grab those lines that contain clusters with
single copy protein coding genes (4 genes, 4 taxa)
        #$1 =cluster name $2= protein_id from taxon specified by -a option
        if(m/ORTHOMCL([0-9]+) \(4 genes,4 taxa\):.*jgi\|$output_header\|([0-9]
+)\|.*/){
                push @target_list, $2;
        }

}#end cluster processing loop


###loop over the protein input file and , line by line, and populate the
protein global hashes
while(<PROT_FILE>){
        chomp;
        #OK, first check to see if this line is defining an mRNA. If so, map
the mRNA id to the protein ID that it represents
        #in this block $1=Parent_scaffold_ID $2=mRNA_start $3=mRNA_end
$4=strand $5=mRNA_ID $6=Protein_ID
        if(m/scaffold_([0-9]+)\tprediction\tmRNA\t([0-9]+)\t([0-9]+)\t.*\t([+-
])\t.*\tID=mRNA_([0-9]+);Name=jgi\.p\|Rhivi1\|([0-9]+);/){
                #print "$1 $2 $3 $4 $5 $6\n" and die();
                if($Protein_map_ID{$5}){
                        die("mRNA map ID $5 is duplicated in the .gff3 file: $!
\n");
                }else{
                        #OK, let's catalog the information for this protein
                        $Protein_map_ID{$5}=$6;
                        $Protein_info{$6}= "$1 $2 $3 $4"
                }
        }

        ## if we are in a protein region, get the exon regions and the
sequence and enter them in to the protein hashes
        ## this block assumes that ($Temp_protein ne "") because CDS regions
can only be found after a (#start ) pattern
        ## as recognized in the next code block
        ## There is another emergent factor here, due to the match condition
below, some CDS regions are being incorrectly applied
        ## to the wrong protein (or those that do not exist...). Store the
exon in direct association with the gene name on the same line
        ## to ensure that the name of the protein in the CDS is the protein
that is recieving the information in the array.
        ## The assumption above was not sufficient to prevent error.
        ## as I've seen in the past, the pattern matching approach below will
incorrectly apply some information to the wrong protein
        ## in this block $1 = CDS_start, $2=CDS_end, $3=strand, $4 = exon_ID
(#in_parent_mRNA), $5 = Parent_mRNA_ID
        if(m/prediction\texon\t([0-9]+)\t([0-9]+)\t.*\t([+-])\t.*\tID=exon_
[0-9]+_([0-9]+);Parent=mRNA_([0-9]+)/){
                #print STDERR "$5\n";
                #print "$1 $2 $3 $4 $5\n" and die();
                if($Protein_exons{$Protein_map_ID{$5}}){
                        #protein has exon info already, add this one to the
list
                        $Protein_exons{$Protein_map_ID{$5}} .= "\t$1 $2 $3 $4";
```

```
                }else{
                        #protein has no exon info, create this entry as the
first
                        $Protein_exons{$Protein_map_ID{$5}} = "$1 $2 $3 $4";
                }
        }

}#end PROTFILE PROCESSING while



###############begin TARGET processing algorithm
@Protein_holder= ();
@Protein_current= ();
@Exon_holder= ();
$active_protein_seq="";
$temp_protein_header="";
$active_seq="";
$temp_header="";
#declare the file output handles
my $Failprotein_outfile= $output_header . "_Failed_proteins.fasta";
my $Failexon_outfile= $output_header . "_Failed_exons.fasta";
my $Passprotein_outfile= $output_header . "_Filtered_proteins.fasta";
my $Passexon_outfile= $output_header . "_Filtered_exons.fasta";
my $protein_handle_Failout=Bio::SeqIO->new( -file=>">$Failprotein_outfile", -
format=>"Fasta");
my $protein_handle_Passout=Bio::SeqIO->new( -file=>">$Passprotein_outfile", -
format=>"Fasta");
my $exon_handle_Failout=Bio::SeqIO->new( -file=>">$Failexon_outfile", -
format=>"Fasta");
my $exon_handle_Passout=Bio::SeqIO->new( -file=>">$Passexon_outfile", -
format=>"Fasta");
my $seq_obj_out;


while(@target_list){
        #############With the information from the cluster files we extracted
earlier, let's screen all proteins
        #############We need to make sure they meet criteria for hyb-seq. print
to 4 files: filtered proteins, filtered exons
        #############rejected proteins, rejected exons
        $Current_protein=trim(shift(@target_list));
        $Exons_totalLength=0;
        $Fail_string="";
        $Fail_flag_120=0;
        $Fail_flag_960=0;
        #print "$Current_protein \n" and die();
        #print "$Protein_info{$Current_protein} \n" and die();


        ########################first load the current protein information, wait
to print the protein until we know if it passed
        @Protein_holder=split(" ", $Protein_info{$Current_protein});
        #print @Protein_holder;
        #print "\n" and die();
        ###############determine if reverse complement needed
        if($Protein_holder[3] eq "+"){
                $Reverse_flag=0;
                $current_strand="forward";
        }else{
```

```
                $Reverse_flag=1;
                $current_strand= "reverse";
        }
        $Current_contig=$Protein_holder[0];
        #print "$Current_contig \n" and die();
        #grab just the protein_sequence for the target protein being examined
        ##############reverse complement if neccessary, and then print to
file
        $temp_protein_header="scaffold_". $Current_contig . "." .
$Current_protein ."." $Protein_holder[1] . "." . $Protein_holder[2] . "." .
$current_strand;
        $active_protein_seq=substr($Contig_sequence{$Current_contig},
($Protein_holder[1] - 1), ($Protein_holder[2] - $Protein_holder[1]+1));

        #note, this used to be $active_seq... which was incorrect. Changed on
2-6-15 to be $active_protein_seq which is the correct item
        #to reverse complement
        if($Reverse_flag){
                $active_protein_seq=reverse_complement
($active_protein_seq);
        }


        ############second deal with the exons from this protein.
        ############Grab each exon, and determine if this protein is both long
enough (960bp total exon length)
        ############and all of its exons are over 120bp. Reverse complement if
neccessary, and print the exons to file the appropriate file
        @Exons_toPrint= ();
        @Exon_holder=split("\t", $Protein_exons{$Current_protein});
        while(@Exon_holder){
                @Exon_current=split(" ",trim(shift(@Exon_holder)));
                $temp_header="scaffold_". $Current_contig . "."
$Current_protein. "." . $Exon_current[3]. "." $Exon_current[0] . "." .
$Exon_current[1] . "." .$current_strand;
                $active_seq=substr($Contig_sequence{$Current_contig},
($Exon_current[0] - 1), ($Exon_current[1] - $Exon_current[0]+1));
                #is the protein on the negative strand? if so, reverse
complement the sequence
                if($Reverse_flag){
                        $active_seq=reverse_complement($active_seq);
                }
                ############is this exon greater than or equal to 120bp? If not
set the 120bp fail flag
                #print(length($active_seq) . "\n");
                if(int(length($active_seq)) < 120){
                        $Fail_flag_120 = 1;
                        $Fail_string .=".lt120";
                }
                #print("120bp condition flag = $Fail_flag_120 \n");
                $Exons_totalLength += length($active_seq);
                push(@Exons_toPrint, ("$temp_header". "\t" . "$active_seq"));
        }#end Exon Processing Loop

        ###############is the total length of exons in this protein less than
960? if so, fail this protein
        if($Exons_totalLength < 960){$Fail_flag_960=1; $Fail_string .=
".lt960";}

        ##############OK, let's print the exons to the appropriate file, pass
```

```
or fail file
        while(@Exons_toPrint){
                @Exon_current=split("\t",trim(shift(@Exons_toPrint)));
                $active_seq=$Exon_current[1];
                $temp_header= "$Exon_current[0]" . "$Fail_string";
                $seq_obj_out= Bio::Seq->new(-seq=>"$active_seq", -
display_id=>"$temp_header", -alphabet=>"dna");
                #print("960bp condition flag = $Fail_flag_960 \n");
                if($Fail_flag_120 || $Fail_flag_960){
                        $exon_handle_Failout->write_seq($seq_obj_out);
                }else{
                        $Passed_exons_totallength += length($active_seq);
                        $exon_handle_Passout->write_seq($seq_obj_out);
                }
        }


        ############print the entire protein to the appropriate file: filtered
proteins if it passed, rejected proteins if it failed
        $temp_protein_header .= $Fail_string;
        $seq_obj_out= Bio::Seq->new(-seq=>"$active_protein_seq", -
display_id=>"$temp_protein_header", -alphabet=>"dna");
        if($Fail_flag_120 || $Fail_flag_960){
                $Failed_proteins++;
                $protein_handle_Failout->write_seq($seq_obj_out);
        }else{
                $Passed_proteins++;
                $protein_handle_Passout->write_seq($seq_obj_out);
        }

}#end TARGET processing while

print STDERR "Printed $Failed_proteins failed proteins to $Failprotein_outfile
\nPrinted $Passed_proteins filtered proteins to $Passprotein_outfile\n";
print STDERR "Total length of all filtered exons = $Passed_exons_totallength
base pairs\n";

close(PROT_FILE);
close(TARGET_FILE);

##################Begin PERL subroutine definitions##################
#####################################################################


#subroutine to reverse complement a string of DNA sequence
sub reverse_complement($) {
        my $dna = shift;

        # reverse the DNA sequence
        my $revcomp = reverse($dna);

        # complement the reversed DNA sequence
        $revcomp =~ tr/NACGTacgt/NTGCAtgca/;
        return $revcomp;
}
```

```
# Perl trim function to remove whitespace from the start and end of the string
sub trim($)
{
        my $string = shift;
        $string =~ s/^\s+//;
        $string =~ s/\s+$//;
        return $string;
}

