UNEARTHING THE MECHANISMS OF THE MYCORRHIZAL-BACTERIAL SYMBIOSIS IN PLANT ROOTS USING A METATRANSCRIPTOMIC APPROACH

BY

GAUTAM NAISHADHAM

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Bioinformatics
with a concentration in Crop Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Master's Committee:

Professor Matthew Hudson
Senior Research Scientist Liudmila Mainzer
Professor Gustavo Caetano-Anolles

Abstract

Arbuscular mycorrhizal fungi participate in a widely conserved symbiosis with a majority of land plants which provides plant hosts with increased capability for soil nutrient uptake. These endosymbiotic fungi are themselves colonized by a diverse group of bacteria, including both parasitic and symbiotic species. Recently several obligate endosymbionts of the arbuscular mycorrhizal fungi have been identified, and these bacteria have been shown to modulate both the metabolism and morphology of the fungal symbionts. However, molecular and functional characterization of these bacterial endosymbionts has been limited by an inability to isolate and culture such obligate symbionts, which have significant metabolic dependencies on the host fungi. In this work, a metatranscriptomic approach is applied in order to determine the transcriptional mechanisms underlying this multilayered symbiosis. Different mycorrhizal fungal species were found to be colonized by distinct communities of bacteria, and the study identified bacterial genes with significant differential abundance in mycorrhiza-inoculated plant roots as well as bacterial genes with varying abundance across the life cycle of the symbiosis. Overall, arbuscular mycorrhizal fungi harbor a diverse and metabolically active community of bacteria, and metatranscriptomics provides a capable tool to uncover the functional basis of such complex, obligate symbioses.

Table of Contents

Introduction

Arbuscular mycorrhizal (AM) fungi are a group of endosymbiotic fungi associated with the roots of a majority of plants belonging to the phylum *Glomeromycota*. They participate in a mutual symbiosis with over 80% of species of terrestrial plants (Lee et al. 2013), providing plants with essential nutrients such as phosphorous. In return, the AM fungi receive photosynthesis-derived carbon. Arbuscular mycorrhizal fungi confer these benefits by greatly expanding plant access to soil phosphorous. The fungi penetrate root cortical cells and form dense, branched structures called arbuscules within root cells to allow for nutrient exchange. Outside of the roots, AM fungi form extensive networks of hyphae that protrude into the soil. These hyphal networks provide greatly increased absorption of soil-derived nutrients by both exploring a greater volume of soil and by increasing the surface area over which these nutrients may diffuse (Bolan 1991). The fungi also provide more efficient nutrient uptake in a variety of nutrient-deficient conditions, primarily due to the increased surface-to-volume ratio of the mycorrhiza and a fungal capacity for rapid growth (Tuomi, Kytoviita, and Hardling 2001). This symbiosis is highly conserved, and it is believed that the arbuscular mycorrhiza evolved in conjunction with the first land plants (Brundrett 2002). With the ability to increase plant access to nutrients in deficient soils, the mycorrhizal symbiosis is of direct importance to maximizing crop yield and resilience for the global food supply and bioenergy feedstocks.

Full description of the mycorrhizal symbiosis must consider the microbial community associated with the fungi. The AM fungi are readily colonized by several bacterial species, including both parasitic and symbiotic organisms. Several species of the *Burkholderia* genus, which are largely soil based pathogens that can also infect humans, are able to invade AM fungal spores (Levy et al. 2003). Additionally, an endosymbiont of the AM fungi has been identified

(V. Bianciotto et al. 1996). Candidatus *Glomeribacter gigasporarum*, which relies on its fungal host for survival, can modulate the metabolic profile of the mycorrhiza (Salvioli et al. 2010) and may influence hyphal branching in the fungi (Lumini et al. 2007) by enhancing responsiveness to branching signals from the plant host. As hyphal branching directly affects nutrient uptake by the fungi, an understanding of these bacteria and the mechanisms of their symbiosis with the fungi is essential. A second group of obligate endosymbionts of the mycorrhizal fungi has been described more recently (Naumann, Schüßler, and Bonfante 2010). Unlike the *Ca. G. gigasporarum*, which exclusively colonize arbuscular mycorrhizal fungi of the family *Gigasporaceae*, the *Mollictures*-related endobacteria (MRE) are able to colonize arbuscular mycorrhizal fungi from a much more diverse lineage. These endobacteria are similarly nutritionally dependent on their hosts and show a much reduced genome (Naito, Morton, and Pawlowska 2015).

The obligate nature of the bacterial symbiosis, however, has made the study of these communities challenging, as the bacteria cannot be cultured. Indeed, a genomic sequencing survey of *Ca. G. gigasporarum* showed a greatly reduced genome with a metabolic dependence on the fungal host (Ghignone et al. 2012). The bacteria have a reduced metabolic capacity for the synthesis of several amino acids as well as the degradation of complex sugars. This results in an inability to propagate in host-free media, although the bacteria can be kept alive after removal from the host (Jargeat et al. 2004). The evolutionary trajectory of the *Ca. G. gigasporarum* genome suggests an ancient symbiosis with the AM fungi (Castillo and Pawlowska 2010) and it is likely that *Ca. G. gigasporarum* became increasingly metabolically dependent on the host as the symbiosis evolved. Although the symbiosis exacts an energetic cost on the host fungi, the age and extent of the symbiosis indicates that the fungi enjoy a selective benefit from the morphological influence of these bacteria.

In order to determine the molecular mechanisms involved in the AM fungal-bacterial symbiosis, we conducted a metatranscriptomics study on the root system of *Brachypodium distachyon.* The study included three different arbuscular mycorrhizal symbioses with the fungal symbionts *Gigaspora gigantea, Glomus intraradices, and Glomus versiforme*. Transcriptomic approaches provide a survey of the full complement of genomic products (transcripts) being transcribed in a particular tissue sample. Although ideal transcriptomic experiments require isolation of an individual species' tissue and the exclusion of all contaminants and associated organisms, there exist situations where this may not be feasible or possible. The symbiosis of *Ca. G. gigasporarum* with arbuscular mycorrhizal fungi is a clear example of such a situation, as both the bacteria and its fungal host are obligate symbionts of their respective hosts, and both cannot be isolated and cultured individually. To accommodate this inability to sequence the fungi alone, a metatranscriptomic approach was utilized wherein RNA from all species present in an environment is sampled together, and transcriptomic sequencing is applied to this agglomerate sample. This type of analysis benefits from an ability to easily sample gene expression in complex environments from any number of known or unknown species. However, it entails the additional difficulty of computationally post-processing the sequences to individually analyze gene expression for the constituent species. In the present study, both root and shoot samples were studied. Root samples included RNA sequences from the host plant *B. distachyon*, one of the three fungal symbionts studied, as well as an undetermined number of bacterial symbionts or parasites of the fungal host. Although metatranscriptomic sequencing allowed observation of the yet-unexamined plant-fungal symbiosis as well as the fungal-bacterial symbiosis, functional and expression pattern analysis required robust isolation of sequences from the individual organisms present.

There are a large number of tools available that attempt to resolve the problem of separating metagenomic (or the derivative metatranscriptomic) sequencing reads into the respective taxonomic bins. This problem, known as taxonomic binning, is largely defined by the difficulties resulting from extensive sequence homology throughout the tree of life due to conservedness as well as convergent evolution. These homologous regions make it challenging to confidently place sequences within a particular taxonomic group, as high levels of similarity or even perfect identity may exist between sequences of sometimes widely divergent species. Therefore, the most robust approaches to the problem attempt to place the sequences in bins that are at higher levels of the taxonomic tree, resulting in less specific but more confident taxonomic assignments. The desire to place sequences in the most specific (lowest) taxonomic bin possible while maintaining a high level of confidence in the assignment has led to a number of tools being developed to address this problem, with differing focuses.

These tools can be placed into two overarching categories: similarity-based and composition-based binning. In similarity based binning, sequences are first compared to a comprehensive database of sequences, such as NCBI's nr database, which contains sequences from across the tree of life for which the taxonomic annotations are available. For each query sequence with matches in the database, the taxa of the database sequences for which there were matches is compiled, and a most-likely taxonomy for the query is chosen based on this set of taxa. Composition based binning techniques do not rely on the sequences themselves, but rather consider features of the sequences such as GC content, kmer abundance, and codon usage. Using a set of taxonomically-annotated sequences from across the tree of life, a model is trained using these features as predictors, and each query sequence is placed in the taxonomy based on its values for these features. Although these composition based techniques work well for longer

sequences, short sequences such as the reads produced by modern high-throughput sequencing machines do not contain enough compositional information to be accurately binned using these methods (Peabody et al. 2015). Therefore, similarity based methods are usually used for taxonomic annotation of short reads.

Similarity based binning methods, however, generally require a computationally intensive and time consuming alignment of the query sequences to a reference database. As the methods rely on matches to query sequences for taxonomic assignment, increasingly comprehensive databases may allow assignment of a greater proportion of queries. For this reason, widely inclusive databases such as the NCBI nr database are often used that contain hundreds of millions of reference sequences, resulting in a vastly increased computational workload. In addition, the alignment of query sequences to such a database requires an aligner capable of aligning fairly dissimilar sequences, and so aligners such as BLASTx and BLASTn are commonly used. Although capable of generating informative alignments between sequences of relatively low identity, these aligners usually require a costly dynamic programming step and are therefore incapable of processing the massive number of sequences produced by current high throughput sequencing runs in a reasonable amount of time. More recent tools have been developed to alleviate this problem in the large-scale sequencing projects currently being conducted, but the adoption of these tools in taxonomic binning software has been limited. Several available similarity based binning methods, such as CARMA3 (Gerlach and Stoye 2011) and DiScRIBinATE (Ghosh, Monzoorul Haque, and Mande 2010), provide excellent precision and sensitivity in their taxonomic assignments (Peabody et al. 2015), but they rely on the BLASTx software that is essentially unusable for aligning sequencing libraries numbering in the millions or tens of millions of reads. One similarity based tool for taxonomic binning, MEGAN

(Huson et al. 2007), provides moderate precision and sensitivity, and uses the DIAMOND (Buchfink, Xie, and Huson 2014) aligner to identify matches in the reference database. This aligner allows the alignment of dissimilar sequences at a very high rate, and is appropriate for aligning libraries numbering in the tens of millions. Very recent methods, such as KRAKEN (Wood and Salzberg 2014) and CLARK (Ounit et al. 2015), abandon the alignment step entirely and select taxonomic bins based on kmer distributions, but these generally show reduced sensitivity or precision compared to alignment based methods (Peabody et al. 2015). In order to assign a taxonomic annotation to the greatest proportion of reads with the highest accuracy possible in a reasonable amount of time, MEGAN was used for taxonomic binning in this experiment.

With a robust method for separating sequences in a metagenomic or metatranscriptomic study by their taxonomy of origin, it is possible to conduct standard genomic or transcriptomic analyses independently on each operational taxonomic unit (OTU). In the current study, bacterial sequences were separated from eukaryotic sequences and functional and differential expression analyses were conducted on the bacterial population in order to characterize genomic activity in the mycorrhizal microbiome. Additionally, the population structure of the fungal-root microbiome was determined using phylogenetic methods aided by the robust taxonomic binning provided by MEGAN. These analyses show a complex microbial community providing an array of functions related to symbiosis and endosymbiotic colonization in the mycorrhizal environment.

Methods

Experimental design

Brachypodium distachyon roots were inoculated with spores of either *Glomus intraradices*, *Glomus versiforme*, or *Gigaspora gigantea*. In addition, control treatments were grown which were not inoculated with any fungal symbionts. Plants were grown in a greenhouse at Cornell University in the Harrison lab. Tissue samples were collected from roots and shoots for each treatment at five and nine weeks after inoculation. Roots were washed to remove soil and surface-associated fungi and bacteria prior to sample preparation. Three biological replicates were collected for each treatment, and in total 48 samples were collected. Tissue samples were then ground under liquid nitrogen using mortar and pestles, and RNA was extracted from ground cell matter. Samples were sequenced using pyrosequencing from the Roche 454 protocol.

cDNA was also prepared from the RNA samples using the Illumina TruSeq SBS sequencing kit v3 and quantitated using qPCR before sequencing . These libraries were sequenced on an Illumina HiSeq2000 using 100 cycles. Root samples were using paired-end sequencing with an average insert size of 250bp, while shoot samples were sequenced using single-end sequencing. Sequences were first trimmed of any Illumina adapter sequences remaining, and reads were quality trimmed to remove low-quality bases from the 3’ ends. After reads were trimmed of adapters and low-quality bases, libraries were filtered to remove reads shorter than 25 bases. Adapter trimming, quality trimming, and length filtering were performed using a Perl script written by Gopal Battu in the Hudson laboratory.

Assembly using Roche 454 reads

Contigs were assembled from the Roche 454 sequencing reads using the ABySS sequence assembly software, and this work was performed by Liudmila Mainzer. As a high quality *B. distachyon* genome has already been published, assembly of *Brachypodium* transcripts was not the intent of this project, and so contigs were aligned to the *Brachypodium* genome and those with high confidence *Brachypodium* alignments were removed from further analyses.