# PERL usage function for this program, SNP_density_calc.pl
sub usage
{
        print STDERR << "EOF";

        Name $prg - Extracts proteins (listed by JGI protein ID) for the
target taxon from a cluster
                        input file containing a newline delimited list of
orthologous proteins clusters.
                        Only those proteins belonging to single copy protein
coding clusters, with all
                        exons greater than 120bp and a total exon length of
greater than 960 bp will be
                        extracted. Four files are printed: 1) complete
nucleotide sequence of proteins
                        that passed exon length filters, 2)exons from proteins
that passed exon length
                        filters, 3) complete nucleotide sequence of proteins
that failed exon length
                        filters, 4)exons from proteins that failed exon length
filters.
                        protein header format:
scaffold_id.protein_id.protein_start.protein_end.strand
                                        .reason_for_fail
                        exon header format:
scaffold_id.protein_id.exon_id.protein_start.protein_end
                                                .strand.reason_for_fail
        usage:  $prg (-l input) (-g JGI_input) (-c fasta_input) [-a header]

        -h              :       print this help message
        -l input        :       Cluster file that contains a newline delimited
list of protein
                                clusters. This file is the ".end" file
generated as output of
                                the FastOrtho program, sets inputfile
        -g JGI_input    :       JGI format .gff3 input file with exon
coordinates of target taxon,
                                sets inputfile
        -c fasta_input  :       Fasta format file defining nucleotide sequence
of assembly contigs
                                for the targe taxon, sets inputfile
        -a header       :       The target taxon abbreviation to use for
searching cluster file
                                and naming output files (Note: this
abbreviation must exactly match
                                the jgi taxon from the second field of jgi
protein headers)

        ex: $prg -l FastOrthoOutput.end -g JGI_predictions.gff3 -c contigs.fa -
a MyTaxonAbbreviation
```

```
EOF
        exit;

}
```

```
#!/usr/bin/perl -w

#Name insert_intoAlignment.pl - Inserts proteins (listed by JGI protein ID)
for the target
#               taxon into a alignments for those proteins.
#See usage() function at the end of program for details and usage instructions.

#use warnings;
use strict;
use Getopt::Std;
use File::Basename;
use Bio::SeqIO;
use Bio::DB::Fasta;


sub reverse_complement($);
sub trim($);
sub usage;


#declare the global variables
#hashes

my %Insert_sequence;

#arrays
my @Protein_holder;
my @Protein_current;
my @Exon_holder;
my @Exon_current;
my @Exons_toPrint;
my @target_list;

#strings
my $current_strand;
my $Temp_contig;
my $Current_contig;
my $Last_contig;
my $Current_protein;
my $next_line;
my $Fail_string;
my $active_seq;
my $temp_header;
my $active_protein_seq;
my $temp_protein_header;
my $Seq_directory;
my $Reference_file;

#integers
my $jgi_flag=0;
my $maker_flag=0;
my $Reverse_flag=0;
my $Fail_flag_120=0;
my $Fail_flag_960=0;
my $Failed_proteins=0;
my $Passed_proteins=0;
my $Exons_totalLength=0;
my $Passed_exons_totallength=0;
my $Print_count=0;
```

```
#declare the options string
my $opt_string='hd:r:';
my $prg = basename $0;
my %opt;
my $reference_header;
my $target_header;
my $list_file;
my $reference_file;
my $gff_file;
my $contig_file;

####process the options and set global variables for options
getopts("$opt_string", \%opt) or usage();
$opt{h} and usage();
usage() unless ($opt{d} && $opt{r});
print "Could not open file: $opt{r}\n" and die(usage()) unless (-e $opt{r});
print "Could not read directory: $opt{d}\n" and die(usage()) unless (-e $opt{d}
);

######inform the user of program settings
$opt{d} and print STDERR "Inserting sequence data for target taxon into files
stored in the directory: $opt{d}\n";
$opt{c} and print STDERR "using sequences from input file: $opt{r}\n";


#set the input files
$Seq_directory = $opt{d};
$Reference_file = $opt{r};
my @Reference_file_array = split(/_/,$Reference_file);
my $Reference_ID = $Reference_file_array[0];

# open up BLAST and ASSEMBLY directories and reference file
opendir(SEQDIR, $Seq_directory) or die "can't open directory: $Seq_directory $!
\n";

# if input directory does not end with '/', add it.
#This is for the purposes of printing files later on, no need for the opening
of the directory.
if ($Seq_directory !~ /\/$/){
        $Seq_directory .= "/";
}

#declare the sequence handle and container stuff
my $sequence_handle=Bio::SeqIO->new( -file=>$Reference_file, -format=>"Fasta");
my $db = Bio::DB::Fasta->new($Reference_file);
my @active_header;


#populate the reference sequence hash with all of the reference sequences
while(my $seq_obj=$sequence_handle->next_seq){
        $temp_header=$db->header($seq_obj->id);
        chomp($temp_header);
        @active_header=split(/\./, $temp_header);
        $active_seq= $seq_obj->seq;
        #insert the sequence into a new hash
        $Insert_sequence{$active_header[1]}=$active_seq;
}#end while

###loop through all of the files in the input directory and input sequences
where appropriate
```

```
foreach my $f (readdir (SEQDIR)){
        next if ($f =~ /^\./); # skip if this file is default '.' or '..'
directory
        next if ($f !~ /\.fasta$/); # skip if this file does not end with the
right file suffix
        my ($id,$jnk) = split(/_/,$f);
        chomp($id);
        my $in_file  = $Seq_directory.$f;
        #declare the sequence handle and container stuff
        my $infile_handle=Bio::SeqIO->new( -file=> ">>$in_file", -
format=>"Fasta");

        #check if this file header matches any of the protein ids in the
$Ref_sequence hash
        #if so, print the sequence to this file
        if(exists($Insert_sequence{$id})){

                my $temp_protein_header=$Reference_ID . "." . $id;
                my $active_protein_seq = "$Insert_sequence{$id}";
                my $seq_obj_out= Bio::Seq->new(-seq=>"$active_protein_seq", -
display_id=>"$temp_protein_header", -alphabet=>"dna");

                $infile_handle->write_seq($seq_obj_out);
                $Print_count++;
                print STDERR "Printed protein with id: $id for sample ID
$Reference_ID into file $f\n";
        }
}
print STDERR "Printed a total of $Print_count sequences from sample ID
$Reference_ID into the appropriate files\n in directory $Seq_directory\n";

##################Begin PERL subroutine definitions##################
#######################################################################


#subroutine to reverse complement a string of DNA sequence
sub reverse_complement($) {
        my $dna = shift;

        # reverse the DNA sequence
        my $revcomp = reverse($dna);

        # complement the reversed DNA sequence
        $revcomp =~ tr/NACGTacgt/NTGCAtgca/;
        return $revcomp;
}




# Perl trim function to remove whitespace from the start and end of the string
sub trim($)
{
        my $string = shift;
        $string =~ s/^\s+//;
        $string =~ s/\s+$//;
        return $string;
}
```

```
# PERL usage function for this program
sub usage
{
        print STDERR << "EOF";

        Name $prg - Inserts proteins (listed by JGI protein ID) for the target
taxon into a alignments
                        for those proteins. Takes a directory as input which
contains alignments for
                        the proteins of interest. Prints a file with newline
delimited information
                        showing the protein files each protein id was printed
to.
                        protein header format expected: TaxonID.proteinID-
inREF.proteinID-inTARGET.protein_start.protein_end.strand
                        File format expected: ProteinID_SomeInformation.fasta

                        NOTE: file names of fasta files must end in .fasta

        usage:  $prg -d directory -r sequences_to_insert.fasta

        -h              :       print this help message
        -d directory    :       Path to a directory which contains nucleotide/
protein alignments that
                                with have sequence data inserted into them,
sets input directory
        -r seq.fasta    :       A fasta format sequence file with multiple
sequences to be inserted
                                into the files stored in directory, sets input
file.

        ex: $prg -d Some_directory/ -r my_sequences.fasta

EOF
        exit;

}
```

```
#!/usr/bin/perl -w

# This script will search a multiple sequence nucleotide fasta file for open
reading frames that contain CaaX motifs typical
# of small phermone precursor genes.  Both forward and reverse strands are
searched.

use strict;
use Bio::SeqIO;
use Bio::DB::Fasta;

sub trim($);

my $usage="usage: pheromone_seeker.pl inputfile";
#require file input from the command line
die("$usage\n") unless(@ARGV > 0);
my $input=$ARGV[@ARGV-1];


#declare the pheromone motif we are looking for...
#pattern is structured as such: M|L protein-body C (V|L|I|M|){2} X
my $pheromone_pattern="ATG ((?!TAA|TAG|TGA)[TCAG]{3})+ TG[TC] (GT[TCAG] | CT
[TACG] | AT[TCAG]){2} [TCAG]{3} (TAA|TAG|TGA)";

#Beginning of pattern that will match novel L and M start codons (replaces
ATG): ((ATG)|(CT[TCAG])|(TT[AG]))
#old beginning of match pattern(matches stop codons indescriminantly in middle
of query): ATG ([TCAG]{3})+
#"C[A-Z]{6,14}C[A-Z]{4,5}C{3}H";

#declare the sequence handle and container stuff
my $active_seq;
my $seq_obj_revcom;
my $active_seq_revcom;
my $sequence_handle=Bio::SeqIO->new( -file=>$input, -format=>"Fasta");
my $db = Bio::DB::Fasta->new($input);
my $temp_header;
my @active_header;


#ok, now loop over all of the sequences in the file and output the appropriate
stuff
while(my $seq_obj=$sequence_handle->next_seq){
        $temp_header=$db->header($seq_obj->id);
        chomp($temp_header);
        @active_header=split(/_/, $temp_header);
        $active_seq= $seq_obj->seq;
        $seq_obj_revcom=$seq_obj->revcom;
        $active_seq_revcom=$seq_obj_revcom->seq;
        while($active_seq =~ /$pheromone_pattern/gix){
                #$match_pos=pos();
                #prints: fasta id \t match \t start_of_match \t end_of_match \n
                #my $translator=Bio::PrimarySeq->new(-seq=>$& , -
alphabet=>'dna');
                print $seq_obj->id . "\t" . $& . "\t" . (length($`) + 1) .
"\t" . (length($`) + 1 + length($&)) . "\t" . "+" .  "\n";
                #print $translator->translate(-frame=>1) . "\n";
        }
        while($active_seq_revcom =~ /$pheromone_pattern/gix){
                #$match_pos=pos();
                print $seq_obj->id . "\t" . $& . "\t" . (length($`) + 1) .
```