Taxonomic composition

To determine the taxonomic composition of species present in the root and shoot tissue systems, samples were compared against a comprehensive database of one highly conserved gene, chaperonin 60 (Cpn60). Although 16S ribosomal RNA (rRNA) genes are commonly used to assess population structure in metagenomic samples, metatranscriptomics does not reliably capture these sequences due to the polyadenylation filtering used in mRNA library preparation. Cpn60 is a protein-coding gene and has been shown to have nearly universal coverage among the bacterial and eukaryotic domains of life as well as some coverage among archaeal lineages, and so it represents an ideal marker gene for transcriptomic phylogenetic analysis. Additionally, cpn60 genes may provide greater discriminatory power between closely related bacterial genomes (Brousseau et al. 2001) than 16S rRNA. To identify the population structure of the mycorrhizal microbiome, reads were aligned to a database of chaperonin 60 sequences (Hill 2004) from a diverse collection of organisms representing a large portion of the tree of life. STAR aligner (Dobin et al. 2013) was used with a minimum of 70% similarity required. Initially, these alignments themselves were used for assessing OTU abundance in the samples, but this showed a very high type I error rate likely due to sequencing errors. For this reason, the more robust taxonomic predictions provided by the Lowest Common Ancestor (LCA) algorithm, as

implemented in MEGAN (Huson et al. 2007), were assessed. LCA assignments using the cpnDB alignments were conducted using a "top-percent" filter of 10% (default).

Additional taxonomic classification was conducted using taxonomic binning with the comprehensive NCBI nr database. MEGAN's LCA algorithm was used for its ability to produce robust and efficient taxonomic assignment of the reads, and sequences annotated as bacterial were separated from eukaryotic sequences in order to exclude *Brachypodium* and fungal reads from expression analysis. Although more robust methods for this task exist (Peabody et al. 2015), such as CARMA3 (Gerlach and Stoye 2011) and DiScRIBinATE (Ghosh, Monzoorul Haque, and Mande 2010), the LCA algorithm was chosen for its ability to process tens of millions of sequences while providing acceptable Type I and Type II error rates for short reads. Blast2lca, a command line tool provided with MEGAN, was used to process the reads in parallel on a cluster. The DIAMOND aligner was used with the "fast" setting to compare the sequences against the NCBI nr database, and these alignments were processed using Blast2lca. In order to maintain specificity while using the more inclusive NCBI nr database, a minimum bit-score threshold of 15% of the top hit score ("--topPercent 15") was used to balance the precision and sensitivity of assignments. Default parameters used include an absolute minimum bit-score of 50 ("-ms 50") required for annotation as well as a 0.01 cutoff for the E-value of alignments ("-me 0.01").

Taxa to which reads were assigned were then filtered to retain only those taxa with a high likelihood of being truly present. The filtering of taxa was performed by first compiling counts-per-million (CPM) of the reads present in each taxonomic bin, normalizing raw read counts by the total number of reads in each respective library. Taxa were then selected which had a

minimum abundance at least 1CPM in at least three libraries. Specifically, only reads from side 1 of the paired-end reads from root samples were used in this portion of the study.

As LCA binning allows quantification of taxa at all levels, many of the taxa output are redundant, less-specific higher level parent taxa to lower taxa. If these taxa occur in the lineage of only one significantly abundant descendant taxon, then they are uninformative, since the lower taxon provides a more specific description of the lineage's presence in the sample. For this reason, these uninformative higher-level parent taxa were removed from further analysis to avoid assessing extraneous information as well as to maximize power while correcting test results for multiple comparisons. Taxa were selected for removal if they contained only one descendant taxon with significant abundance, as determined above, and if the total CPM across all libraries for the parent taxon was within 30% of the total CPM for the descendant taxon. This relative-abundance filter was included to ensure that parent taxa with high numbers of descendant taxa present in a sample, yet only one *significantly abundant* descendant taxon, were retained for further analysis, as the abundance of these higher level parent taxa may be of interest on their own. Due to the nature of output of the MEGAN Blast2lca utility, taxa which do not fall into a canonical rank (superkingdom, kingdom, phylum, class, order, family, genus, or species) could not be removed regardless of their redundancy.

With this filtered set of taxonomic abundances, expressed as CPM quantities, tests were performed to identify those taxa whose mean abundance varied significantly between different conditions. Two-sided Student's t-test, as implemented in the R function "t.test", was used for these comparisons, and p-values for each test were corrected using the False Discovery Rate (FDR) (Benjamini and Hochberg 1995) correction for multiple comparisons, implemented in the "p.adjust" function in R. Comparisons performed included those between individual mycorrhizal

fungal treatments and non-inoculated (control) treatments, as well as those between individual fungal-inoculated treatments at nine weeks-post-inoculation and five weeks-post-inoculation. Further, a comparison of mean CPM between root and shoot samples was conducted, and those taxa with significantly higher binned read counts in shoots were excluded from the results of the root-focused comparisons. Radial plots from MEGAN were used to visually survey microbial populations for various treatments.

Taxonomic binning of contigs

In order to study the cellular function of mycorrhizal endobacteria, the contigs produced by assembly of the Roche 454 reads were filtered to retrieve a set of contigs for which bacterial origin could be confidently assigned, and these are referred to as "bacterial contigs". This was done by first aligning all of the taxonomically annotated Illumina reads to the contigs. STAR aligner was used, aligning sequences with at least 70% identity ("--outFilterMismatchNoverLmax 0.3") and providing at most 20 alignments for multimapped sequences ("--outFilterMultimapNmax 20"). With the taxonomically binned read alignments, contigs were selected which had only alignments from reads assigned to the bacterial bin by Blast2lca and none from reads assigned to the eukaryotic bin.

Contigs with a high likelihood of being of bacterial origin were also identified using direct taxonomic assignment with MEGAN. For this, contigs were aligned to NCBI-nr with DIAMOND, and a top percent filter of 15% ("--topPercent 15") was used. This provided a set of contigs for which assembled-kmer information could be included to enhance taxonomic assignment.

Functional enrichment

Contigs selected as being of bacterial origin were then functionally annotated using the software package Blast2GO (Conesa and Götz 2008, 2). This software package annotates sequences by (1) aligning reads to the NCBI nr database using blastx (Altschul et al. 1990), (2) mapping GO terms to hits in the database, and (3) selecting the high-confidence GO terms based on alignment similarity, GO term quality, and GO term graph structure. Additionally, the software performs an InterProScan (Jones et al. 2014) annotation to predict protein function.

Differential expression of contigs

The edgeR package (M. D. Robinson, McCarthy, and Smyth 2010) was used to compare bacterial gene expression between time points, tissues, and fungal symbioses. The STAR alignments of the reads to the contigs were used to quantify contig expression levels, and read counts were compiled using the eXpress (Roberts and Pachter 2012) tool, which applies an expectation-maximization algorithm to more accurately quantify RNA-Seq gene expression from short read sequencing. Effective read counts from eXpress were used to quantify gene expression for downstream analyses. However, raw read counts were used to the filter bacterial contigs expressed at lower levels in order to remove contigs whose expression could not be accurately quantified. This filter required that at least 5 reads be aligned in at least 3 samples for a contig to be retained for further analysis. Significantly expressed bacterial contigs meeting this criterion were then assessed for differential expression using edgeR. The effective counts were normalized using the TMM normalization (Mark D. Robinson and Oshlack 2010). Differential expression tests were conducted between root samples for each fungal symbiosis and the mock symbiosis, between root samples at 9 weeks and root samples at 5 weeks for all symbioses, and between root samples and shoot samples in all symbioses and time points. A false discovery rate of 0.1 was used to identify significantly differentially expressed bacterial contigs.

Results

Sequencing levels

Illumina HiSeq sequencing of the 48 samples collected yielded 1,039,663,930 total reads. After removing adapter sequences, low quality bases, and length filtering, 1,035,181,767 reads remained across all libraries. Sequencing runs had an error rate ranging from 0.42% to 0.51%, indicating an overall high level of quality. Single-end libraries from the shoot samples had a mean read count of 13,878,151, while paired-end libraries from root samples had a mean read-pair count of 7,254,049.

Table 1: Total high quality read counts for treatments

	Ggigantea	Gintraradices	Gversiforme	Mock
root, 5wk.	89,408,325	91,252,024	87,951,867	85,131,658
root, 9wk.	77,424,690	97,630,600	79,049,266	92,590,681
shoot, 5wk.	40,885,742	47,958,959	41,896,057	42,845,854
shoot, 9wk.	35,230,095	44,662,458	41,030,483	40,233,008
Total	242,948,852	281,504,041	249,927,673	260,801,201

Taxonomic classification of samples

Phylogenetic classification of the samples using the naïve approach of simple alignment to cpnDB and use of the top hit taxon resulted in a large number of spurious alignments identified from the taxa predicted. For example, a significant number of reads were aligned to the cpn60 gene assigned to the genus *Alligator*, large animals unlikely to be responsible for the contamination of these samples grown in a Cornell greenhouse. After refining these cpn60-based taxonomic predictions using MEGAN, the number of predicted taxa reduced significantly, and a large portion of the most suspect taxa were removed. However, due to the relatively low number of alignments to cpnDB (Table 2), as well as the more limited coverage of the tree of life in

cpnDB compared to other sequence repositories (Table 3), the cpn60 alignment did not allow adequate characterization of the bacterial population in the samples.

Table 2: Number of reads binned to relevant upper clades

	Viridiplantae	Fungi	Bacteria
cpnDB	75,255	941	8
NCBI-nr	395,272,860	10,851,671	386,410

Table 3: Subtree coverage for relevant upper clades

	Viridiplantae	Fungi	Bacteria
Subtaxa in cpnDB	286	912	17,647
Subtaxa in NCBI-nr	99,935	24,607	91,171
Protein sequences in NCBI-nr	6,954,379	8,974,862	235,859,143

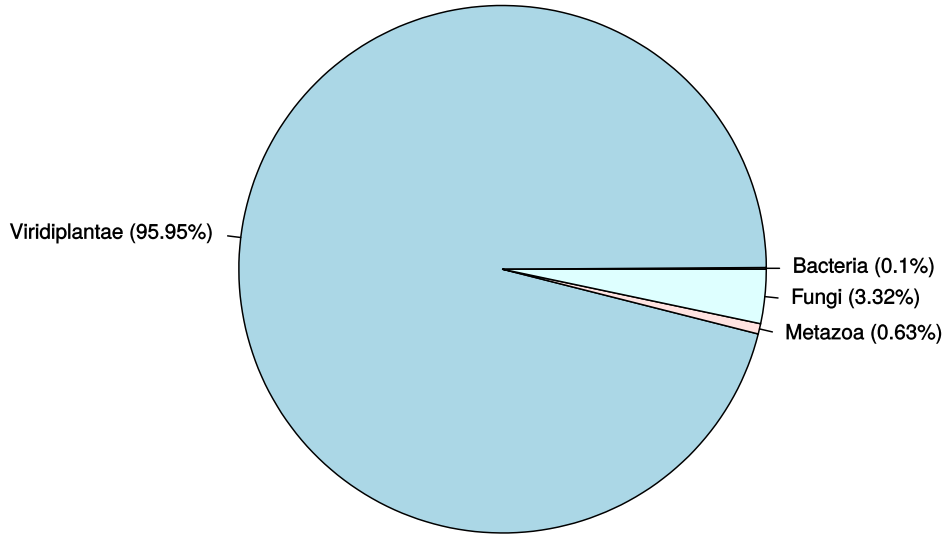
To obtain a more informative survey of the taxonomic makeup of the microbial population in the root samples, taxonomic characterization using the comprehensive NCBI nr database was conducted. This showed much greater sensitivity than the cpn60-based characterization, with bacterial taxa containing a significant number of assigned reads (Table 2). Due to the higher sensitivity, a minimum threshold filter was used to remove low abundance taxa which removed more than 80% of taxa identified by MEGAN (Table 4). Removal of redundant, non-informative taxa further reduced the taxa count by 4.5%.

Table 4: Taxa counts after filtering

Filter	Viridiplantae	Fungi	Bacteria	Total	Percent of total taxa
All predicted taxa	9,957	4,691	11,281	39,527	-
Minimally abundant taxa	2,170	1,343	985	6,780	17.15
Minimally abundant, informative taxa	1,752	881	751	5,016	12.69

The non-redundant, significantly abundant taxa remaining provide a census of the transcriptionally active species present in the intracellular mycorrhizal environment. It is difficult to make direct conclusions about the relative proportions of individual taxa from a survey that strictly considers RNA transcripts, as individual cells may have differing total RNA productivity under different conditions (Traganos, Darzynkiewicz, and Melamed 1982) and different clades can vary considerably in their total RNA production per cell. However, such a survey does provide an informative, although qualitative, view of the overall taxonomic makeup of an environment. The existence of a significant number of reads assigned by the LCA algorithm to a taxon is often strong evidence that the taxon is present and transcriptionally active in the sample environment. Therefore, observation of these high-confidence taxa gives an approximate census of the members present in the mycorrhizal community. In the root samples which were inoculated with mycorrhizal fungi in this study, plant taxa contained the overwhelming majority of short read assignments (Figure 1). Expectedly, fungal taxa were assigned the next greatest number of reads, while Metazoan taxa contained the fewest predicted read assignments of the eukaryotic kingdoms. Bacterial taxa comprised a lower fraction of the assigned reads than any eukaryotic kingdom. The relative reduction in transcript abundance between host and endosymbiont for both the plant-fungal and fungal-bacterial symbioses was approximately 30 fold.

Figure 1. Taxonomic distribution of reads binned to Eukaryotic kingdoms and Bacteria in fungal-inoculated root samples



In addition to qualitatively surveying the taxa present in the mycorrhiza, a quantitative comparison of transcriptome abundance levels between various taxa across samples can be made with robust taxonomic read assignments. Although transcriptome abundance does not directly correlate with organismal abundance as genome abundance does, it does measure the relative transcriptomic activity of the various clades within an environment. Because proteins translated from transcripts are responsible for the biomolecular activity in an environment including structural and metabolic processes, an organism's transcriptome can have significant effects on the environment and other organisms present. For this reason, it is still of interest to compare the transcriptome abundances of the organisms in an environment across different conditions, and

the abundances of taxonomically binned reads provide a good measure of these transcriptome abundances.

In the root samples studied, several bacterial taxa showed elevated transcriptome levels, as measured by binned-read counts, in samples treated with arbuscular mycorrhizal fungi compared to those that were not. Of the 751 minimally abundant, non-redundant bacterial taxa, 649 did not have significantly higher mean binned read abundances in shoots at an FDR of 0.05, and these taxa were considered for differential transcriptome abundance comparisons in the root samples. Student's t-tests comparing the mean CPM of all inoculated root samples to that of all control root samples showed that 190 (29.3%) of the 649 taxa had significantly differentially abundant transcriptomes at a FDR of 0.05. All of these taxa had higher numbers of reads assigned to them in the fungal inoculated samples than in the control samples. The twenty taxa with the most significant differences in mean binned read counts in inoculated roots is presented in Table 5.

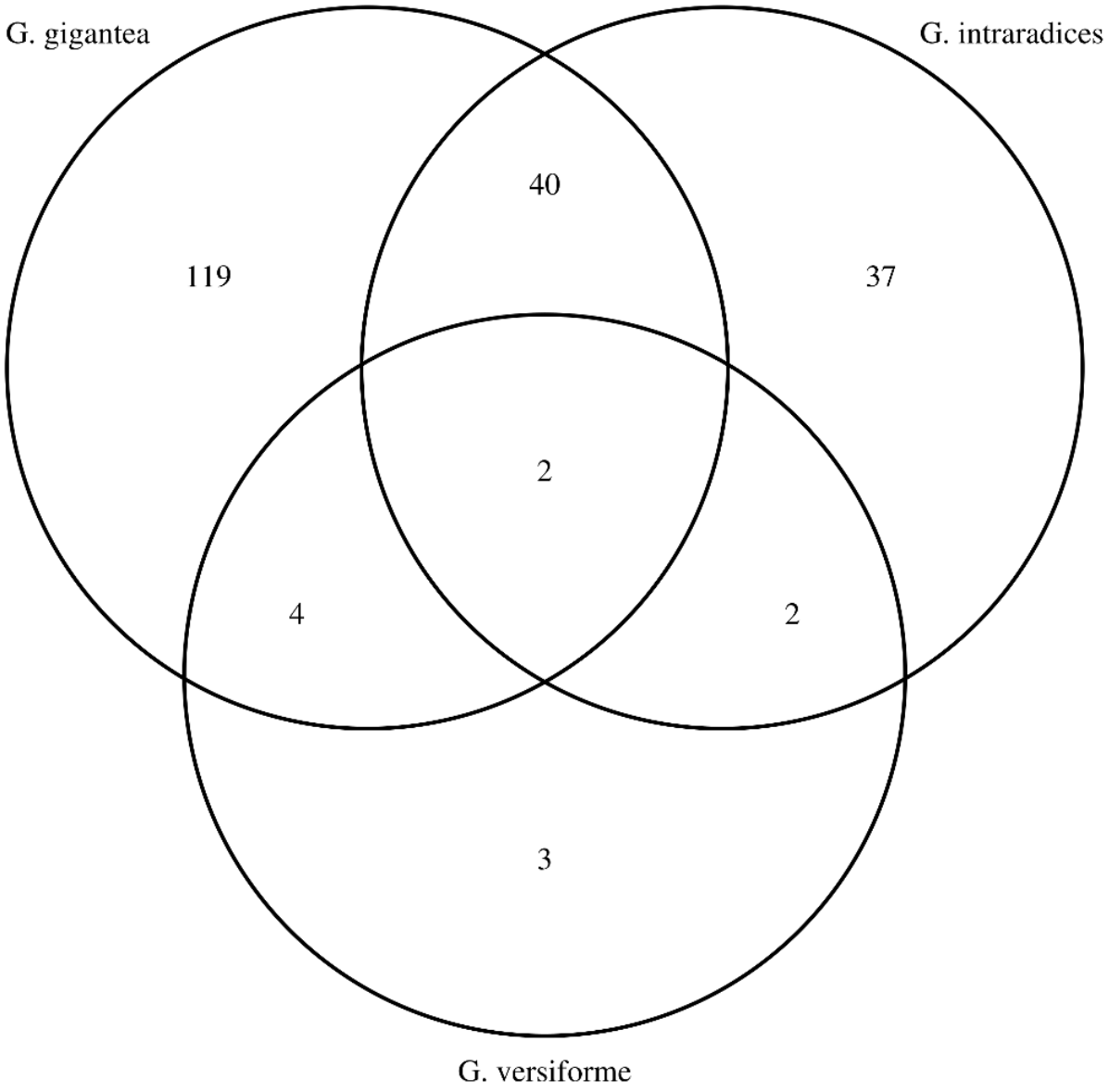
Table 5: Taxa with most significantly binned-read abundance differences between inoculated and control roots

	Taxon	Class	FDR	Difference in mean CPM	Mean CPM in roots
1	[Species+] Dyadobacter fermentans DSM 18053	Cytophagia	3.00×10^{-6}	0.77	0.61
2	[Genus] Gillisia	Flavobacteriia	4.04×10^{-5}	0.59	0.64
3	[Genus] Cellulophaga	Flavobacteriia	5.35×10^{-5}	0.58	0.49
4	[Species] Arenibacter algicola	Flavobacteriia	5.84×10^{-5}	0.80	0.60
5	[Family] Cytophagaceae	Cytophagia	6.10×10^{-5}	6.12	6.74
6	[Order] Cytophagales	Cytophagia	6.70×10^{-5}	7.97	8.54
7	[Species] Spirosoma linguale	Cytophagia	7.80×10^{-5}	0.97	1.06
8	[Genus] Dyadobacter	Cytophagia	9.79×10^{-5}	1.43	1.78
9	[Genus] Sphingobacterium	Sphingobacteriia	9.85×10^{-5}	2.43	1.91
10	[Species] Nonlabens sediminis	Flavobacteriia	9.97×10^{-5}	0.42	0.32
11	[Genus] Dickeya	Gammaproteobacteria	1.39×10^{-4}	1.08	0.92
12	[Species+] Enterococcus casseliflavus 14-MB-W-14	Bacilli	1.39×10^{-4}	0.61	0.46
13	[Family] Sphingobacteriaceae	Sphingobacteriia	1.45×10^{-4}	3.43	2.88
14	[Species] Rhodopseudomonas palustris	Alphaproteobacteria	1.56×10^{-4}	0.72	0.78
15	[Family] Cyclobacteriaceae	Cytophagia	1.65×10^{-4}	1.60	1.49
16	[Class+] unclassified Alphaproteobacteria	Alphaproteobacteria	2.23×10^{-4}	0.84	0.90
17	[Genus] Marinobacter	Gammaproteobacteria	3.45×10^{-4}	1.50	1.56
18	[Family] Anaplasmataceae	Alphaproteobacteria	3.57×10^{-4}	0.50	0.61
19	[Species+] Bacillus cereus m1293	Bacilli	3.89×10^{-4}	3.33	2.53
20	[Class++] OMG group	Gammaproteobacteria	4.66×10^{-4}	0.47	0.37

There were several clades that were heavily represented in the taxa with the most elevated transcriptome abundance in the roots. The *Bacteroidetes* phylum contains a large portion of these taxa, with the children classes *Cytophagia*, *Flavobacteriia*, and *Sphingobacteriia* containing all of the ten most significantly different taxa. Several *Proteobacteria* clades were also among these taxa with transcriptome abundances elevated in roots, including the *Dickeya* genus, which contains several plant pathogens *Rhodopseudomonas palustris*, a ubiquitous microbe notable for its diverse metabolic capabilities, *Marinobacter*, a sea water-based genus able to degrade hydrocarbons, and *Bacillus cereus*, a common food-borne pathogen in humans which is also associated with the rhizosphere of certain plants (Halverson, Clayton, and Handelsman 1993).

To determine which bacterial taxa exhibit different transcriptome abundances in the mycorrhiza of different fungal symbionts, comparisons were made of the read counts of taxa between the individual fungal treatments and the mock non-inoculated treatment. Taxa which showed increased transcriptome abundances in only a subset of the fungal symbioses were then examined further, as these taxa may show selective transcriptomic responses or colonization in certain mycorrhizal fungal symbioses. Overall, *G. gigantea* symbioses showed the greatest number of unique transcriptomically over-abundant taxa, as well as an overall larger number of taxa with over-abundant transcriptomes, while *G. versiforme* showed the fewest in both categories (Figure 2)

Figure 2. Fungal treatment-specific counts of bacterial taxa with significant differential transcriptome abundances in inoculated roots relative to control roots



The full results of the comparisons between bacterial transcriptome abundances in the various fungal inoculations with those in the control treatment for all bacterial taxa, excluding those with a higher transcriptome abundance in shoots, are provided in Appendix C, and the results for a selection of these tests is shown in Table 6. The *G. gigantea* symbiosis was uniquely enriched in transcriptome sequences for the *Actinobacteria* genera *Mycobacterium* and *Amycolatopsis*, the *Proteobacteria* families *Methylobacteriaceae, Bradyrhizobiaceae, Rhodobacteraceae*, and *Xanthomonadaceae*, and the *Proteobacteria* genera *Citrobacter* and *Pseudomonas*. Root samples inoculated with *G. intraradices* were uniquely enriched for transcriptomes of the *Bacteroidales* order, *Streptococcus mutans* and the *Lachnoclostridium* genus from the *Firmicutes* phylum, the *Neisseriaceae* family from the *Proteobacteria* phylum, and the *Spirochaetales* order.

Table 6: Significance of comparisons of taxonomically-binned read counts for fungal-inoculated samples against non-inoculated samples in roots. Red cells indicate increasingly significant (lower) t-test FDR values, with color scaling beginning at FDR 0.05. Green cells indicate increasing mean abundances (mean CPM) of reads binned to the taxon in root samples. (A) Taxa with significantly higher mean binned-read counts in the G. gigantea symbiosis, but not in the G. intraradices or G. versiforme symbioses, compared to non-inoculated roots. (B) Taxa with significantly higher mean binned-read counts in only the G. intraradices symbiosis compared against non-inoculated roots. (C) Taxa with significantly higher mean binned-read counts in both the G. gigantea and G. intraradices mycorrhiza, but not in the G. versiforme symbiosis, measured against non-inoculated roots.