```
"\t" . (length($`) + 1 + length($&)) . "\t" . "-" . "\n";
        }
}#end while


# Perl trim function to remove whitespace from the start and end of the string
sub trim($)
{
        my $string = shift;
        $string =~ s/^\s+//;
        $string =~ s/\s+$//;
        return $string;
}
```

```
#!/usr/bin/perl -w

#Name protein_coordinates.pl - Selects and prints to file all protein
coordinates from a .gff3 file.
#               Can be restricted to printing only those proteins from a
particular contig
#               or those proteins supplied in a newline delimited file
#See usage() function at the end of program for details and usage instructions.


use strict;
use Getopt::Std;
use File::Basename;

sub trim;
sub usage;

my $prg = basename $0;
my %opt;
my $opt_string='hac:l:mj:';
my $taxon_abbr="";
my $augustus_flag=0;
my $maker_flag=0;
my $jgi_flag=0;
my $contig_flag=0;
my $list_flag=0;
my $gff_infile;
my $list_infile;
my %protein_list;

####process the options and set global variables for options
getopts("$opt_string", \%opt) or usage();
$opt{h} and usage();
usage() unless ($opt{a} || $opt{m} || $opt{j});
$augustus_flag = $opt{a} ? 1 : 0;
$maker_flag = $opt{m} ? 1 : 0;
if($opt{j}){
        $jgi_flag=1;
        $taxon_abbr= $opt{j};
}else{
        $jgi_flag=0;
}
$jgi_flag = $opt{j} ? 1 : 0;
$contig_flag = $opt{c} ? 1 :0;
$list_flag = $opt{l} ? 1 : 0;
if(($augustus_flag + $maker_flag + $jgi_flag) != 1 || (($contig_flag ||
$list_flag) && ($augustus_flag + $maker_flag + $jgi_flag) != 1)){
        print "exactly 1 input format option must be set\nIf the contig or
list option is set, you must still specify an input format\n";
        usage();
}
if($contig_flag && $list_flag){
        print "Contig and list output options are mutually exclusive and
cannot both be specified\n";
        usage();
}

#require file input and contig from the command line
usage() unless(@ARGV == 1);
my $contig_num;
$gff_infile=$ARGV[@ARGV-1];
```

```
$list_infile= $opt{l} ? $opt{l} : "";
$contig_num= $contig_flag ? $opt{c} : "All-Proteins";
if(!$contig_flag){$contig_num= $list_flag ? "List-Proteins" : "All-Proteins";}
my @basename=split(/\./, basename($gff_infile));
my $outfile = $basename[0] . "_coordinates_" . $contig_num . ".txt";

#make an input file handle to read coordinate/contig information

if($list_flag){
        open(LIST_FILE, $list_infile) or die("could not open protein list file
$list_infile: $!\n");
}

open(PROT_FILE, $gff_infile) or die("could not open protein coordinates file
$gff_infile: $!\n");

#make an output handle
open(OUT_FILE, ">$outfile")or die("could not open output coordinates file
$outfile: $!\n");


#in list mode? get the list and load it into a hash.
my $temp_id;
if($list_flag){
        while(<LIST_FILE>){
                chomp;
                $temp_id=trim($_);
                #print "the protein id is $temp_id\n";
                $protein_list{$temp_id}=1;
        }
}


# now loop over all of the lines in the file and output the appropriate stuff

while(<PROT_FILE>){
        #loop over the protein gff3 file line by line, get those protein
coordinates that are on the appropriate contig
        if($augustus_flag && m/NODE_([0-9]+)_length_[0-9]+_cov_[0-9+]\.[0-9]+
\tAUGUSTUS\tgene\t([0-9]+)\t([0-9]+)\t.*\t([+-])\t.*\tID=.+\.(g[0-9]+)\./){
                #print STDERR "$5\n";
                if($contig_flag && ($contig_num == $1)){
                        #protein is on the appropriate contig, print it
                        print OUT_FILE "NODE_$1\t$5\t$2\t$3\t$4\n";
                }
                if($list_flag && exists($protein_list{$5})){
                        print OUT_FILE "NODE_$1\t$5\t$2\t$3\t$4\n";
                }
                #if not in contig mode or list mode, just print it all
                if(!$contig_flag && !$list_flag){print OUT_FILE "NODE_$1\t$5\t
$2\t$3\t$4\n";}
        }

        if($jgi_flag && m/[a-zA-Z]+_([0-9]+)\tprediction\tmRNA\t([0-9]+)\t
([0-9]+)\t.*\t([+-])\t.*\tID=mRNA_([0-9]+);Name=jgi\.p\|$taxon_abbr\|([0-9]+);/
){

                if($contig_flag && ($contig_num == $1)){
                        #protein is on the appropriate contig, print it
                        print OUT_FILE "scaffold_$1\t$6\t$2\t$3\t$4\n";
                }
```

```
                if($list_flag && exists($protein_list{$6})){
                        print OUT_FILE "scaffold_$1\t$6\t$2\t$3\t$4\n";
                }
                #if not in contig mode or list mode, just print it all
                if(!$contig_flag && !$list_flag){print OUT_FILE "scaffold_$1\t
$6\t$2\t$3\t$4\n";}
        }

        ##add a special case here to deal with Coprinopsis cinerea
        if($jgi_flag && ($taxon_abbr eq "Copci1")){


        }

        if($maker_flag && m/NODE_([0-9]+).*\tmaker\tgene\t([0-9]+)\t([0-9]+)
\t.*\t([+-])\t.*\tID=(Rves056\.[0-9]+);/){
                #print"Entered maker flag section correctly. Contig ID = $1\n";
                #print "the contig flag is set to: $contig_flag, the contig
number specified was: $contig_num and the current contig is: $1\n";
                if($contig_flag && $contig_num == $1){
                        #print "Contig flag condition was met\n";
                        #protein is on the appropriate contig, print it
                        print OUT_FILE "NODE_$1\t$5\t$2\t$3\t$4\n";
                }
                if($list_flag && exists($protein_list{$5})){
                        #print "the protein id is $5
\n";
                        print OUT_FILE "NODE_$1\t$5\t$2\t$3\t$4\n";
                }
                #if not in contig mode or list mode, just print it all
                if(!$contig_flag && !$list_flag){print OUT_FILE "NODE_$1\t$5\t
$2\t$3\t$4\n";}
        }
}

close(LIST_FILE);
close(OUT_FILE);
close(PROT_FILE);

# Perl trim function to remove whitespace from the start and end of the string
sub trim($)
{
        my $string = shift;
        $string =~ s/^\s+//;
        $string =~ s/\s+$//;
        return $string;
}

sub usage
{
        print STDERR << "EOF";

        Name $prg -     Selects and prints to file all protein coordinates
from a .gff3 file.
                        Output includes scaffold/contig, protein_id, start,
end, and strand
                        in tab delimited format.
                        If the -c option is set it will print only those
proteins on a particular
                        assembly contigs/scaffolds.
                        If the -l option is set it will print only those
```

```
proteins from a newline
                        delimited list of protein ids
                        Accepts gff3 format from jgi, maker, or augustus.
        usage:  $prg (-a | -m | -j ) [-c contig_num]  input_proteins.gff3

        -h              :       print this help message
        -a              :       Specify augustus input format
        -m              :       Specify maker input format
        -j  taxon_abbr  :       Specify jgi input format and the taxon
abbreviation of the target taxon
        -c  contig_num  :       Specify that output should be for only a
specific contig.
        -l  list.txt    :       Specify that output should be only for a
specific list of proteins
                                as specified by a newline delimited list of
protein ids in the file
                                list.txt. Sets inputfile

        ex: $prg -m -c 35 input_proteins.gff3
                will print all protein coordinates from contig/scaffold 35
contained within the file input_proteins.gff3

EOF
        exit;

}
```

```
#!/usr/bin/perl -w

#Name protein_select_by_contig.pl - Selects and prints to file all proteins
#               on a particular assembly contig.
#               Can also be restricted to print only those contigs from a
newline
#               delimited list
#See usage() function at the end of program for details and usage instructions.

use strict;
use Getopt::Std;
use Bio::SeqIO;
use Bio::DB::Fasta;
use File::Basename;

sub usage;

my $prg = basename $0;
my %opt;
my $opt_string='hal:m:j:';
my $augustus_flag=0;
my $maker_flag=0;
my $jgi_flag=0;
my $list_flag=0;
my $gff_infile;
my $current_protein;
my %Protein_sequence;
my $temp_header;
my @active_header;
my @active_protline;
my @temp_seq_obj;

####process the options and set global variables for options
getopts("$opt_string", \%opt) or usage();
$opt{h} and usage();
usage() unless ($opt{a} || $opt{m} || $opt{j} || $opt{l});
$augustus_flag = $opt{a} ? 1 : 0;
$maker_flag = $opt{m} ? 1 : 0;
$jgi_flag = $opt{j} ? 1 : 0;
$list_flag = $opt{l} ? 1 :0;
if(($augustus_flag + $maker_flag + $jgi_flag) != 1 || ($list_flag &&
($augustus_flag + $maker_flag + $jgi_flag) != 1)){
        print "exactly 1 input format option must be set\nIf the list option
is set, you must still specify an input format\n";
        usage();
}


if($opt{m}){$gff_infile= $opt{m};}
if($opt{j}){$gff_infile= $opt{j};}
if($opt{l}){$gff_infile= $opt{l};}

#require file input and contig from the command line

if($list_flag){usage() unless(@ARGV == 1);}else{usage() unless(@ARGV == 2);}
my $contig_num;
my $input=$ARGV[@ARGV-1];
$contig_num= $list_flag ? "Selected-Proteins" : $ARGV[@ARGV-2];
my @basename=split(/\./, basename($input));
my $outfile = $basename[0] . "_" . $contig_num . ".fasta";
#open(OUTFILE, $outfile);
```

```
#close(OUTFILE);