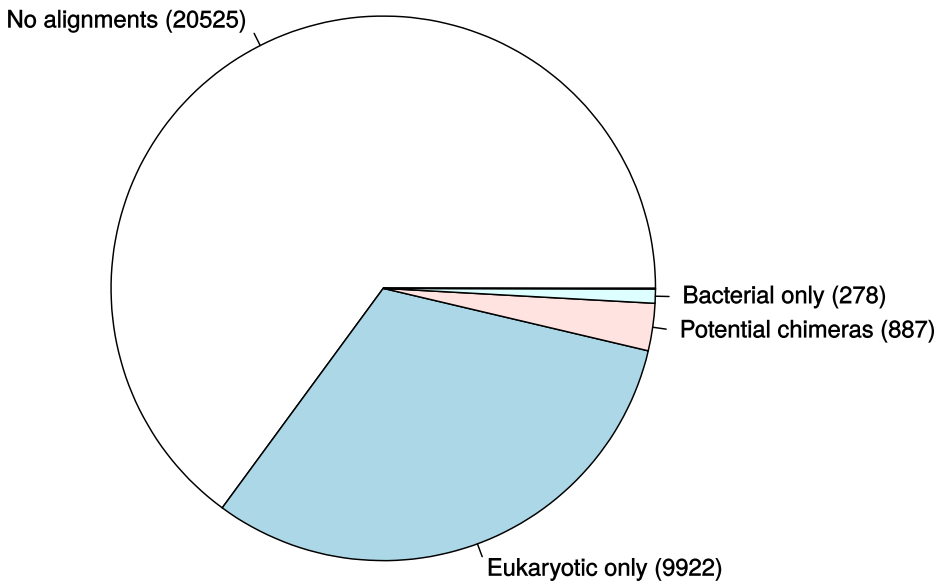
Taxon	Phylum	Class	Order	Average CPM in root samples	FDR (all inoculations)	FDR (G. gigantea)	FDR (G. intraradices)	FDR (G. versiforme)	
[Genus] Mycobacterium	Actinobacteria	Actinobacteria	Corynebacteriales	3.23	6.81E-02	3.52E-02	4.04E-01	4.57E-01	A
[Family] Methylobacteriaceae	Proteobacteria	Alphaproteobacteria	Rhizobiales	3.92	4.16E-02	1.59E-02	2.97E-01	3.48E-01	A
[Genus] Mesorhizobium	Proteobacteria	Alphaproteobacteria	Rhizobiales	3.39	5.26E-02	3.90E-02	8.08E-01	2.15E-01	A
[Family] Rhodobacteraceae	Proteobacteria	Alphaproteobacteria	Rhodobacterales	6.22	2.84E-02	4.15E-02	5.66E-01	1.97E-01	A
[Species] Citrobacter farmeri	Proteobacteria	Gammaproteobacteria	Enterobacteriales	17.96	2.14E-01	4.56E-02	8.38E-01	4.64E-01	A
[Genus] Pseudomonas	Proteobacteria	Gammaproteobacteria	Pseudomonadales	6.35	1.54E-02	2.29E-02	4.81E-01	3.12E-01	A
[Order] Bacteroidales	Bacteroidetes	Bacteroidia	Bacteroidales	12.20	5.45E-02	1.79E-01	1.28E-02	4.18E-01	B
[Genus] Lachnoclostridium	Firmicutes	Clostridia	Clostridiales	8.09	2.65E-01	1.19E-01	2.42E-02	5.51E-01	B
[Species] Streptococcus mutans	Firmicutes	Bacilli	Lactobacillales	2.51	1.73E-01	9.29E-01	7.80E-03	4.15E-01	B
[Family] Neisseriaceae	Proteobacteria	Betaproteobacteria	Neisseriales	4.42	1.54E-01	8.80E-01	3.91E-04	3.13E-01	B
[Order] Spirochaetales	Spirochaetes	Spirochaetia	Spirochaetales	2.26	1.25E-03	1.24E-01	1.75E-03	1.16E-01	B
[Family] Cytophagaceae	Bacteroidetes	Cytophagia	Cytophagales	6.74	6.10E-05	2.79E-02	6.13E-03	8.88E-02	C
[Family] Flavobacteriaceae	Bacteroidetes	Flavobacteriia	Flavobacteriales	14.64	6.16E-04	2.61E-02	2.46E-03	1.55E-01	C
[Family] Sphingobacteriaceae	Bacteroidetes	Sphingobacteriia	Sphingobacteriales	2.88	1.45E-04	2.84E-02	5.02E-03	6.86E-02	C
[Species] Bacillus cereus	Firmicutes	Bacilli	Bacillales	4.20	2.27E-03	2.14E-02	7.91E-03	3.89E-01	C

Taxonomic binning of the assembled contigs

Isolation of bacterial contigs using taxonomically-binned reads

The assembly of Roche 454 reads contained 31,612 contigs after the removal of *B. distachyon* contigs. Reads from side 1 of the paired-end Illumina libraries were used to separate these taxonomically and perform differential expression comparisons. After selecting only those contigs which had bacterial read alignments but no eukaryotic read alignments, 278 contigs (0.88%) remained (Figure 3). Of these, 10 contigs met a minimum abundance threshold of 1 CPM for bacterial read mappings in at least 3 libraries. 9,922 (31.4%) of the contigs had alignments to only eukaryotic reads, and 2,850 of these met the minimum abundance threshold. 887 (2.81%) contigs had both bacterial and eukaryotic reads align, while 20,525 (64.9%) had no alignments from bacterial or eukaryotic side 1 reads with the STAR parameters used.

Figure 3. Superkingdom distribution of contigs with taxonomically-binned read alignments

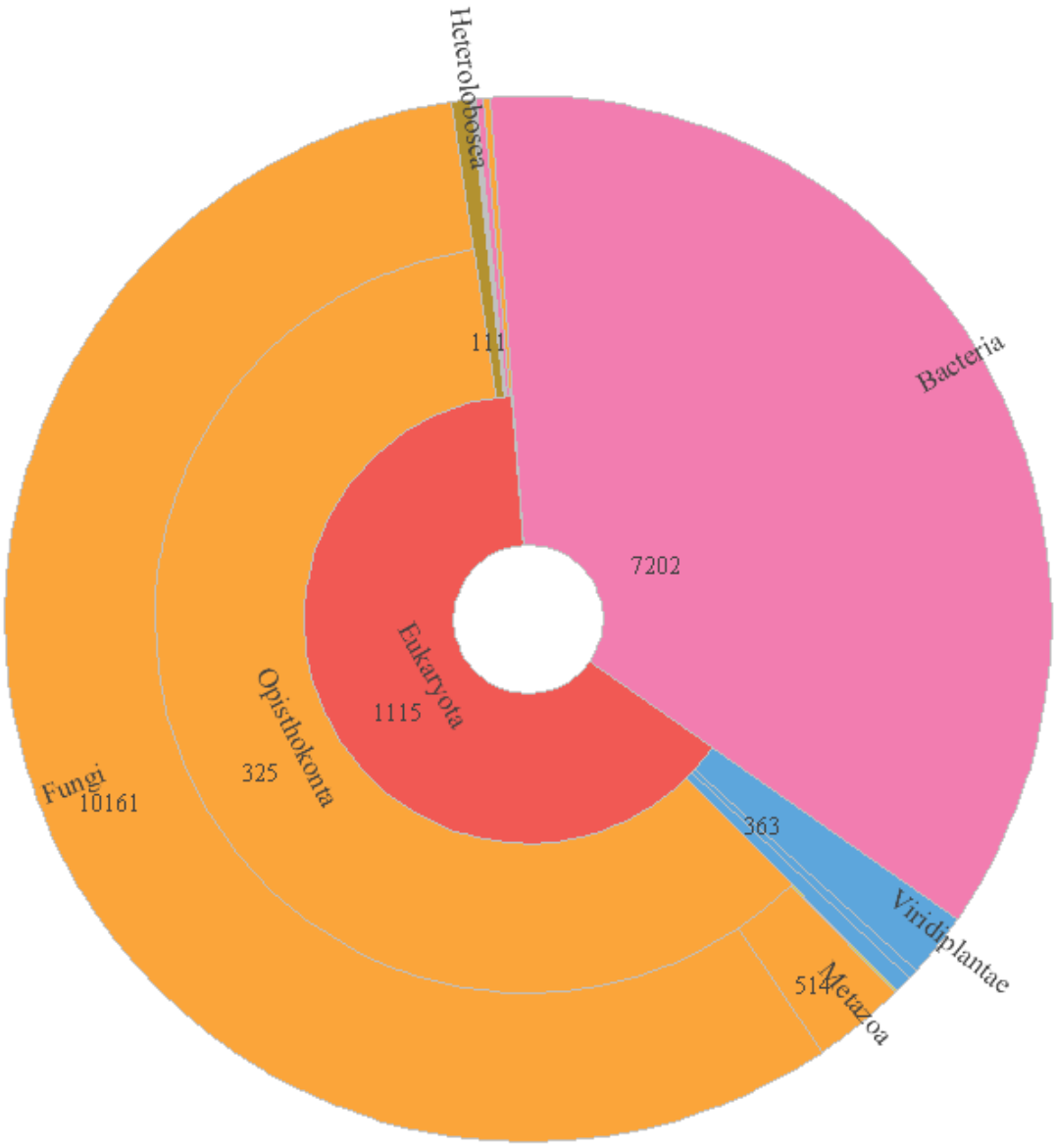


Isolation of bacterial contigs using MEGAN

An alternative method for selecting the contigs likely to be of bacterial origin used MEGANs LCA algorithm to bin the contigs themselves. This resulted in 20,406 of the 31,612 original contigs being labeled as either eukaryotic or bacterial in origin, with 7,202 (35.3%) of these labeled as bacterial and 12,867 (63.1%) labeled as eukaryotic. Figure 4 shows the

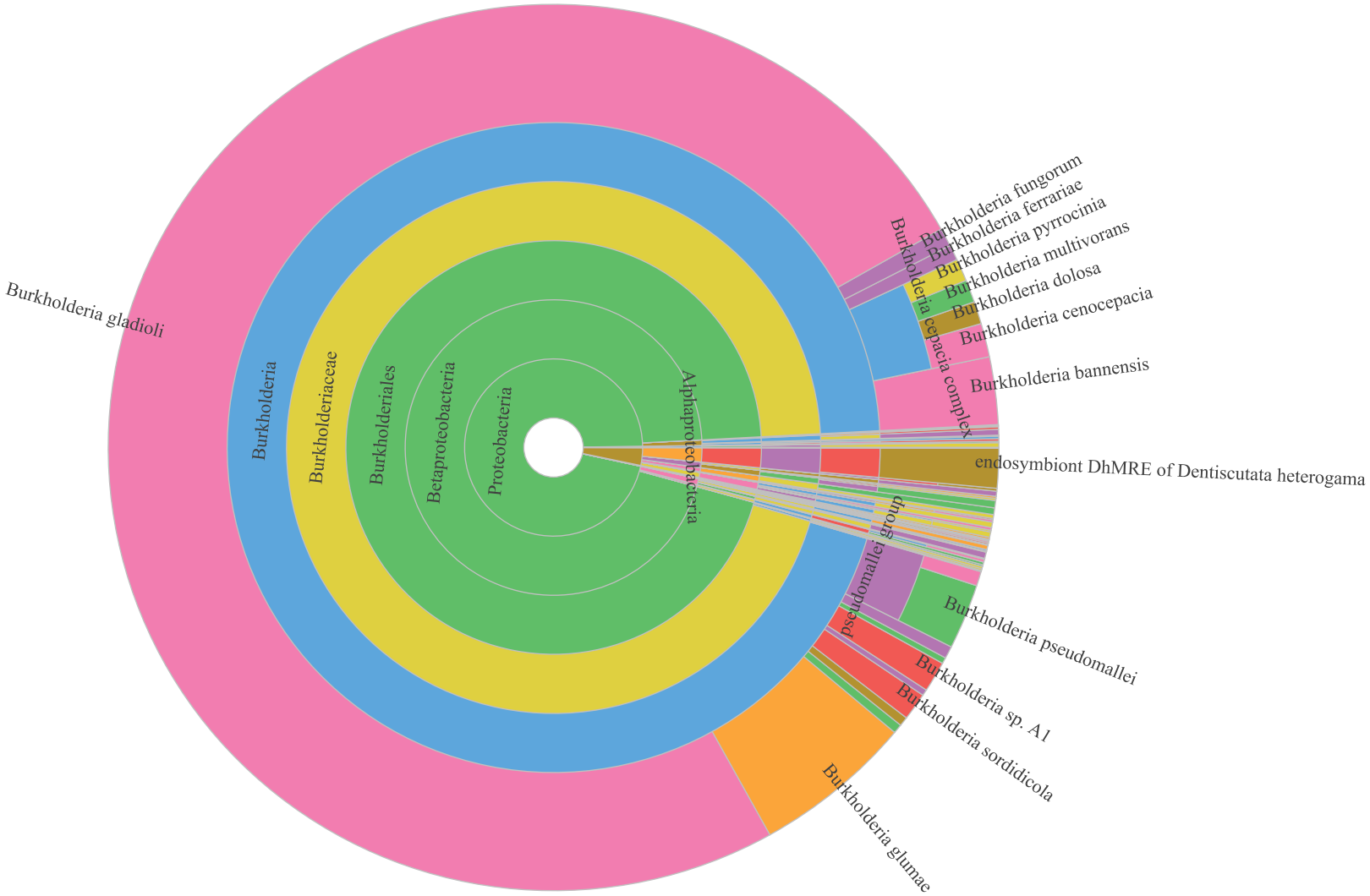
taxonomic distribution of the contigs labeled as either bacterial or eukaryotic to the kingdom level. Of the 7,202 bacterial contigs, 41 had a minimum abundance of 1 CPM among side 1 reads which had been annotated as bacterial in at least 3 root samples.

Figure 4. Taxonomic assignments of all contigs to kingdom level, after filtering *Brachypodium* contigs



Among the contigs predicted to be bacterial, an overwhelming majority (6,288; 87.3%) were assigned at or below the *Burkholderia* genus (Figure 5). Of these, *Burkholderia gladioli* was the most highly represented species (670; 9.3%). Interestingly, the only taxonomic unit outside of the *Burkholderia* clade with significant representation among the bacterial contigs was the recently discovered *Mollicutes*-related endobacteria (MRE), which is shown to be an obligate symbiont of a majority of arbuscular mycorrhizal fungi. Although it contained significantly fewer taxonomically annotated contigs than *B. gladioli*, the *Dentiscutata heterogama* MRE (*Dh*MRE) clade included 83 contigs, representing over 1% of the annotated bacterial contigs. Between the *Burkholderiaceae* family (containing *Ca. G. gigasporarum*) and the *Dh*MRE clades, both of the two bacterial groups known to participate in obligate endosymbioses with AM fungi are represented among the contigs assembled (Valeria Bianciotto et al. 2003; Naumann, Schüßler, and Bonfante 2010). Blast2GO gene ontology maps for the *Dh*MRE contigs are provided in Appendix A (molecular function) and Appendix B (biological process).

Figure 5. Full taxonomic assignments of contigs predicted to be bacterial



Concordance between two methods for isolating bacterial contigs

Overall, identifying contigs likely to be of bacterial origin by direct annotation using MEGAN proved to be much more sensitive than selecting those contigs with only bacterially-annotated short reads aligning. The two methods did agree on their predictions for most of the bacterial contigs predicted by compiling bacterial read-counts, and 228 of the 278 contigs identified by this method were also annotated as bacterial by MEGAN. Fifty contigs were not binned as bacterial by MEGAN, but contained reads annotated as bacterial and none as eukaryotic. Of the 10 contigs identified using the bacterial read-count method and meeting the minimum abundance threshold among bacterial reads, all were also annotated as bacterial in MEGAN. For this reason, differential abundance tests comparing contig abundance at various treatments was performed once, using the more inclusive set of minimally-abundant contigs predicted by MEGAN. Contigs that were predicted to be bacterial by both MEGANs LCA algorithm as well as by having alignments strictly to bacterially-binned reads are marked in Table 7 with bold contig names.

Differential abundance of contigs annotated as bacterial across treatments

Comparison of roots inoculated with mycorrhizal fungi with control roots

Roots which were inoculated with fungal symbionts showed enrichment for several of the bacterial contigs annotated by MEGAN. All fungal symbioses showed significantly more abundant RNA of a contig annotated as a type I glyceraldehyde-3-phosphate dehydrogenase in Blast2GO for at least one time point when compared to the non-inoculated samples. Additionally, a serine acetyltransferase was overexpressed in G. versiforme roots at both five and nine weeks post-inoculation (WPI). In G. gigantea, a carbonic anhydrase, an aminocarboxypropyltransferase, an alpha galactosidase, a glycosyl hydrolase, and an alpha mannosidase were more abundant in roots at both five and nine WPI. When all mycorrhizal symbioses were jointly compared against mock-inoculated roots, a cell wall associated hydrolase showed significantly increased abundance only at week nine.

Table 7: Summary of differential abundance tests done on contigs predicted to be bacterial by MEGAN. Bold contig names indicate the absence of eukaryotic read alignments to the contig.

	Significant differential expression in mycorrhizal vs. control roots	**Significant differential expression at 9 weeks vs. 5 weeks-post-inoculation**	**Phylum predicted**	**Most specific taxonomic prediction**	**Description**	**GO Names list**	**InterPro IDs**
Contig74221	↑G. intraradices (Week 5), ↑G. intraradices (Week 9), ↑G. versiforme (Week 9), ↑G. gigantea (Week 5), ↑G. gigantea (Week 9), ↑All symbioses (Week 9)	↑G. intraradices, ↑G. versiforme, ↑G. gigantea			type I glyceraldehyde-3-phosphate dehydrogenase		G3DSA:3.30.360.10 (GENE3D); IPR020829 (PFAM); SSF55347 (SUPERFAMILY)
Contig55250		↓G. gigantea	Proteobacteria	Betaproteobacteria (Class)	cell wall-associated hydrolase	P:metabolic process; F:hydrolase activity	
Contig82905			Proteobacteria	Vibrio (Genus)	dehydration responsive partial		
Contig95292			Proteobacteria	Magnetospirillum gryphiswaldense	conserved hypothetical protein		
Contig86041	↓G. gigantea (Week 5)			Terrabacteria group	AE001886_6hypothetical protein DR_0254		
Contig89600		↓Control	Proteobacteria	Rickettsiales (Order)	hypothetical protein mgI388	C:membrane; C:integral component of membrane	SIGNAL_PEPTIDE_H_REGION (PHOBIUS); NON_CYTOPLASMIC_DOMAIN (PHOBIUS); SIGNAL_PEPTIDE_N_REGION (PHOBIUS); SIGNAL_PEPTIDE (PHOBIUS); SIGNAL_PEPTIDE_C_REGION (PHOBIUS)
Contig91948		↓G. versiforme, ↓G. gigantea	Proteobacteria	Magnetospirillum gryphiswaldense	conserved hypothetical protein		

Table 7 (cont.)

Contig75991		↓G. versiforme	Proteobacteria	Xanthomonas citri	dehydration responsive partial		SIGNAL_PEPTIDE_C_REGION (PHOBIUS); SIGNAL_PEPTIDE_H_REGION (PHOBIUS); SIGNAL_PEPTIDE (PHOBIUS); NON_CYTOPLASMIC_DOMAIN (PHOBIUS); SIGNAL_PEPTIDE_N_REGION (PHOBIUS)
Contig69753		↓G. gigantea	Actinobacteria	Streptomyces (Genus)	leucine rich	C:membrane; C:integral component of membrane	
Contig64303			Firmicutes		hypothetical CTC00065		
Contig88913	↑G. versiforme (Week 9)	↑G. intraradices, ↑G. versiforme			Cell wall-associated hydrolase	P:metabolic process; F:hydrolase activity	
Contig8507	↑G. gigantea (Week 5), ↑G. gigantea (Week 9)	↓G. gigantea			carbonic anhydrase	F:zinc ion binding; P:metabolic process; F:carbonate dehydratase activity; P:carbon utilization	Coil (COILS); IPR001765 (SMART); IPR001765 (PFAM); IPR001765 (G3DSA:3.40.1050.GENE3D); PTHR11002:SF20 (PANTHER); IPR001765 (PANTHER); IPR015892 (PROSITE_PATTERNS); IPR015892 (PROSITE_PATTERNS); IPR001765 (SUPERFAMILY)
Contig25689			Proteobacteria	Burkholderiales (Order)	hypothetical protein SRAA_2358		
Contig70961					Cell wall-associated hydrolase	P:metabolic process; F:hydrolase activity	
Contig93335		↑G. intraradices	Proteobacteria	Magnetospirillum gryphiswaldense	conserved hypothetical protein	C:membrane; C:integral component of membrane	

Table 7 (cont.)

Contig91037	↑G. intraradices (Week 9), ↑G. gigantea (Week 9)	↓Control			hypothetical protein AUK64_2547		
Contig91401			Proteobacteria	Alphaproteobacteria (Class)	unnamed protein product, partial		
Contig73374		↑G. gigantea	Proteobacteria		ORF16-lacZ fusion partial		PTHR34890:SF2 (PANTHER); PTHR34890 (PANTHER)
Contig76612			Proteobacteria		ORF16-lacZ fusion partial		
Contig79960					Conserved		
Contig81971	↑G. versiforme (Week 9)	↑G. versiforme	Proteobacteria		hypothetical protein H845_49		
Contig60224			Proteobacteria	Alphaproteobacteria (Class)	hypothetical protein		PTHR34890:SF1 (PANTHER); PTHR34890 (PANTHER)
Contig93341		↓G. intraradices, ↓G. versiforme, ↓G. gigantea	Proteobacteria	Betaproteobacteria (Class)	dehydration responsive partial		SIGNAL_PEPTIDE_N_REGION (PHOBIUS); SIGNAL_PEPTIDE_C_REGION (PHOBIUS); NON_CYTOPLASMIC_DOMAIN (PHOBIUS); SIGNAL_PEPTIDE (PHOBIUS); SIGNAL_PEPTIDE_H_REGION (PHOBIUS)
Contig87270			Proteobacteria	agnetospirillum gryphiswalder	secreted {ECO:0000313	C:membrane	
Contig39090	↓G. gigantea (Week 9)	↓G. gigantea	Firmicutes	Clostridium (Genus)	hypothetical CTC00065		
Contig94383			Proteobacteria		serine acetyltransferase		
Contig92863			Proteobacteria	Acinetobacter towneri	hypothetical protein F947_00014, partial	C:membrane	TRANSMEMBRANE (PHOBIUS); NON_CYTOPLASMIC_DOMAIN (PHOBIUS); CYTOPLASMIC_DOMAIN (PHOBIUS)
Contig80641		↑G. versiforme			cell wall-associated hydrolase domain	P:metabolic process; F:hydrolase activity	TRANSMEMBRANE (PHOBIUS); CYTOPLASMIC_DOMAIN (PHOBIUS); NON_CYTOPLASMIC_DOMAIN (PHOBIUS); TMhelix (TMHMM)

Table 7 (cont.)

Contig24059			Proteobacteria	Rhizobiales (Order)	conserved hypothetical protein		
Contig90510			Proteobacteria		secreted {ECO:0000313	C:membrane	
Contig60189	↑G. versiforme (Week 9), ↑All symbioses (Week 9)	↑G. intraradices, ↑G. versiforme, ↑G. gigantea	Proteobacteria		cell wall-associated hydrolase	P:metabolic process; F:hydrolase activity	
Contig81177			Proteobacteria	Escherichia coli	3,4-dihydroxy-2-butanone-4-phosphate synthase	F:catalytic activity	TRANSMEMBRANE (PHOBIUS); NON_CYTOPLASMIC_DOMAIN (PHOBIUS); CYTOPLASMIC_DOMAIN (PHOBIUS)
Contig84762					cell wall-associated hydrolase	P:metabolic process; F:hydrolase activity	TRANSMEMBRANE (PHOBIUS); NON_CYTOPLASMIC_DOMAIN (PHOBIUS); CYTOPLASMIC_DOMAIN (PHOBIUS); TRANSMEMBRANE (PHOBIUS); CYTOPLASMIC_DOMAIN (PHOBIUS)
Contig75353			Proteobacteria	Pararhodospirillum photometricum	cell wall-associated hydrolase	P:metabolic process; F:transferase activity	
Contig84292	↓G. intraradices (Week 9)	↓G. intraradices			PG1 homology to Homo sapiens		PTHR38146 (PANTHER)
Contig36978	↑G. versiforme (Week 9), ↑G. gigantea (Week 5), ↑G. gigantea (Week 9)	↑G. versiforme, ↑Control			O-acetylhomoserine aminocarboxypropyltransferase	F:pyridoxal phosphate binding; P:'de novo' L-methionine biosynthetic process; P:cysteine biosynthetic process; F:cysteine synthase activity; F:O-acetylhomoserine aminocarboxypropyltransferase activity; F:hydrolase activity	IPR000277 (PFAM); IPR000277 (PIRSF); IPR015422 (G3DSA:3.90.1150.GENE3D); IPR015421 (G3DSA:3.40.640.GENE3D); IPR006235 (TIGRFAM); IPR000277 (PANTHER); IPR006235 (PTHR11808:PANTHER); IPR006235 (PTHR11808:PANTHER); IPR000277 (PANTHER); IPR015424 (SUPERFAMILY)

Of the minimally abundant bacterial contigs that were differentially abundant between treatments in this comparison, two were placed into the *Proteobacteria* phylum by MEGAN, two were assigned to the *Firmicutes* phylum, including a contig annotated to the genus *Clostridium*, and one was assigned to the *Actinobacteria* class. Another contig differentially abundant in one of these comparisons was annotated as belonging to the *Terrabacteria* group, which includes the *Firmicutes* and *Actinobacteria* phyla, among others. The majority of these contigs showed significantly higher abundance in the inoculated roots relative to the non-inoculated hosts, but four of these contigs showed reduced abundance relative to control. Two annotated only as *Bacteria* had reduced abundance in *G. intraradices* mycorrhiza compared to control roots at week nine. Interestingly, one of these contigs also showed significantly *increased* abundance in *G. gigantea* roots at week five relative to control. Another contig annotated as belonging to the *Terrabacteria* group showed reduced abundance in *G. gigantea* treated roots at week five, and the contig assigned as *Clostridium* showed reduced abundance in *G. gigantea* roots at week nine.

Mycorrhiza across time points

To observe changes in the mycorrhizal microbiome through the progression of the plant life cycle, fungal inoculated root metatranscriptomes were compared at five and nine weeks after inoculation. In all three symbioses studied, a type I glyceraldehyde-3-phosphate dehydrogenase and one cell wall-associated hydrolase were present at higher levels in roots at nine WPI. Additionally in all three symbioses, a gene that highly conserved among many bacterial species, including several *Burkholderia* species, and involved in dehydration response had significantly lower abundance in roots at nine WPI compared to roots at five WPI. Another cell wall-associated hydrolase had increased abundance at the later time point in both the *G. intraradices* and *G. versiforme* mycorrhiza, while a third cell wall-associated hydrolase showed increased

abundance at the later time point in *G. versiforme* and *G. gigantea* treatments. Another contig with similarity to a dehydration responsive protein had reduced abundance at nine weeks in only *G. versiforme*. Contigs annotated for a type I glyceraldehyde-3-phosphate dehydrogenase, a serine acetyltransferase, and an O-acetylhomoserine aminocarboxypropyltransferase were more abundant only in the *G. versiforme* symbiosis at 9 weeks. Of these, however, both the type I glyceraldehyde-3-phosphate dehydrogenase and the O-acetylhomoserine aminocarboxypropyltransferase showed the same significant differential abundance pattern in the non-inoculated roots. Contigs with differential abundance in only the *G. gigantea* symbiosis include a cell wall-associated hydrolase, a leucine-rich protein, a carbonic anhydrase, an alpha-galactosidase, an O-glycosyl hydrolase, and an alpha-mannosidase, all with decreased abundance at the later time point.

Contigs with more abundant RNA at nine WPI relative to five WPI which had taxonomic assignments specific below the superkingdom level were all annotated as belonging to the phylum *Proteobacteria*. One of these contigs was assigned more specifically to the species *Magnetospirillum gryphiswaldense*, a freshwater sediment-inhabiting Gram-negative bacteria with the unusual ability to orient along Earth's magnetic axis.