#declare the sequence handle and container stuff
my $active_seq;
my $sequence_handle;
my $db;
my $split_char;
my $current_id;

#make a seqio object to read in the sequence information, make an input file
handle to read coordinate/list information
$sequence_handle=Bio::SeqIO->new( -file=>$input, -format=>"Fasta");
$db = Bio::DB::Fasta->new($input);
if($maker_flag || $jgi_flag || $list_flag){
        open(PROT_FILE, $gff_infile) or die("could not open protein
coordinates file $gff_infile: $!\n");
}



#make an output handle
my $output_handle=Bio::SeqIO->new( -file => ">$outfile", -
format=>"Fasta");#need to figure out how to make the file

#populate the protein sequence hash with all of the protein sequences if we
are in maker or jgi mode
if($augustus_flag){$split_char="_";}
if($maker_flag){$split_char=" ";}
if($jgi_flag){$split_char='\|';}

##########if the list file is specified, then scan through it and load the
proteins into a hash
if($list_flag){
        while(<PROT_FILE>){
                chomp;
                $Protein_sequence{$_}=1;
        }
}


while(my $seq_obj=$sequence_handle->next_seq){
        #print($seq_obj->id ."\n") and die();
        $temp_header=$seq_obj->id;
        $active_seq= $seq_obj->seq;
        chomp($temp_header);
        #print($split_char ."\n") and die();
        @active_header=split( /$split_char/ , $temp_header);
        #print($active_header[2] ."\n") and die();
        #if in list format, see if the protein is on the list and print it if
that is the case. Parse for the appropriate format
        if($list_flag){
                #$match_pos=pos();
                if($augustus_flag && exists($Protein_sequence{$active_header
[1]})){
                        $output_handle->write_seq($seq_obj);
                        next;
                }
                if($jgi_flag && exists($Protein_sequence{$active_header[2]})){
                        $output_handle->write_seq($seq_obj);
                        next;
```

```
                }
                if($maker_flag && exists($Protein_sequence{substr
($active_header[0], -8, 5)})){
                        $output_handle->write_seq($seq_obj);
                        next;
                }
         }
        #if in augustus format, just get the contig number from  the protein
header and print if it is on the appropriate contig
        if($augustus_flag && !$list_flag && ($active_header[1] == $contig_num))
{
                #$match_pos=pos();
                $output_handle->write_seq($seq_obj);
                next;
         }

        #insert the sequence into the sequence storage hash
        #parse input for JGI format
        if($jgi_flag){
                #print($active_header[2] ."\n") and die();
                $Protein_sequence{$active_header[2]}=$temp_header . " " .
$active_seq;
                next;
        }
        #parse input for maker input
        if($maker_flag){
                $current_id=substr($active_header[0], -8, 5);
                $Protein_sequence{$current_id}=$temp_header . " " .
$active_seq;
        }
}#end while


####################if augustus or list flag, stop here, we are done
if($augustus_flag || $list_flag){exit(0)};


##not in augustus mode? use a pattern matching routine to determine the
proteins to print instead of direct parsing based upon header
# now loop over all of the sequences in the file and output the appropriate
stuff
my $seq_obj;
@active_header=();
while(<PROT_FILE>){
        #loop over the protein gff3 file line by line, get those proteins that
are on the appropriate contig
        if($jgi_flag && m/scaffold_([0-9]+)\tprediction\tmRNA\t([0-9]+)\t([0-9]
+)\t.*\t([+-])\t.*\tID=mRNA_([0-9]+);Name=jgi\.p\|Rhivi1\|([0-9]+);/){
                if($1 == $contig_num){
                        #print($6 . "\n") and die();
                        @active_header=split(" ", $Protein_sequence{$6});
                        #print($active_header[0] . "\n") and die();
                        $temp_header=$active_header[0];
                        $active_seq= $active_header[1];
                        $seq_obj= Bio::Seq->new(-seq=>$active_seq, -
display_id=>$temp_header, -alphabet=>'protein');
                        $output_handle->write_seq($seq_obj);
                }
        }

        if($maker_flag && m/NODE_([0-9]+).*\tmaker\tgene\t([0-9]+)\t([0-9]+)
```

```
\t.*\t([+-])\t.*\tID=Rves056\.([0-9]+);/){

                if($1 == $contig_num){
                        #print $5 . "\n" and die();
                        @active_header=split(" ", $Protein_sequence{$5});
                        $seq_obj= Bio::Seq->new(-seq=>$active_header[1], -
display_id=>$active_header[0], -alphabet=>'protein');
                        $output_handle->write_seq($seq_obj);
                }
        }

}


sub usage
{
        print STDERR << "EOF";

        Name $prg -     Selects and prints to file all proteins on a
particular assembly contigs. Accepts
                        gff3 format from jgi, maker, or augustus
        usage:  $prg (-a | -m input.gff3 [-l list.txt] | -j input.gff3 [-l
list.txt] ) contig# input.fasta

        -h              :       print this help message
        -a              :       Specify augustuts input format
        -m   input.gff  :       Specify maker input format
        -j   input.gff  :       Specify jgi input format
        -l   input.txt  :       list input file that contains a newline
delimited list of protein
                                IDs to fetch, this option preempts output by
contig, sets inputfile

        ex: $prg -m my_proteins.gff3 35 input_proteins.aa.fasta
                will print all proteins from a contig/scaffold 35 contained
within the file input_proteins.aa.fasta

EOF
        exit;

}
```

```
#!/usr/bin/perl -w

#Name SNP_density_calc - Calculates the density of SNPs at the level of
protein, contig,
#               and determines rates of synonymous and non-synonymous mutations
#See usage() function at the end of program for details and usage instructions.

#use warnings;
use strict;
use Getopt::Std;
use File::Basename;
use Bio::SeqIO;
use Bio::DB::Fasta;
use Switch;

sub SNP_status();
sub protein_on_contig(@);
sub acid_test(@);
sub reverse_complement($);
sub trim($);
sub usage;


#declare the global variables
#hashes
my %Contig_length;
my %Contig_SNPcount;
my %Contig_proteins;
my %Contig_sequence;
my %Protein_length;
my %Protein_SNPcount;
my %Protein_exons;
my %Frame_array;
my %Contig_norms;
my %Protein_norms;
my %Protein_map_ID;

#arrays
my @Protein_holder;
my @Protein_current;
my @Exon_holder;
my @Exon_current;
my @Contig_holder;
my @SNP_holder;
#strings
my $Temp_contig;
my $Current_contig;
my $Last_contig;
my $Temp_protein;
my $next_line;
my $Codon_old;
my $Codon_new;
my $SNP_value;
my $Nuc_match;
my $Nuc_ref;
my $split_char;

#integers
my $Contig_count=0;
my $Protein_count=0;
```

```
my $Contig_w_protein=0;
my $Reverse_flag=0;
my $SNP_count=0;
my $SNP_count_protein=0;
my $SNP_count_exonic=0;
my $SNP_count_nonsyn=0;
my $Total_protein=0;
my $Total_intergene=0;
my $SNP_pos;
my $SNP_posINcodon;
my $Codon_upperbound;
my $Codon_lowerbound;
my $Augustus_flag=0;
my $Maker_flag=0;
my $JGI_flag=0;


#declare the options string
my $opt_string='hs:v:c:f:amj';
my $prg = basename $0;
my %opt;
my $frame_size;
my $cutoff;
my $snp_file;

####process the options and set global variables for options
getopts("$opt_string", \%opt) or usage();

$opt{h} and usage();
($opt{v} && $opt{s}) and usage();
usage() unless ($opt{v} || $opt{s});
usage() unless ($opt{a} || $opt{m} || $opt{j});
$Augustus_flag = $opt{a} ? 1 : 0;
$Maker_flag = $opt{m} ? 1 : 0;
$JGI_flag = $opt{j} ? 1 : 0;
if(($Augustus_flag + $Maker_flag + $JGI_flag) != 1){usage();}


$frame_size = $opt{f} ? $opt{f} : 10000;
$cutoff = $opt{c} ? $opt{c} : 10;

######inform the user of program settings
print STDERR "Frame size set to: $frame_size\n";
print STDERR "Cutoff value set to: $cutoff\n";
$opt{v} and print STDERR "VarScan input format specified, calculating SNP
Density from input file: $opt{v}\n";
$opt{s} and print STDERR "SAMtools input format specified, calculating SNP
Density from input file: $opt{s}\n";

#set the SNP input file
$snp_file = $opt{v} ? $opt{v} : $opt{s};

#require file input from the command line
usage() unless(@ARGV == 2);

open(PROT_FILE, "$ARGV[@ARGV-2]") or die("could not open protein file $ARGV
[@ARGV-2]: $!\n");
#open(CONTIG_FILE, "$ARGV[@ARGV-1]") or die("could not open contig file $ARGV
[@ARGV-1]: $!\n");
open(SNP_FILE, $snp_file) or die("could not open SNP definition file
```

```
$snp_file: $!\n");

#declare the sequence handle and container stuff
my $contig_input=$ARGV[@ARGV-1];
my $active_seq;
my $sequence_handle=Bio::SeqIO->new( -file=>$contig_input, -format=>"Fasta");
my $db = Bio::DB::Fasta->new($contig_input);
my $temp_header;
my @active_header;
my $current_id;

#what input mode are we in? Set the split character to be used in header
parsing appropriately
#if($Augustus_flag || $Maker_flag){$split_char="_";}
#if($Maker_flag){$split_char=" ";}
#if($JGI_flag){$split_char='\|';}

#populate the Contig sequence hash with all of the contig sequences
while(my $seq_obj=$sequence_handle->next_seq){
        $Contig_count++;
        $temp_header=$db->header($seq_obj->id);
        chomp($temp_header);
        @active_header=split(/_/, $temp_header);
        $active_seq= $seq_obj->seq;

        #parse input for Augustus and Maker (These both use VELVET contigs)
format
        #if($Augustus_flag || $Maker_flag){$current_id=$active_header[1];}
        #parse input for JGI format
        #if($JGI_flag){$current_id=$active_header[1]; }
        #parse input for maker input
        #if($Maker_flag){$current_id=substr($active_header[0], -8, 5);}