The contigs with reduced abundance at nine WPI compared to five WPI represented taxa from the *Firmicutes* (2 contigs), *Actinobacteria* (2 contigs), and *Proteobacteria* (4 contigs) phyla. In *G. gigantea* inoculated roots, contigs with significantly reduced abundance at the later time point included one contig annotated as belonging to *Clostridium*, one annotated as *Streptomyces*, one annotated at the class level as a *Betaproteobacteria*, another annotated at the class level as *Actinobacteria*, and one annotated at the phylum level as *Firmicutes*. In *G. versiforme* inoculated roots, a contig predicted to have been produced by *Xanthomonas citri*, a

common pathogen of both citrus and cotton plants, showed reduced RNA abundance at nine weeks. In all three symbioses, a contig annotated to the *Betaproteobacteria* class had decreased abundance at nine weeks. Interestingly, one of the contigs present at lower levels in both *G. versiforme* and *G. gigantea* at nine weeks was placed into the *Magnetospirillum gryphiswaldense* by MEGAN, in contrast to another contig annotated as *Magnetospirillum gryphiswaldense* but more abundant at nine weeks in the *G. intraradices* symbiosis.

Discussion

Taxonomic characterization of a metatranscriptomic sample

A natural first step in any meta-omics survey is to determine which operational taxonomic units (OTUs) are present in the samples and what the overall population structure of the community is. Many methods exist for this task, and the most common generally involve the use of a highly conserved taxonomic marker gene to survey taxa and estimate genome abundances. While 16S rRNA genes are commonly used for this task in metagenomic studies, where both coding and non-coding sequences are represented in the sequenced material, metatranscriptomic surveys require the use of polyadenylated coding sequences since current mRNA sequencing protocols use an oligo-dT purification step. Numerous transcribed candidate genes are available for this purpose, which are both extensively conserved yet still variable enough to be able to differentiate between closely related taxa. One transcribed gene useful for this task is the chaperonin 60 (cpn60) gene, which has been found to exist in virtually all bacteria and eukaryotes and a number of archaeal species (Hill 2004). Indeed, for closely related prokaryotic populations, the use of the cpn60 gene as a taxonomic marker may provide greater power than the conventional 16S rRNA markers even when those genes are available to use, since cpn60 is significantly less conserved than 16S rRNA genes (Brousseau et al. 2001; Lan, Rosen, and Hershberg 2016).

Unfortunately, the use of single marker genes for taxonomic characterization proved to be difficult in this study for a few reasons. Firstly, the sheer volume of sequencing output in RNA-Seq experiments causes a large number of false positive hits to be made to the cpn60 database. A standard protocol for taxonomic characterization of a sample using marker genes involves the use of universal PCR primers, whose products are amplified and then sequenced.

This specific targeting and sequencing of the marker genes allows for a much more robust characterization, as sequencing errors and other noise will be vastly reduced relative to the -marker signal. An RNA-Seq characterization, however, sequences the total RNA of a sample, and because of this many marker genes will make up only a small fraction of the sequenced output. In this study, only 107,824 of the 1,035,181,767 total reads (0.01%) aligned to any marker in cpnDB. At these low absolute abundance levels, sequencing errors and subsequent imprecise alignments will produce problematic numbers of spurious taxonomic predictions as an increasing number of markers are matched to. Indeed, a sizeable number of alignments to the alligator cpn60 sequence in the current study caused the taxa to be fairly highly ranked among the predicted taxa, and this suggested the limitations of this type of population characterization for a metatranscriptomic dataset.

One tool for limiting the impact of spurious matches to a conserved database is the Lowest Common Ancestor (LCA) algorithm (Huson et al. 2007). This algorithm takes all the alignments of a query sequence to a taxonomically-annotated database, and it uses these to push the predicted taxonomy for the query up the tree until it is inclusive of all the taxa to which the sequence aligned in the database. By making the taxonomic prediction less specific, the LCA algorithm can reduce the number of false positive predictions for a sample.

However, applying this filter on the alignments to the cpn60 database still failed to produce acceptable taxonomic predictions because of the overall low number of alignments for a majority of the taxa in the samples. While taxa belonging to the *Viridiplantae* clade had over 75,000 reads binned by the LCA algorithm, and *Brachypodium distachyon* was correctly predicted to be the most abundant plant species, fungal taxa had only 941 reads binned to them, and bacterial taxa had only 8. Even though 941 reads may have provided adequate coverage for

taxonomic characterization of the fungal population, the cpnDB database contains only one cpn60 sequence from the entire *Glomeromycota* phylum, to which all of the fungal symbionts studied here belong and to which all mycorrhizal fungal reads will be binned. Marker-based taxonomic classification therefore suffers from two major limitations to its applicability in metatranscriptomic studies, even if measures are taken to limit spurious predictions. Taxa with low levels of abundance, such as the bacteria studied here, may be absent or inadequately represented in the resulting taxonomic predictions if the marker gene is not extremely highly expressed. Furthermore, although such marker databases are routinely updated with sequences from newly characterized taxa, the inherent incompleteness of a database can lead to unacceptable biases for a particular taxonomic group of interest (Table 3).

For these reasons, a full taxonomic characterization of all reads was conducted using alignments to the NCBI-nr database and the LCA algorithm. A comprehensive sequence database like NCBI-nr will be significantly less biased towards well-studied species than a manually curated, single-marker specific database like cpnDB. This increases the specificity of taxonomic predictions. Additionally, the presence of genes from across entire genomes allows the identification of lowly abundant or lowly transcriptionally active taxa, since this increases the sequence space over which sequencing reads may be sampled from for use in the taxonomic characterization. This increased sensitivity can be seen in Table 2, where taxonomic characterization of the samples using the full set of all reads provided taxonomic assignments for a much greater portion of reads from the lowly-abundant bacteria than did characterization with the cpn60 marker gene.

It is important to consider exactly what is being represented by the abundance estimates provided by counting reads binned to various taxa in a metatranscriptomic experiment. Although

metagenomic surveys are able to use marker or read abundances to approximate taxon abundances, read abundances in metatranscriptomic experiments are affected by more than just the abundance of the originating organism. While the genome count-per-cell is effectively invariant or minimally variable when considering a randomly dividing population of a specific organism, the total RNA output per cell can vary significantly across treatments (Traganos, Darzynkiewicz, and Melamed 1982). It is also expected that certain taxa can be more or less transcriptionally active at any given time relative to other taxa, as when organisms of different trophic capabilities occupy the same environment through changes in nutritional availability. Lastly, the abundance of RNA-Seq reads binned to a particular taxon can of course vary with the differential expression of a small number of highly abundant genes, because the distribution of reads along a transcriptome is not uniform or constant like it generally is with genomic reads. Therefore, the most accurate description of the phenomenon estimated by metatranscriptomic read counts for a taxon is the proportion of the total transcribed RNA in the sample contributed by the taxon's transcriptome. Importantly, RNA abundance can be affected either by alterations in the abundance of an organism, and / or by gene expression levels. Thus, RNA does not provide information on the taxon's cellular abundance in the sample beyond presence or absence. It also does not necessarily suggest anything about the taxon's transcriptomic activity per cell, which is what read counts are commonly interpreted as representing in standard RNA-Seq experiments.

Despite this ambiguity in the cause for a particular read count for a particular taxon in a metatranscriptomic experiment, the phenomenon represented by the read count is still of value for the survey. The significant abundance of metatranscriptomic reads binned to a particular taxon is good evidence that both the taxon is present and that a given gene is transcriptionally

active in the environment. This evidence can therefore be used to form a qualitative characterization of the taxonomic makeup of the environment to determine which taxa are present and transcriptionally active in the environment. There is another measure, however, that can be quantitatively compared in this type of dataset and also provides useful information about the environmental biota.

An organism's proteome can have significant effects on the biochemical activity in an environment, and this is true whether the elements of that proteome are produced by a few individuals or by many. Because a proteome derives from the transcriptome, the same may be said of an organism's transcriptome. There is not necessarily a correlation between the abundance of a transcriptome and the abundance of a proteome; in the case of individual transcripts and proteins, this is generally a weak correlation (Maier, Güell, and Serrano 2009), but the correlation becomes significant and positive when considering only differentially expressed transcripts (Koussounadis et al. 2015). Regardless, given the lack of a direct high-throughput measure of even a single taxon's proteomic output, changes in the transcriptomic output of the various taxa in a community provide interesting insight into the relative contributions of the taxa to the biochemical activity in an environment. For this reason, the *transcriptomic abundance* of individual taxa was compared quantitatively using Student's t-test between fungal inoculated roots and the mock (control) roots. As stated above, changes in this transcriptomic abundance for a taxon are equally efficacious from a biochemical standpoint whether they derive from a fixed number of individuals transcribing a greater number of transcripts or whether they are caused by an increase in the number of individuals expressing transcripts at a constant rate.

Overall, several taxa showed higher transcriptome abundance in inoculated compared to non-inoculated roots (Table 5). The *Bacteroidetes* phylum was highly overrepresented among the taxa with the most differentially-abundant transcriptomes. Bacteria of this phylum are prevalent environmental bacteria, and one notable sub taxa is the soil-inhabiting *Cytophagaceae*. These bacteria possess novel mechanisms to digest insoluble cellulose, and they may use cell-surface proteins for the initial cellulose digestion (McBride et al. 2014).

Among the *Proteobacteria*, both the *Dickeya* genus and *Bacillus cereus* are pathogenic, with *Dickeya* causing soft rot in certain plants while *B. cereus* can infect the rhizosphere of soybean plants (Halverson, Clayton, and Handelsman 1993). However, two other *Proteobacteria* with significantly differentially-abundant transcriptomes, *Rhodopseudomonas palustris* and *Marinobacter* are both noteworthy for their unique metabolic capabilities. *R. palustris* has extremely versatile metabolic capabilities, able to grow as a photoautotroph, photoheterotroph, chemoheterotroph, or as a chemoautotroph and able to thrive in both aerobic and anaerobic environments (Larimer et al. 2004). It is also known to degrade plant biomass. *Marinobacter* is notable for comprising a genus that includes species which possess the metabolic capacity to degrade hydrocarbons (Brito et al. 2006). Overall, these bacterial taxa whose transcriptomes may be enriched in fungal-symbiotic roots show a variety of unique metabolic features, and often share the ability to degrade cell walls. This trend is supported by Blast2GO annotation of the differentially abundant contigs in Table 7, which includes multiple contigs annotated as “cell wall-associated hydrolases” and with significantly increased expression in at least one symbiosis relative to mock roots.

Tests for differential transcriptome abundances between individual mycorrhizal symbioses and mock-inoculated roots showed that *G. gigantea* mycorrhiza are enriched for the most diverse microbial communities, while *G. versiforme* mycorrhiza harbor very few microbial taxa with transcriptomic abundance significantly different from mock roots (Figure 2). The reason for this is not an overall low level of transcriptomic read counts in *G. versiforme*-inoculated roots; mean transcriptomic abundances for the minimally-abundant bacterial taxa was actually higher in *G. versiforme* mycorrhiza than it was for the other two symbionts. Therefore, it seems that *G. versiforme* harbors a set of microbial taxa which truly show less selectivity for the mycorrhizal environment over the mock-inoculated root environment. It is interesting that *G. versiforme* mycorrhiza are colonized by fewer selective taxa than *G. intraradices* mycorrhiza; previous surveys (Naumann, Schüßler, and Bonfante 2010) have shown that *G. intraradices* is actually ones of very few arbuscular mycorrhizal fungi which do not harbor the *Mollicutes*-related endobacteria (MREs) that are highly prevalent among other AM fungi, including those in the *Glomus* genus to which *G. versiforme* falls within and which *G. intraradices* has been removed from (Krüger et al. 2012). Simultaneously, an explanation for the very high prevalence of taxa with significantly overabundant transcriptomes in the *G. gigantea* symbiosis may come from the unique ability for arbuscular mycorrhizal fungi in the *Gigasporaceae* family to be colonized by both MRE species as well as the *Gigasporaceae*-specific endosymbiont *Candidatus Glomeribacter gigasporarum* (Desirò et al. 2014). Indeed, the *Burkholderiaceae* clade under which *Ca. G. gigasporarum* has been assigned was highly represented among the contigs which MEGAN annotated as bacterial (Figure 5), indicating a high level of transcriptomic sequence diversity amongst the clade in the mycorrhizal community.