        ###oops! Augustus, maker, and JGI all use the same header format for
contig sequences...
        #the parsing done in commented lines above was not neccessary
        $current_id=$active_header[1];

        #insert the contig sequence into a new hash
        $Contig_sequence{$current_id}=$active_seq;
        #initialize the SNP count of this contig
        $Contig_SNPcount{$current_id}=0;
        $Contig_length{$current_id}=length($active_seq);
        $Total_intergene+=length($active_seq);
}#end while

###loop over the protein input file and , line by line, and populate the
protein and Contig global hashes
while(<PROT_FILE>){

        chomp;
        #populate the Contig Hashes THIS "IF" CODE BLOCK IS DEPRICATED. THIS
CODE IS NOW IMPLEMENTED
        #IN THE TRAVERSAL OF THE CONTIG FILE. SEE ABOVE CODE BLOCK
        #if($Augustus_flag && m/length = ([0123456789]+), name = NODE_([0-9]
+)_/){
        #       $Contig_count++;
        #       $Contig_SNPcount{$2}=0;
        #       $Contig_length{$2}=$1;
        #       $Total_intergene+=$1;
```

```
        #       next;
        #}


        #OK, first check to see if this line is defining an mRNA. If so, map
the mRNA id to the protein ID that it represents

        #parse for JGI format
        #in this block $1=Parent_scaffold_ID $2=mRNA_start $3=mRNA_end
$4=strand $5=mRNA_ID $6=Protein_ID
        if($JGI_flag && m/scaffold_([0-9]+)\tprediction\tmRNA\t([0-9]+)\t([0-9]
+)\t.*\t([+-])\t.*\tID=mRNA_([0-9]+);Name=jgi\.p\|Rhivi1\|([0-9]+);/){
                #print "$1 $2 $3 $4 $5 $6\n" and die();
                if($Protein_map_ID{$5}){
                        die("mRNA map ID $5 is duplicated in the .gff3 file: $!
\n");
                }else{
                        #OK, let's catalog the information for this protein
                        $Protein_map_ID{$5}=$6;

                        #Increment protein count
                        $Protein_count++;
                        #initialize the hash bin for this protein in %
Protein_SNPcount
                        $Protein_SNPcount{$6}=0;
                        #set the hash bin for this protein in %Protein_length
                        $Protein_length{$6}=($3 - $2);
                        #adjust the sum of total length of protein coding
sequence
                        $Total_protein+= ($3 - $2);
                        #set the protein string in the %Contig_proteins hash
to contain the appropriate information
                        if($Contig_proteins{$1}){#if there is an entry for the
current contig
                                $Contig_proteins{$1} .=  "\t$6 $2 $3";
                        }else{#if there is not an entry for the current contig
                                $Contig_w_protein++;
                                $Contig_proteins{$1} =  "$6 $2 $3";
                        }
                }
                next;
        }#end $JGI_flag block


        #parse for Maker format
        #in this block $1=Parent_Contig_ID $2=mRNA_start $3=mRNA_end $4=strand
$5=Protein_ID
        if($Maker_flag && m/NODE_([0-9]+).*\tmaker\tgene\t([0-9]+)\t([0-9]+)
\t.*\t([+-])\t.*\tID=Rves056\.([0-9]+);/){
                #print "$1 $2 $3 $4 $5\n" and die();

                #OK, let's catalog the information for this protein
                #Increment protein count
                $Protein_count++;
                #initialize the hash bin for this protein in %Protein_SNPcount
                $Protein_SNPcount{$5}=0;
                #set the hash bin for this protein in %Protein_length
                $Protein_length{$5}=($3 - $2);
                #adjust the sum of total length of protein coding sequence
                $Total_protein+= ($3 - $2);
```

```
                #set the protein string in the %Contig_proteins hash to
contain the appropriate information
                if($Contig_proteins{$1}){#if there is an entry for the current
contig
                        $Contig_proteins{$1} .=  "\t$5 $2 $3";
                }else{#if there is not an entry for the current contig
                        $Contig_w_protein++;
                        $Contig_proteins{$1} =  "$5 $2 $3";
                }
                next;
        }#end $Maker_flag_output



        ## if we are in a protein region, get the CDS regions and the sequence
and enter them in to the protein hashes
        ## this block assumes that ($Temp_protein ne "") because CDS regions
can only be found after a (#start ) pattern
        ## as recognized in the next code block
        ## There is another emergent factor here, due to the match condition
below, some CDS regions are being incorrectly applied
        ## to the wrong protein (or those that do not exist...). Store the
exon in direct association with the gene name on the same line
        ## to ensure that the name of the protein in the CDS is the protein
that is recieving the information in the array.
        ## The assumption above was not sufficient to prevent error.
        ## as I've seen in the past, the pattern matching approach below will
incorrectly apply some information to the wrong protein

        ##parse for augustus format
        ## in this block $1 = CDS_start, $2=CDS_end, $3=strand, $4 = phase, $5
= protein ID
        if($Augustus_flag && m/AUGUSTUS\tCDS\t([0-9]+)\t([0-9]+)\t.*\t([+-])\t
([012])\tID=.+\.(g[0-9]+)\./){
                #print STDERR "$5\n";
                if($Protein_exons{$5}){
                        #protein has exon info already, add this one to the
list
                        $Protein_exons{$5} .= "\t$1 $2 $3 $4";
                }else{
                        #protein has no exon info, create this entry as the
first
                        $Protein_exons{$5} = "$1 $2 $3 $4";
                }
        }
        ##parse for JGI format
        ## in this block $1 = CDS_start, $2=CDS_end, $3=strand, $4 = phase,
$5= CDS_ID(#in_parent_exon), $6 = Parent_mRNA_ID
        if($JGI_flag && m/prediction\tCDS\t([0-9]+)\t([0-9]+)\t.*\t([+-])\t
([012])\tID=CDS_([0-9]+);Parent=mRNA_([0-9]+)/){
                #print STDERR "$5\n";
                #print "$1 $2 $3 $4 $5 $6\n" and die();
                if($Protein_exons{$Protein_map_ID{$6}}){
                        #protein has exon info already, add this one to the
list
                        $Protein_exons{$Protein_map_ID{$6}} .= "\t$1 $2 $3 $4";
                }else{
                        #protein has no exon info, create this entry as the
first
                        $Protein_exons{$Protein_map_ID{$6}} = "$1 $2 $3 $4";
```

```
                }
        }
        ##parse for Maker format
        ## in this block $1 = CDS_start, $2=CDS_end, $3=strand, $4 = phase,
$5= Protein_ID
        if($Maker_flag && m/maker\tCDS\t([0-9]+)\t([0-9]+)\t.*\t([+-])\t([012])
\tID=Rves056\.([0-9]+)-RA/){
                #print STDERR "$5\n" and die();
                if($Protein_exons{$5}){
                        #protein has exon info already, add this one to the
list
                        $Protein_exons{$5} .= "\t$1 $2 $3 $4";
                        #if($5 == 13783){print STDERR "$Protein_exons{$5}\n";}

                }else{
                        #protein has no exon info, create this entry as the
first
                        $Protein_exons{$5} = "$1 $2 $3 $4";
                        if($5 == 13783){print STDERR "$Protein_exons{$5}\n";}
                }
        }


        #if we are dealing with the start of a new protein line, break it down
to the protein header and update the global variables.
        if($Augustus_flag && m/^# start/){
                $Protein_count++;
                #get the protein header and put it in @Protein_holder
                @Protein_holder=split(" ", trim($_));
                @Protein_holder=split("_" , $Protein_holder[3]);
                $Temp_contig=$Protein_holder[1]; #get the contig name
                $Temp_protein=substr($Protein_holder[5], (rindex
($Protein_holder[5], ".") + 1 )); #get the protein name
                #initialize the hash bin for this protein in %Protein_SNPcount
                $Protein_SNPcount{$Temp_protein}=0;
                $next_line=<PROT_FILE>;
                chomp($next_line);
                @Protein_current=split("\t", $next_line); #get the next line
that contains protein information
                #set the hash bin for this protein in %Protein_length
                $Protein_length{$Temp_protein}=($Protein_current[4] -
$Protein_current[3]);
                $Total_protein+=$Protein_current[4] - $Protein_current[3];
                #set the protein string in the %Contig_proteins hash to
contain the appropriate information
                if($Contig_proteins{$Temp_contig}){#if there is an entry for
the current contig
                        $Contig_proteins{$Temp_contig} .=  "\t$Temp_protein
$Protein_current[3] $Protein_current[4]";
                }else{#if there is not an entry for the current contig
                        $Contig_w_protein++;
                        $Contig_proteins{$Temp_contig} =  "$Temp_protein
$Protein_current[3] $Protein_current[4]";
                }
        }
}#end while

$Total_intergene= $Total_intergene-$Total_protein;

print "# of contigs with predicted proteins: $Contig_w_protein\n";
```

```
print "# of proteins: $Protein_count\n";
print "Total number of contigs: $Contig_count\n";
print "Total number of nucleotides in gene models: $Total_protein\n";
print "Total number of nucleotides in intergenic regions: $Total_intergene\n";


##############begin SNP processing algorithm
$Temp_contig="";
$Last_contig="";
$Current_contig="";
@Protein_holder= ();
@Protein_current= ();

while(<SNP_FILE>){

        chomp;
        if(m/^Chrom/){ next;}
        @SNP_holder = split("\t", trim($_));
        #this is a test to determine how many SNPs would be eliminated if we
use the number of SNPs
        #predicted as a the denominator in bonferroni correction (ie, 69356
SNPs found... .0001/69356
        #(ie, 483084 SNPs found in Rvin... .0001/483084= 2e-10 for Rvin
        #(ie, 475074 SNPs found in Rves... .0001/475074= 2.1e-10 for Rves
        #if($SNP_holder[11] > 0.00000032){next;}
        #if($SNP_holder[11] > 0.0000000002){next;}
        #if($SNP_holder[11] > 0.00000000021){next;}

        @Contig_holder = split("_", $SNP_holder[0]);
        $Current_contig = $Contig_holder[1];

        #this is a test to determine how many SNPs would be eliminated if we
use the number of SNPs
        #predicted as a the denominator in bonferroni correction (ie, 69356
SNPs found... .0001/69356
        #(ie, 483084 SNPs found in Rvin... .0001/483084= 2e-10 for Rvin
        #(ie, 475074 SNPs found in Rves... .0001/475074= 2.1e-10 for Rves
        if($SNP_holder[11] > 0.00000002){next;}
        #if($SNP_holder[11] > 0.0000000002){next;}
        #if($SNP_holder[11] > 0.00000000021){next;}
        #increment the global and Contig SNP counts
        $Contig_SNPcount{$Current_contig}++;
        $SNP_count++;
        $Last_contig=$Temp_contig;
        $Temp_contig=$Current_contig;
        #print $Temp_contig"";
        #check to see if we are looking at SNPs from a new contig, or the last
contig we looked at
        if($Temp_contig ne $Last_contig && $Contig_proteins{$Current_contig}){
                #we are looking at a new contig! get the proteins list for
that contig
                @Protein_holder=split("\t", $Contig_proteins{$Current_contig});
                #print "%Contig_proteins holds these items for contig number
$Current_contig: @Protein_holder\n" and die();
                #foreach (@Protein_holder){
                #       print $_ . "\n";
                #}
                #$Protein_flag=1;
        }
        #set the $Last_contig examined to be the $Current_contig we are
```

```
examining
        $Last_contig=$Current_contig;

        ## before beginning Protein processing, we need to empty
@Protein_current if the protein we were previously examining is not
        ## on the current contig
        if(@Protein_current){
                #### if the current protein being examined is not on the
current contig, then
                #### set the current protein array (@Protein_current) to empty
                if(!protein_on_contig($Protein_current[0], $Current_contig)){
                        @Protein_current=();
                }
        }

        ####Are we currently examining a protein? if so keep looking to see if
the SNP falls within its range
        if(@Protein_current){
                #check to see if the SNP position is within the current protein
                if(($SNP_holder[1] >= $Protein_current[1]) && ($SNP_holder[1]
<= $Protein_current[2])){
                        $Protein_SNPcount{$Protein_current[0]}++;
                        $SNP_count_protein++;
                        #check to see if the SNP is exonic or intronic and if
exonic, determine the synonymy of the SNP
                        ## This is performed with the SNP_status() subroutine
                        SNP_status();
                        ## we have gotten to the end of the gene model region
and not found any Exon into which the SNP falls,
                        ## go onto the next SNP
                        next;
                }## end if(in gene_model region)
                #if the SNP position is beyond the end of the protein, we can
stop looking at this protein
                if($SNP_holder[1] > $Protein_current[2]){
                        #### set the current protein array (@Protein_current)
to empty
                        @Protein_current=();
                }
        }#end if(@Protein_current)

        ######if, in our current position, we are not currently examining a
protein, get the  next protein
        if(@Protein_holder){
                $Temp_protein = shift(@Protein_holder);
                @Protein_current = split(" ", $Temp_protein);
        }else{
                #there are no proteins to process, go on to the next SNP line
                next;
        }

        #check to see if the SNP position is within the current protein
        if(($SNP_holder[1] >= $Protein_current[1]) && ($SNP_holder[1] <=
$Protein_current[2])){
                $Protein_SNPcount{$Protein_current[0]}++;
                $SNP_count_protein++;
                #check to see if the SNP is exonic or intronic and if exonic,
determine the synonymy of the SNP
                ## This is performed with the SNP_status() subroutine
                SNP_status();
```

```
                ## we have gotten to the end of the gene model region and not
found any Exon into which the SNP falls,
                ## go onto the next SNP
                next;
        }
        #if the SNP position is less than the start of the current protein,
get the next protein
        while(($SNP_holder[1] < $Protein_current[1]) && @Protein_holder){
                $Temp_protein=shift(@Protein_holder);
                @Protein_current = split(" ", $Temp_protein);
        }
        if($SNP_holder[1] <= $Protein_current[2]){
                $Protein_SNPcount{$Protein_current[0]}++;
                $SNP_count_protein++;
                #check to see if the SNP is exonic or intronic and if exonic,
determine the synonymy of the SNP
                ## This is performed with the SNP_status() subroutine
                SNP_status();
                ## we have gotten to the end of the gene model region and not
found any Exon into which the SNP falls,
                ## proceed to the the next SNP (automatic since we are at the
end of this code block)
        }
}#end SNP while

print STDERR "A total of $SNP_count SNPs were identified with\n";
print STDERR ($SNP_count-$SNP_count_protein) . " intergenic SNPs and ".
$SNP_count_protein . " SNPs in gene models\n";
print STDERR "Of the SNPs in gene models $SNP_count_exonic fall within coding
sequences (exons)\n";
print STDERR "with " . ($SNP_count_exonic-$SNP_count_nonsyn). " synonymous
SNPs and $SNP_count_nonsyn nonsynonymous SNPs\n";

#####now calculate the normalized values for Contigs  and print out the top
number up to the cutoff
#normalize the contig values
while((my $contig, my $length) = each(%Contig_length)) {
        $Contig_norms{$contig}= ($Contig_SNPcount{$contig}/
$length);
}
#remove all of the zero values from the contig length array
foreach my $contig (keys %Contig_norms){
        if($Contig_norms{$contig} == 0) {delete $Contig_norms{$contig}};
}
#$Contig_count=1;
$Temp_contig= basename($snp_file) . ".sorted_contigs.snp";
open(CONTIG_OUT, ">", $Temp_contig) or die("could not open SNP Contig ouput
file $Temp_contig: $!\n");

#sort the contig length array and print all entries to an output file
foreach my $contig (sort {$Contig_norms{$b} <=> $Contig_norms{$a} }
           keys %Contig_norms)
{
    #if($Contig_count > $cutoff){last;}
    print CONTIG_OUT "$contig $Contig_SNPcount{$contig} $Contig_length
{$contig} $Contig_norms{$contig}\n";
    #$Contig_count++;
}

#####now calculate the normalized values for Proteins and print out the top
```

```
number up to the cutoff
#normalize the contig values
while((my $protein, my $length) = each(%Protein_length)) {
        $Protein_norms{$protein}= ($Protein_SNPcount{$protein}/
$length);
}
#remove all of the zero values from the contig length array
foreach my $protein (keys %Protein_norms){
        if($Protein_norms{$protein} == 0) {delete $Protein_norms{$protein}};
}
#$Contig_count=1;
$Temp_protein= basename($snp_file) . ".sorted_proteins.snp";
open(PROTEIN_OUT, ">", $Temp_protein) or die("could not open SNP protein ouput
file $Temp_protein: $!\n");

#sort the contig length array and print all entries to an output file
foreach my $protein (sort {$Protein_norms{$b} <=> $Protein_norms{$a} }
           keys %Protein_norms)
{
    #if($Contig_count > $cutoff){last;}
    print PROTEIN_OUT "$protein $Protein_SNPcount{$protein} $Protein_length
{$protein} $Protein_norms{$protein}\n";
    #$Contig_count++;
}

close(CONTIG_OUT);
close(PROTEIN_OUT);
close(PROT_FILE);
close(SNP_FILE);

##################Begin PERL subroutine definitions##################
#####################################################################

## This subroutine uses global variables to determine the position (intronic
or exonic) and synonymy if SNP is exonic
sub SNP_status(){

        @Exon_holder=split("\t", $Protein_exons{$Protein_current[0]});
        #print STDERR "We are dying on protein $Protein_current[0] and contig
$Current_contig\n" and die();
        #print STDERR "The exon holder array holds these values, and then we
die... : @Exon_holder \n" and die();
        while(@Exon_holder){
                @Exon_current=split(" ",trim(shift(@Exon_holder)));
                if(($SNP_holder[1] >= $Exon_current[0]) && ($SNP_holder[1] <=
$Exon_current[1])){
                        #SNP is exonic! check to see if it is synonymous or
non-synonymous and break the loop
                        #if($Exon_current[0]){print STDERR "It's the
Exon_current array\n" and die();}
                        $SNP_count_exonic++;
                        #if($Exon_current[2] eq "-"){last;}
                        #grab just the exon
                        $active_seq=substr($Contig_sequence{$Current_contig},
($Exon_current[0] - 1), ($Exon_current[1] - $Exon_current[0]+1));
                        #set a variable to hold the SNP_position relative to
the start of the exon
                        #this value is 0 based
                        $SNP_pos=$SNP_holder[1]-$Exon_current
[0];
```

```
                        #is the protein on the negative strand? if so, reverse
complement the sequence
                        #and transform the $SNP_pos to start from the
beginning of the negative strand
                        #keep this value 0 based
                        if($Exon_current[2] eq "-"){
                                $active_seq=reverse_complement($active_seq);
                                $SNP_pos=($Exon_current[1]-$Exon_current[0]+1)-
($SNP_pos+1);
                                $Reverse_flag=1;
                        }else{
                                $Reverse_flag=0;
                        }
                        #OK, even though the SNP is in the coding sequence, it
might be placed
                        #outside of the actually coding area (if it is less
than the phase).
                        #if this is the case, it is not a true exonic SNP,
skip it (and keep the
                        #program from choking) by exiting the loop and looking
for the next SNP
                        if($SNP_pos < $Exon_current[3])
{last;}

                        #if($Current_contig == 26965){print $active_seq . " is
the active sequence and $SNP_pos is the active snp\n" and die();}

                        $Nuc_match= substr($active_seq, $SNP_pos, 1);
                        $Nuc_ref=$SNP_holder[2];
                        if($Reverse_flag){$Nuc_ref=reverse_complement
($Nuc_ref);}
                        ##this test block is choking on line 3837 of the snp
file, NODE_2851 with length 433, SNP at 338
                        ##this is a negative strand protein, the substring is
being created beyond the upper bound
                        ##also choking on line 13222 of SNP file, this is a
positive strand protein, the substring is being
                        ## created below the lower bound. This is due to the
substr on 266 that is calling the beginning of the
                        ## codon before the start of the protein
                        ## OK, problem on 13222 solved, problem on 3837 is not
due to snp processing, but rather to inproper
                        ## gene associations. There are no gene models on
contig 2851!, yet some how we found our way into this
                        ## code block, I suspect that the pattern matching for
gene names needs to be refined.