Another potential reason that *G. gigantea* mycorrhiza may harbor such a diverse microbial community is that the fungus may reduce plant host defenses against bacterial infection. *B. distachyon* plants inoculated with *G. gigantea* spores had the lowest biomass between the three mycorrhiza tested (personal communication from Liudmila Mainzer). This reduction in plant biomass production may be due to general health detriment caused by the nutritional and metabolic stresses of increased microbial infection in plants colonized by *G. gigantea.*

Specific taxa which exhibited increased transcriptomic abundance in the *G. gigantea* symbiosis included a number of taxa spanning three different phyla. Some, such as *Xanthomonadaceae*, are common plant pathogens, while others are unrelated to plant or soil environments, including the human pathogen *Mycobacterium* and the seawater microbial family *Rhodobacteraceae*. Both the *Bradyrhizobiaceae* family and *Citrobacter* genus include species roles in the soil nitrogen cycle. *Bradyrhizobiaceae* includes endophytic nitrogen fixing species, and so transcriptomes of these organisms would necessarily be sampled with the collection mechanisms used in this study. It is noteworthy that such an important microbe is providing higher levels of transcriptomic products in one of the mycorrhizal symbioses studied, although there does not appear to be any published evidence for interactions between a mycorrhizal fungal population and microbial endophytes. Such interactions have, however, been reported between mycorrhizal fungi and fungal endophytes (Park and Eom 2007). Citrobacter species, when found in soil, are not located within plant tissue. Their presence in the root samples may be due to partially incomplete washing of the soil from root samples, which will necessarily be less than perfectly efficient. Most interestingly, the *Pseudomonas* genus was among the taxa which were significantly more transcriptomically abundant in only the *G. gigantea* symbiosis. At least two

species of *Pseudomonas*, *Pseudomonas fluorescens* and *Pseudomonas montelilii*, have been shown to modulate the morphology of mycorrhizal fungi. The effects of *Pseudomonas* species on mycorrhizal fungi have been observed in both ectomycorrhizal fungi (Deveau et al. 2007) and arbuscular mycorrhizal fungi (Duponnois and Plenchette 2003), and the specific morphological changes caused by *Pseudomonas* in these fungi warrants further attention to this microbe. Treatment with the bacteria was observed to correlate with an increase in nodulation and colonization of plant roots by mycorrhizal fungi as well as increases in hyphal growth. Specifically, *P. fluorescens* has been shown to enhance the hyphal extension, branching angle and branching density of mycorrhizal fungi. As the mycorrhizal hyphae are the morphological trait which provides the plant with increased nutrient uptake by increasing the available absorptive surface area, this modulation of the length, shape, and density of the hyphae has important implications for the effectiveness of a particular mycorrhizal symbiosis.

Isolation of bacterial contigs

As the focus of this study was limited to the microbiota of the arbuscular mycorrhizal symbiosis, differential expression analysis and functional annotation was conducted on only those contigs which were likely to be of bacterial origin. Two methods were used to isolate this set of contigs. The first compiled the number of alignments of taxonomically-binned short reads to all contigs, and then selected only those contigs which had reads from the bacterial bin align and none from the eukaryotic bin align. A second method taxonomically binned the assembled contigs directly using MEGAN, selecting those contigs which the LCA algorithm assigned at or under the bacterial node. Of the two methods, the LCA-binned contig method proved to have much greater power for separating the contigs at the superkingdom level, assigning over twenty times more contigs to the bacterial group than the binned-read method. However, the LCA-

binned contig method may have also assigned a greater fraction of chimeric contigs to the bacterial bin, since a majority of the contigs assigned by this method had alignments to short reads which were binned as eukaryotic and were therefore not placed in the bacterial bin by the binned-read method. However, due to the both the significant type I error rate of the taxonomic binning of short reads and the potential for incorrect and ambiguous alignments caused by sequencing errors and the conservedness of sequences, the binned-read method for isolating bacterial contigs may be overly conservative for this purpose, and many of the contigs predicted to be chimeric by the binned-read method may actually be homophyletic. Indeed, after applying a minimum bacterial read-abundance filter to the bacterially-binned contigs, the number of contigs remaining for the LCA-binned method was reduced to within an order of magnitude from the number produced by the binned-read isolation method.

Taxonomic binning of bacterial contigs

In a metatranscriptomic assembly, ideal contigs provide a deduplicated representation of all of the transcripts produced by organisms in the sample environment. The taxonomic makeup of the set of all contigs, therefore, depicts the relative transcriptomic sequence diversity of each clade in the environment, assuming that no two clades share identical transcripts. For example, a clade that includes one species with a genome containing 100 transcribed genes will produce half the number of contigs as a clade that includes two species each with a genome of 100 transcribed genes or a clade that includes one species with a genome of 200 transcribed genes. However, it should be noted that taxa with very low transcriptomic output with likely be underrepresented among the assembled contigs, as their transcripts will be under-sampled in the sequencing and may not be able to assemble with adequate quality.

In the mycorrhizal root environments studied, fungi produced a more diverse set of transcripts than bacteria, as seen in Figure 4. Additionally, bacteria produced a more diverse set of transcripts than the *Metazoa*, which contribute only 514 unique transcripts to the metatranscriptome. This indicates that the relatively high proportion of reads binned to *Metazoa* in Figure 1 originate from only a few highly expressed genes.

Among the bacterial contigs, an overwhelming majority were assigned to the *Burkholderia* genus. This genus contains a number of soil and fungal pathogens and symbionts, notably including the obligate endosymbiont of *Gigasporaceae* mycorrhizal fungi, *Ca. G. gigasporarum*. Additionally, the clade containing the other well-described obligate endosymbionts of the arbuscular mycorrhizal fungi, the *Mollicutes*-related endobacteria (MRE), contributed a substantial number of contigs. Together, these two clades, the *Burkholderia* and MRE, provided over 75% of the contigs annotated as bacterial in this study.

Functional annotation, differential expression of bacterial contigs

A number of functions were encoded by the bacterial contigs present in these samples at a minimal level of abundance, as identified by bacterial read counts. Among these were contigs involved in amino acid synthesis, carbohydrate metabolism, and cell wall-associated hydrolases. A number of these showed statistically significant differential expression between inoculated and control treatments and between time points. One contig was annotated as an O-acetylhomoserine aminocarboxypropyltransferase gene. This protein is involved in the synthesis of both methionine and cysteine, and the contig was expressed at significantly higher levels in all three of the fungal symbioses studied as well as in roots at nine weeks-post-inoculation. A serine acetyltransferase coding gene was also expressed higher in *G. versiforme* symbioses when compared to non-inoculated roots as well as across time points. There were also several contigs

annotated as being involved in carbohydrate metabolism, including starch, mannose, galactose, and glucose metabolism. Both alpha-galactosidase and alpha-mannosidase genes were overexpressed in *G. gigantea* roots at both time points relative to control, and were expressed significantly higher in roots at 5 weeks-post-inoculation. Previous characterizations of the obligate endosymbionts of the mycorrhizal fungi indicate that the bacteria are heavily dependent on the host fungi for nutritional needs, and the presence of bacterial genes involved in certain essential biochemical pathways, including certain carbon utilization and amino acid synthesis pathwayss, communicate important information about the specific metabolic capabilities of the endosymbionts in a particular mycorrhiza.

A number of contigs were annotated as coding for cell wall-associated hydrolases. As endosymbionts, these proteins would be essential for colonization of the host. Additionally, these have been shown to be used by certain bacteria as anti-microbial agents, allowing control over the local microbiome. Interestingly, two of these cell wall-associated hydrolases code for cytoplasmic, transmembrane, and non-cytoplasmic domains, suggesting that these are extracellular hydrolases. Of these contigs, two were expressed significantly higher in only the *G. versiforme* symbiosis compared to control roots, but both were expressed higher in nine-week roots relative to five-week roots. One of the cell wall-associated hydrolases showed the opposite time-dependent expression pattern in *G. gigantea*, being expressed higher in five week roots.

The presence of a secreted protein among the bacterial contigs is of particular interest. It is known that plant hosts control arbuscular mycorrhizal fungi using the secreted strigolactone hormones—, and more recently it has been shown that *Ca. G. gigasporarum* can modulate the responsiveness of these fungi to this hormone (Salvioli et al. 2010). The mechanism for this

modulatory function is not currently known, and the presence of secreted bacterial proteins in the metatranscriptome suggests that bacterial exudates may potentially be involved.

Gene ontology characterization of *Dh*MRE contigs

There was a significant presence of contigs which were taxonomically annotated specifically to the *Dh*MRE endosymbionts of *Dentiscutata heterogama*. To determine the specific genes and functions that were encoded by transcripts present in the metatranscriptome of this survey, gene ontology annotation of these contigs was performed. A large number (75.9%) of the contigs were annotated for kinase activity, indicating that functional regulatory mechanisms play a significant role in the activity of these bacterial symbionts. A large portion were annotated as exhibiting hydrolase activity, of which a number were specifically predicted to act on carbon-nitrogen, but not peptide, bonds. This family of proteins is believed to be involved in the reduction of organic nitrogen compounds (Bork and Koonin 1994), and the presence of these contigs suggests that the bacterial symbionts may be participating in the nitrogen cycle. This is a function that arbuscular mycorrhizal fungal symbiosis has recently been shown to act in (Whiteside et al. 2012), and the existence of genes encoding parts of this pathway in the bacterial symbionts suggest that this process may involve multiple members of the mycorrhizal symbiosis.

Biological process annotations of the *Dh*MRE contigs showed that a majority were involved in both metabolic and cellular processes. Of those involved in cellular processes, phosphate-containing compound metabolic processes were the major functional group, supporting the involvement of bacteria in the phosphate-uptake pathway. Additionally, contigs annotated as containing metabolic activity were all involved in macromolecule modification,

protein metabolic processes, and cellular macromolecule metabolic processes. As the endobacteria are heavily nutritionally dependent on the host, these contigs may encode for genes involved in the bacterial metabolic processes as well as the host-invasion processes.

Challenges with a metatranscriptomic analysis of a multilayered symbiosis

Extremely skewed abundance distribution

Metatranscriptomic experiments hold great promise to expose the molecular activity of diverse populations, but a number of challenges remain. In this study, three levels of symbiotic organisms were studied, including plant, fungal, and bacterial clades. The complexities of analyzing the transcriptomic activity of such divergent organisms pooled together is compounded by the extreme differences in RNA abundances between the three clades. Although the sequencing depth of the RNA-Seq libraries used in this study was fairly high, the overwhelmingly higher number of reads from the plant host compared to the fungal symbiont, as well as for the fungal host compared to its bacterial symbionts, results in a vast range of sequencing coverage for each clade. These differences are a result of the orders of magnitude differences in cellular abundance between the different symbionts, with the bulk of the mycorrhiza consisting of root tissue, a lesser portion of consisting of fungal spores, and an even lesser fraction containing bacterial cells. As the power to detect expression and changes in expression of genes in an RNA-Seq experiment increases with the number of reads, and requires a minimum number of reads due to error rates in sequencing—, this stratified population structure makes the detection of bacterial, and to smaller extent fungal, gene expression much less sensitive. In the symbioses studied here, the root mycorrhizal samples provided 700,439,111

total reads, of which 408,390,641 could be reliably annotated by taxonomic assignment. Of these 408,390,641 reads, an overwhelming 96.5% were assigned to *Viridiplantae*, likely originating from the *B. distachyon* host. This provides more than adequate coverage for analysis of the plant host. However, the fungal symbionts contributed only 3.23% of the taxonomically-annotated reads in the root samples, effectively reducing the coverage of their genes several fold. The bacterial symbionts merely provide 0.094% of the taxonomically-annotated root sample reads, so the detection of bacterial genes will necessarily be limited to only the most highly abundant genes expressed by the most abundant bacterial species in the samples. This problem is exacerbated by the existence of genes in all organisms with an overall higher level of expression than most genes, causing the coverage of other low expression genes to be further depressed.

In this study, 31,612 contigs had no significant matches to *Brachypodium*, and therefore could be presumed to be of either fungal or bacterial origin. However, of these contigs only 33.4% could be confidently predicted to be of either bacterial or eukaryotic origin. This leaves 21,042 contigs whose gene products provide potentially significant contributions to the mycorrhizal symbiosis for which the transcriptional activity cannot be confidently credited to one taxon, even between two different domains of life. Again, due to the obligate natures of both the fungal-plant symbiosis and the bacterial-fungal symbiosis, neither the fungi nor the bacterial endosymbionts may be cultured or enriched independently of the plant host, and so this problem can only be alleviated with increased sequencing depth. The realities of current generation sequencing set limits on this, as the cost of increasing the depth by several fold may be prohibitive in many cases.

High conservation of short read sequences

It is important to note that a major limitation to accurate taxonomic analysis in contemporary metagenomic or metatranscriptomic studies is the short read lengths available with current sequencing techniques. Illumina HiSeq and MiSeq sequencing protocols allow for between 100 and 250 cycles, depending on the machine used, and reads of this length often do not contain enough information for accurate taxonomic assignment. In this study, which used 100 cycles of Illumina HiSeq sequencing, over 41% of filtered, high quality reads could not be confidently assigned to either the bacterial or eukaryotic domains. Although some of this can be caused by sequencing errors resulting in unsuccessful alignments, the major cause of this lack of sensitivity is the existence of extensive conserved regions in the genomes of all species—. Short read sequencing is generally motivated by a desire to assemble full length sequences of genes from a single organism using high, redundant coverage and the resulting overlap of short read fragments. Metagenomic sequencing, however, relies largely on divergent regions of genes and genomes in order to differentiate genetic elements belonging to particular species in a pooled sample. Although short read sequencing can and has been applied to metagenomic characterizations, its utility is significantly limited by the fact that although short read fragments contain enough sequence information for kmer and overlap-based assembly, they are often of insufficient length to span across highly conserved regions in genes and genomes. This results in vast regions of the genes of organisms present in metagenomic samples being unable to be reliably assigned a taxonomic origin. While it is informative to observe transcriptional activity as a holistic representation of the genetic activity of a local environment, it is often significantly more useful to assess the genetic activity of individual clades present in an environment independently. This is especially true of environments containing organisms spanning multiple

domains, as was the case in the current mycorrhizal survey, where both fungal and bacterial genic function were of interest as separate cladistic entities operating in different symbioses.