                        #if($Nuc_ref ne $Nuc_match){
                        #       print("the nucleotides $Nuc_ref and " .
$Nuc_match . " at position " . ($SNP_pos +1) . " don't match\n");
                        #}else{
                        #       print("the nucleotides $Nuc_ref and " .
$Nuc_match . " at position " . ($SNP_pos +1) . " match\n");
                        #}
                        #now get the codon that the SNP is in, and see if it
is changed by the SNP
                        #get the position in codon (zero based) ie, 0=1st
position
                        $SNP_posINcodon= ($SNP_pos - $Exon_current[3])%3;
                        $Codon_lowerbound = $SNP_pos - $SNP_posINcodon;
```

```
                        $Codon_old = substr($active_seq,$Codon_lowerbound,3);
                        $Codon_new = $Codon_old;
                        $SNP_value = $SNP_holder[18];
                        if($Reverse_flag){$SNP_value=reverse_complement
($SNP_value);}
                        substr($Codon_new, $SNP_posINcodon, 1, $SNP_value);
                        if(!acid_test($Codon_old, $Codon_new))
{$SNP_count_nonsyn++}
                        #print "the old reference codon is $Codon_old and the
SNP codon is $Codon_new \n";
                        #print("The Codon sequence is $Codon_old starting at
postion ".($Codon_lowerbound +1). " with a phase of " . $Exon_current[3].  "
and a SNP position at site ". ($SNP_posINcodon + 1). " of the codon, and the
strand is " . $Exon_current[2] . "\n");

                        ## we have located the snp position and determined its
exonic status, break the loop
                        last;
                }## end if(inexon)
        }#end exonic SNP Loop



}#end subroutine SNP_status


## this subroutine takes a gene ID and a contig ID as arguement
## returns true if the gene is found on the current contig
## invoked as such protein_on_contig(geneID, contigID)
sub protein_on_contig(@){
        my $return_value=0;
        my @temp_proteins;
        my $test_protein;
        my @protein_array;
        if($Contig_proteins{$_[1]}){
                @temp_proteins=split("\t", $Contig_proteins{$_[1]});
                while(@temp_proteins){
                        $test_protein=shift(@temp_proteins);
                        @protein_array=split(" ", $test_protein );
                        if($protein_array[0] eq $_[0]){$return_value=1;}
                }
        }
        return $return_value;
}#end subroutine protein_on_contig()


#subroutine that will return true if the SNP is synonymous, false otherwise
#takes two codons as arguement and basically runs a switch statement
#invoked as follows acid_test($reference_codon_string, $SNP_codon_string)
sub acid_test(@) {
        my $reference_AA;
        my $SNP_AA;

        switch($_[0]){
                case "TTT"      {$reference_AA="F"}
                case "TTC"      {$reference_AA="F"}
                case "TTA"      {$reference_AA="L"}
                case "TTG"      {$reference_AA="L"}
                case "CTT"      {$reference_AA="L"}
                case "CTC"      {$reference_AA="L"}
```

```
        case "CTA"      {$reference_AA="L"}
        case "CTG"      {$reference_AA="L"}
        case "ATT"      {$reference_AA="I"}
        case "ATC"      {$reference_AA="I"}
        case "ATA"      {$reference_AA="I"}
        case "ATG"      {$reference_AA="M"}
        case "GTT"      {$reference_AA="V"}
        case "GTC"      {$reference_AA="V"}
        case "GTA"      {$reference_AA="V"}
        case "GTG"      {$reference_AA="V"}
        case "TCT"      {$reference_AA="S"}
        case "TCC"      {$reference_AA="S"}
        case "TCA"      {$reference_AA="S"}
        case "TCG"      {$reference_AA="S"}
        case "CCT"      {$reference_AA="P"}
        case "CCC"      {$reference_AA="P"}
        case "CCA"      {$reference_AA="P"}
        case "CCG"      {$reference_AA="P"}
        case "ACT"      {$reference_AA="T"}
        case "ACC"      {$reference_AA="T"}
        case "ACA"      {$reference_AA="T"}
        case "ACG"      {$reference_AA="T"}
        case "GCT"      {$reference_AA="A"}
        case "GCC"      {$reference_AA="A"}
        case "GCA"      {$reference_AA="A"}
        case "GCG"      {$reference_AA="A"}
        case "TAT"      {$reference_AA="Y"}
        case "TAC"      {$reference_AA="Y"}
        case "TAA"      {$reference_AA="*"}
        case "TAG"      {$reference_AA="*"}
        case "CAT"      {$reference_AA="H"}
        case "CAC"      {$reference_AA="H"}
        case "CAA"      {$reference_AA="Q"}
        case "CAG"      {$reference_AA="Q"}
        case "AAT"      {$reference_AA="N"}
        case "AAC"      {$reference_AA="N"}
        case "AAA"      {$reference_AA="K"}
        case "AAG"      {$reference_AA="K"}
        case "GAT"      {$reference_AA="D"}
        case "GAC"      {$reference_AA="D"}
        case "GAA"      {$reference_AA="E"}
        case "GAG"      {$reference_AA="E"}
        case "TGT"      {$reference_AA="C"}
        case "TGC"      {$reference_AA="C"}
        case "TGA"      {$reference_AA="*"}
        case "TGG"      {$reference_AA="W"}
        case "CGT"      {$reference_AA="R"}
        case "CGC"      {$reference_AA="R"}
        case "CGA"      {$reference_AA="R"}
        case "CGG"      {$reference_AA="R"}
        case "AGT"      {$reference_AA="S"}
        case "AGC"      {$reference_AA="S"}
        case "AGA"      {$reference_AA="R"}
        case "AGG"      {$reference_AA="R"}
        case "GGT"      {$reference_AA="G"}
        case "GGC"      {$reference_AA="G"}
        case "GGA"      {$reference_AA="G"}
        case "GGG"      {$reference_AA="G"}
        else            {return 1}
}#end of $reference_AA switch
```

```
switch($_[1]){
        case "TTT"      {$SNP_AA="F"}
        case "TTC"      {$SNP_AA="F"}
        case "TTA"      {$SNP_AA="L"}
        case "TTG"      {$SNP_AA="L"}
        case "CTT"      {$SNP_AA="L"}
        case "CTC"      {$SNP_AA="L"}
        case "CTA"      {$SNP_AA="L"}
        case "CTG"      {$SNP_AA="L"}
        case "ATT"      {$SNP_AA="I"}
        case "ATC"      {$SNP_AA="I"}
        case "ATA"      {$SNP_AA="I"}
        case "ATG"      {$SNP_AA="M"}
        case "GTT"      {$SNP_AA="V"}
        case "GTC"      {$SNP_AA="V"}
        case "GTA"      {$SNP_AA="V"}
        case "GTG"      {$SNP_AA="V"}
        case "TCT"      {$SNP_AA="S"}
        case "TCC"      {$SNP_AA="S"}
        case "TCA"      {$SNP_AA="S"}
        case "TCG"      {$SNP_AA="S"}
        case "CCT"      {$SNP_AA="P"}
        case "CCC"      {$SNP_AA="P"}
        case "CCA"      {$SNP_AA="P"}
        case "CCG"      {$SNP_AA="P"}
        case "ACT"      {$SNP_AA="T"}
        case "ACC"      {$SNP_AA="T"}
        case "ACA"      {$SNP_AA="T"}
        case "ACG"      {$SNP_AA="T"}
        case "GCT"      {$SNP_AA="A"}
        case "GCC"      {$SNP_AA="A"}
        case "GCA"      {$SNP_AA="A"}
        case "GCG"      {$SNP_AA="A"}
        case "TAT"      {$SNP_AA="Y"}
        case "TAC"      {$SNP_AA="Y"}
        case "TAA"      {$SNP_AA="*"}
        case "TAG"      {$SNP_AA="*"}
        case "CAT"      {$SNP_AA="H"}
        case "CAC"      {$SNP_AA="H"}
        case "CAA"      {$SNP_AA="Q"}
        case "CAG"      {$SNP_AA="Q"}
        case "AAT"      {$SNP_AA="N"}
        case "AAC"      {$SNP_AA="N"}
        case "AAA"      {$SNP_AA="K"}
        case "AAG"      {$SNP_AA="K"}
        case "GAT"      {$SNP_AA="D"}
        case "GAC"      {$SNP_AA="D"}
        case "GAA"      {$SNP_AA="E"}
        case "GAG"      {$SNP_AA="E"}
        case "TGT"      {$SNP_AA="C"}
        case "TGC"      {$SNP_AA="C"}
        case "TGA"      {$SNP_AA="*"}
        case "TGG"      {$SNP_AA="W"}
        case "CGT"      {$SNP_AA="R"}
        case "CGC"      {$SNP_AA="R"}
        case "CGA"      {$SNP_AA="R"}
        case "CGG"      {$SNP_AA="R"}
        case "AGT"      {$SNP_AA="S"}
        case "AGC"      {$SNP_AA="S"}
```

```
                case "AGA"      {$SNP_AA="R"}
                case "AGG"      {$SNP_AA="R"}
                case "GGT"      {$SNP_AA="G"}
                case "GGC"      {$SNP_AA="G"}
                case "GGA"      {$SNP_AA="G"}
                case "GGG"      {$SNP_AA="G"}
                else            {return 1}
        }#END SNP_AA switch

        ($reference_AA eq $SNP_AA)
}#end of acid_test() subroutine




#subroutine to reverse complement a string of DNA sequence
sub reverse_complement($) {
        my $dna = shift;

        # reverse the DNA sequence
        my $revcomp = reverse($dna);

        # complement the reversed DNA sequence
        $revcomp =~ tr/NACGTacgt/NTGCAtgca/;
        return $revcomp;
}




# Perl trim function to remove whitespace from the start and end of the string
sub trim($)
{
        my $string = shift;
        $string =~ s/^\s+//;
        $string =~ s/\s+$//;
        return $string;
}

# PERL usage function for this program, SNP_density_calc.pl
sub usage
{
        print STDERR << "EOF";

        Name $prg - Calculates the density of SNPs at the
                        level of protein, contig, and determines rates
                        of synonymous and non-synonymous mutations

        usage:  $prg [-c cutoff] [-f frame_size] (-a || -m || -j)
                (-v VarScanSNPinputfile.snp || -s SamtoolsSNPinputfile.snp)
                augustus_proteins.gff3 Contigs.fasta

        -h              : print this help message
        -a              : specify augustus format in .gff3 file
        -m              : specify maker format in .gff3 file
        -j              : specify jgi format in the .gff3 file
        -s input        : SNP input file is in Samtools format, sets inputfile
        -v input        : SNP input file is in Varscan format, sets inputfile
```

```
        -c cutoff     : sets the cutoff value (the number of
                        significant proteins, contigs,
                        frames to report) (default = 10)
        -f frame_size : Sets the frame size to use in sliding window analysis
(default=10000)

        ex: $prg -s Samtools.snp augustus_predictions.gff3 contigs.fa

EOF
        exit;

}
```

```
#!/usr/bin/perl -w

#Name superaligner.pl - This program will concatenate all sequence data for
each taxon
#                       and print a new concatenated super alignment combining
all protein
#                       alignments into a single .fasta file.
#See usage() function at the end of program for details and usage instructions.

#use warnings;
use strict;
use Getopt::Std;
use File::Basename;
use Bio::SeqIO;
use Bio::DB::Fasta;


sub reverse_complement($);
sub trim($);
sub usage;


#declare the global variables
#hashes

my %Super_sequence;
my %Sample_status;

#arrays
my @Protein_holder;
my @Protein_current;
my @Exon_holder;
my @Exon_current;
my @Exons_toPrint;
my @target_list;

#strings
my $current_strand;
my $Temp_contig;
my $Current_contig;
my $Last_contig;
my $Current_protein;
my $next_line;
my $Fail_string;
my $active_seq;
my $temp_header;
my $active_protein_seq;
my $temp_protein_header;
my $Seq_directory;
my $Sample_file;

#integers
my $jgi_flag=0;
my $maker_flag=0;
my $Reverse_flag=0;
my $Fail_flag_120=0;
my $Fail_flag_960=0;
my $Failed_proteins=0;
my $Passed_proteins=0;
my $Exons_totalLength=0;
```

```
my $Passed_exons_totallength=0;
my $Sample_count=0;
my $Print_count=0;
my $Sequence_count=0;
my $Cumulative_length=0;

#declare the options string
my $opt_string='hd:l:';
my $prg = basename $0;
my %opt;
my $reference_header;
my $target_header;
my $list_file;
my $gff_file;
my $contig_file;

####process the options and set global variables for options
getopts("$opt_string", \%opt) or usage();
$opt{h} and usage();
usage() unless ($opt{d} && $opt{l});
print "Could not open sample list file: $opt{l}\n" and die(usage()) unless (-e
$opt{l});
print "Could not read directory: $opt{d}\n" and die(usage()) unless (-e $opt{d}
);

######inform the user of program settings
$opt{d} and print STDERR "Combining target alignment files stored in the
directory: $opt{d}\n";
$opt{l} and print STDERR "using taxon reference list from input file: $opt{l}
\n";


#set the input files
$Seq_directory = $opt{d};
$Sample_file = $opt{l};


# open up BLAST and ASSEMBLY directories and reference file
opendir(SEQDIR, $Seq_directory) or die "can't open directory: $Seq_directory $!
\n";

# if input directory does not end with '/', add it.
#This is for the purposes of printing files later on, no need for the opening
of the directory.
if ($Seq_directory !~ /\/$/){
        $Seq_directory .= "/";
}

#declare the sequence handle and container stuff
#my $sequence_handle=Bio::SeqIO->new( -file=>$Sample_file, -format=>"Fasta");
#my $db = Bio::DB::Fasta->new($Sample_file);
#my @active_header;
open(SAMPLE_LIST, $Sample_file) or die("Could not open the Sample file:
$Sample_file. $!\n");

#populate the reference sequence hash with all of the reference sequences
while(<SAMPLE_LIST>){
        chomp($_);
        $Super_sequence{$_}="";
        $Sample_status{$_}=0;
```

```
        $Sample_count++;
}#end while

###loop through all of the files in the input directory and input sequences
where appropriate
foreach my $f (readdir (SEQDIR)){
        next if ($f =~ /^\./); # skip if this file is default '.' or '..'
directory
        next if ($f !~ /\.fasta$/); # skip if this file does not end with the
right file suffix
        my ($id,$jnk) = split(/_/,$f);
        chomp($id);
        my $in_file  = $Seq_directory.$f;
        #declare the sequence handle and container stuff
        my $infile_handle=Bio::SeqIO->new( -file=>$in_file, -format=>"Fasta");
        my $db = Bio::DB::Fasta->new($in_file);

        #make a variable to store the sequence alignment length for this file
        my $align_length=0;
        my $first_pass=1;

        #populate the reference sequence hash with all of the reference
sequences
        while(my $seq_obj=$infile_handle->next_seq){
                my $temp_header=$db->header($seq_obj->id);
                chomp($temp_header);
                my @active_header=split(/\./, $temp_header);
                $active_seq= $seq_obj->seq;
                #set the alignment length and protein id if they are not set
                if($first_pass){
                        push(@Protein_holder, $active_header[1]);
                        $align_length=length($active_seq);
                        $Cumulative_length+=$align_length;
                        $Sequence_count++;
                        $first_pass=0;
                }
                #insert the sequence into a the Super_sequence hash if an
entry exists, otherwise
                #you should pitch an error and exit.
                if(exists($Super_sequence{$active_header[0]})){
                        #if this sample id is in the the sample list then
insert the sequence into a hash
                        #give passing status to the sample and store the
protein id for the order array
                        $Super_sequence{$active_header[0]}.=$active_seq;
                        $Sample_status{$active_header[0]}=1;

                }else{
                        #crud, we found an ID that does not exist in the
sample list, exit with error
                        print STDERR "Sequence sample id $active_header[0]
from file $f does not match any ids from the sample list\n" and die
("$!");
                }
        }#end while

        #OK, we've populated the hash list with proteins that are in the
alignment, now let's fill it with proteins that
        #are not in the alignment by printing a bunch of "-" (gap) characters
that are equal to the length of the alignment
```

```
        while(my ($key,$value) = each %Sample_status){
                #this Sample id got a sequence printed, move on without
printing anything more
                if($value==1){
                        $Sample_status{$key}=0;
                        next;
                }
                #OK, this sample is not OK, we need to print some gap
information for it
                $Super_sequence{$key}.=('-' x $align_length);
        }
}#end foreach

##declare the sequence handle and stuff
my @basename=split(/\./, basename($Sample_file));
my $outfile = $basename[0] . "_superAlignment.fasta";
my $Outfile_handle=Bio::SeqIO->new( -file=>">$outfile", -format=>"Fasta");


#now we just have to print the super alignment... go for it!
while(my ($key,$value) = each %Super_sequence){

        my $seq_obj_out= Bio::Seq->new(-seq=>"$value", -display_id=>"$key", -
alphabet=>"dna");
        $Outfile_handle->write_seq($seq_obj_out);
        $Print_count++;
}

##OK, just about done, now we just need to print the protein order to a file
my $outfile2= $basename[0] . "_superAlignment_sequenceORDER.txt";
open(ORDER, ">$outfile2") or die("Could not open file $outfile2 for printing:
$!\n");
print ORDER "Printed a total $Sequence_count sequences as $Print_count
concatenated sequences for $Sample_count Samples into the file $outfile\n";
print ORDER "Total length of super alignment = $Cumulative_length\n";
print ORDER "Sequences were concatenated in the following order:\n";

while(@Protein_holder){
        print ORDER shift(@Protein_holder) . "\n";
}

print STDERR "Printed a total $Sequence_count sequences as $Print_count
concatenated sequences for $Sample_count Samples into the file $outfile\n";
print STDERR "Total length of super alignment = $Cumulative_length\n";
print STDERR "Printed list of proteins in the order that they were added to
alignment into file $outfile2\n ";
close(ORDER);

##################Begin PERL subroutine definitions##################
#####################################################################


#subroutine to reverse complement a string of DNA sequence
sub reverse_complement($) {
        my $dna = shift;

        # reverse the DNA sequence
        my $revcomp = reverse($dna);

        # complement the reversed DNA sequence
```

```
        $revcomp =~ tr/NACGTacgt/NTGCAtgca/;
        return $revcomp;
}



# Perl trim function to remove whitespace from the start and end of the string
sub trim($)
{
        my $string = shift;
        $string =~ s/^\s+//;
        $string =~ s/\s+$//;
        return $string;
}

# PERL usage function for this program, SNP_density_calc.pl
sub usage
{
        print STDERR << "EOF";

        Name $prg - Takes as input a newline delimited sample list and a
directory which contains alignments
                        for all those samples for proteins of interest. This
program will concatenate all sequence
                        data for each taxon and print a new concatenated super
alignment combining all protein
                        alignments into a single .fasta file. If a given taxon
is missing from a particular
                        alignment file then superaligner.pl will print N
"-" (gap) character for that taxon
                        where N equals the length of the alignment in question.
                        protein header format expected: TaxonID.proteinID
                        File format expected: ProteinID_SomeInformation.fasta

                        NOTE: file names of fasta files must end in .fasta

        usage:  $prg -d directory -r sequences_to_insert.fasta

        -h              :       print this help message
        -d directory    :       Path to a directory which contains nucleotide/
protein alignments that
                                with have sequence data inserted into them,
sets input directory
        -l list         :       A newline delimited file with one taxon name
per line, sets input file.

        ex: $prg -d Some_directory/ -r my_sequences.fasta

EOF
        exit;

}
```