Although it may initially seem that conserved regions of a gene provide little insight into the genetic activity of a metagenomic sample, as their sequence and protein structure is by definition minimally variable across taxa, these regions do in fact convey unique information about individual species' cellular function if this taxonomic identity is known. In the current study, several bacterial contigs were annotated as coding for cell wall hydrolases. This function is known to be present in the mycorrhizal fungi, which invade the plant root cells in a concerted mutually catalyzed process. Therefore, it is expected that the fungi also express cell wall hydrolase transcripts. However, the existence of cell wall hydrolases with high likelihood of originating from a bacterial transcriptome depicts an aspect of the bacterial invasion of the fungal spores, which is an independent and complementary process to the fungal invasion of plant roots. Fortunately in the case of this protein family, genic sequences were divergent enough to allow their taxonomic separation into the clades of their respective species. Still, a majority of the contigs either did not align to any bacterial- or eukaryotic-binned reads (63.3%) or aligned to both bacterial and eukaryotic reads (3.27%) and therefore could not be included in this bacterial-focused transcriptomic survey, demonstrating the limitations conferred by high levels of conservation and sequence ambiguity. This problem may be alleviated by more sensitive and precise taxonomic binning methods, and there are significant improvements that can be made to the LCA algorithm as implemented by MEGAN. However, even optimal taxonomic binning algorithms cannot correct for perfect conservation among genes and genic regions in distantly related species, and with short reads there will always be a significant number of reads falling in

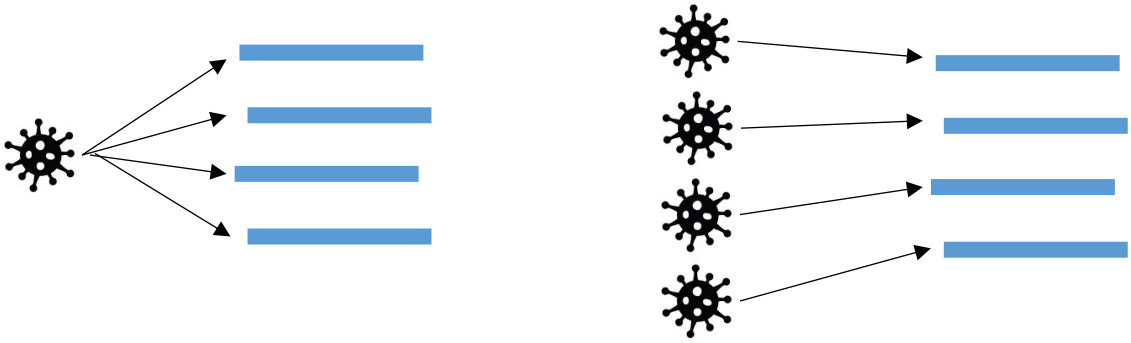
such perfectly or near-perfectly conserved regions which cannot be taxonomically annotated to low ranks.

The single molecule sequencing methods currently being introduced offer significant promise for metagenomic studies, as they will greatly reduce the challenges caused by the conservation of gene sequences and the potentially massive differences in DNA or RNA production between pooled organisms. The likelihood of mutations in a genomic region is proportional to the length of that region, and so longer sequences will tend to be less conserved. Indeed, although there will exist many conserved regions of the size range of shotgun sequencing reads (100-250bp) in any given genome, the number of such regions spanning the multikilobase sizes of current single molecule sequencing technologies will be orders of magnitude lower. These sequencing technologies do, however, tend to have higher error rates. Increased error rates can reduce the alignment scores of query sequences to reference databases that similarity-based taxonomic binning methods depend on for assignment, and therefore may make taxonomic assignment more difficult. However, some single molecule sequencing technologies, such as the SMRT sequencing system by Pacific Biosciences, have random error profiles, and therefore these errors will systematically affect alignment scores more or less equally across all hits to reference databases. As the parameters of taxonomic assignment based on these alignment scores are adjusted either manually, as in MEGAN, or algorithmically, as in CARMA3 and DiScRIBinATE, based on alignment properties, it is possible for taxonomic binning methods to account for increased error rates while still benefiting from the increased read lengths proffered by single molecule sequencing technologies. Also, although these technologies provide lower coverage of the DNA or RNA in a sample, the massively increased proportion of non-conserved sequences within each long read allows for a much more robust and sensitive taxonomic

identification, and the proportion of metagenomic or metatranscriptomic reads able to be taxonomically annotated will be significantly higher with these technologies. Additionally, while increased coverage will still be required to compensate for a greater abundance of host DNA or RNA in symbioses, the fold change increase in sequencing depth required will be lower due to this increased sensitivity of taxonomic assignment.

Convolution of differential expression with changes in population structure

Figure 6. Scenarios producing particular gene read counts in metatranscriptomic samples



An additional challenge exists for metatranscriptomic experiments which does not significantly affect metagenomic studies. In transcriptomics, the focus lies on gene expression, and the abundance of aligned reads is expected to proportionally represent the cellular production of a particular transcript in the particular sample treatment. However, additional factors, systematic to individual samples, may affect the abundance of sequencing reads which are unrelated to the transcriptional activity of a gene and therefore violate this assumption of proportional abundance for individual transcripts. Such systematic factors include differences in the total read counts produced by sequencing runs as well as differences in the total RNA output of certain samples. Several normalization techniques exist in single-species transcriptomic studies which aim to constrain the effects of such technical variation between sequencing

libraries. In particular, the TMM normalization (Mark D. Robinson and Oshlack 2010) uses the assumption that a majority of genes will show constant expression between samples to mitigate technical effects caused by differences in both sequencing output and overall transcriptional activity. However, an additional confounding factor is found in metatranscriptomic studies which is caused by the varying taxonomic makeup of individual samples. When RNA-Seq samples contain multiple species, the assumption that differential read abundances represent proportional differences in transcript abundances between treatments cannot be made, even after correcting for total sequencing output and overall transcriptional activity. The reason for this is the potential for variance in the proportion of cells of different taxa in metatranscriptomic samples. For example, if a particular transcript from some species X has double the read alignments in treatment A compared to treatment B, it is difficult to determine whether the transcript is differentially expressed in species X between the treatments or whether the transcript shows constant expression between the treatments but the taxonomic proportion of species X is doubled in treatment A. This causes some difficulty in interpreting differential read counts for genes in different samples, as these differences may be caused by changes in either transcriptional activity or the taxonomic distribution of cells in the samples. While this challenge exists to some degree in metagenomic studies, as genes or entire genomic regions may be present in duplicate copies in genomes, the extent of the problem is more limited as many genes are expressed in one or relatively few copies. This is especially true of microbial genomes, which are of relatively limited size and complexity compared to eukaryotic genomes. In addition, metagenomic studies are generally focused on the simple presence of individual genes, and quantification is limited to a census survey of taxa present among samples.

One available solution to this problem in metatransciptomics involves sequencing the genomic DNA of pooled samples alongside RNA sequencing. Because transcriptomic studies seek to characterize changes in the regulation of gene expression between treatments, effects on transcript abundance caused by differing source organism abundance must be removed for appropriate interpretation of observed changes. Unlike metatransciptomic RNA, genic DNA content will not vary with transcriptional activity in different treatments. Rather, DNA abundance varies solely as a function of the cellular abundance of a species. For this reason, performing genomic DNA sequencing in addition to RNA sequencing of metatranscriptomic samples allows the removal of abundance effects due to differing cell populations in the samples. One simple computational protocol for using genomic DNA sequencing to reducing population abundance effects requires aligning DNA short reads, ideally using taxonomically binned reads, to assembled contigs, and then normalizing RNA read counts by these DNA read counts. Another method made available through DNA sequencing involves using 16S rRNA sequences to characterize the taxonomic makeup of a sample and then normalizing transcript counts by the relative abundance of their respective taxonomic units. Using these measures of transcripts-per-gene or transcripts-per-genome, it becomes possible to compare transcript production per cell between samples of different treatments.

The current study, however, was performed using solely RNA sequencing, as a genome sequence for the arbuscular mycorrhizal fungi which motivated the study was not available at the design stages of the experiment, and the original focus of the study did not include metatranscriptomic objectives. Although this necessarily alters the interpretation of differential transcript abundances, informative results may still be obtained. While changes in transcript-aligned read abundances may be caused by either population changes or transcriptional changes

in this survey, the end effect of these variations in transcript abundances is still a proportional change in the abundances of the translated proteins. Therefore, metatranscriptomic sequencing can be considered here to provide a quantitative survey of the overall transcriptomic production of an operational taxonomic unit in a sample's environment. While it is difficult to assign a regulatory mechanism for differential transcript abundances in this case, a picture of the molecular and cellular mechanisms driving the symbiosis is still provided. For example, cell wall-associated hydrolases are clearly a major component of this symbiosis, and the importance of these genes in the symbiosis can be inferred strictly from their differential abundance, regardless of whether their increased abundance is caused by constant expression amongst an increased proportion of cells or by increased expression in a constant proportion of cells.

Conclusion

The microbiome has been shown to be an important component of a diverse range of environments, and it plays a heavy role in the plant root ecosystem, forming complex interactions with all members of the community. In this study, metatranscriptomics was applied to elucidate the composition and mechanisms of the microbiome of the mycorrhizal fungi that colonize plant roots. Characterization of the taxonomic composition of the different mycorrhiza showed that different species of mycorrhizal fungi are colonized by both distinct communities and unique diversities of bacteria, with certain species of mycorrhizal fungi being colonized by different and orders of magnitude more numerous bacterial taxa than others. Specifically, *Gigaspora gigantea* is enriched for certain *Actinobacteria* and *Proteobacteria* clades and is colonized by over one hundred unique taxa, while *Glomus intraradices* specifically shows colonization by *Bacteroidales*, *Spirochaetales*, *Proteobacteria*, and *Firmicutes* bacteria among the three fungal symbionts and is colonized by less than fifty unique taxa. *Glomus versiforme*, on the other hand, shows evidence of colonization by only ten total bacterial taxa, suggesting its microbiotic diversity is greatly reduced compared to the other symbionts. In addition to characterizing the taxonomic composition of the mycorrhizal microbiome, metatranscriptomics allowed the assembly of transcript-derived contigs from both of the bacterial groups known to participate in obligate endosymbiosis with the mycorrhizal fungi, including species from *Burkholderiaceae* and the *Dh*MRE clades. The bacterial contigs showed differential abundance patterns, as measured by bacterially-annotated read counts, suggesting that transcripts coding for a range of metabolic and structural proteins are present at different levels between roots colonized by mycorrhizal fungi and non-colonized roots as well between fungal-colonized roots at different times in the life cycle of the symbiosis. Contigs taxonomically assigned to a major

obligate endosymbiont of the mycorrhizal fungi, the *Dh*MRE, were annotated for several functions with significant implications for the mycorrhizal symbiosis, including phosphate uptake, nitrogen reduction, regulatory mechanisms, and host-derived nutrient processing. Of particular note is the potential role of the endobacterial symbiosis in nitrogen and phosphate uptake to the plant roots, which may be of direct significance to the adaptability of agricultural crops to diverse environments and conditions. Overall, metatranscriptomic approaches hold great promise for the study of communities which may be difficult or impossible to study through traditional genomic means, and although several challenges still exist, methods and technologies currently under development will likely increase the power and accessibility of such studies in the future.

References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. "Basic Local Alignment Search Tool." *Journal of Molecular Biology* 215 (3): 403–10. doi:10.1016/S0022-2836(05)80360-2.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300.

Bianciotto, V., C. Bandi, D. Minerdi, M. Sironi, H. V. Tichy, and P. Bonfante. 1996. "An Obligately Endosymbiotic Mycorrhizal Fungus Itself Harbors Obligately Intracellular Bacteria." *Applied and Environmental Microbiology* 62 (8): 3005–10.

Bianciotto, Valeria, Erica Lumini, Paola Bonfante, and Peter Vandamme. 2003. "'Candidatus Glomeribacter Gigasporarum' Gen. Nov., Sp. Nov., an Endosymbiont of Arbuscular Mycorrhizal Fungi." *International Journal of Systematic and Evolutionary Microbiology* 53 (Pt 1): 121–24. doi:10.1099/ijs.0.02382-0.

Bolan, N. S. 1991. "A Critical Review on the Role of Mycorrhizal Fungi in the Uptake of Phosphorus by Plants." *Plant and Soil* 134 (2): 189–207. doi:10.1007/BF00012037.

Bork, P., and E. V. Koonin. 1994. "A New Family of Carbon-Nitrogen Hydrolases." *Protein Science: A Publication of the Protein Society* 3 (8): 1344–46. doi:10.1002/pro.5560030821.

Brito, Elcia Margareth S., Rémy Guyoneaud, Marisol Goñi-Urriza, Antony Ranchou-Peyruse, Arnaud Verbaere, Miriam A.C. Crapez, Julio César A. Wasserman, and Robert Duran. 2006. "Characterization of Hydrocarbonoclastic Bacterial Communities from Mangrove Sediments in Guanabara Bay, Brazil." *Research in Microbiology* 157 (8): 752–62. doi:10.1016/j.resmic.2006.03.005.

Brousseau, R., J. E. Hill, G. Préfontaine, S. H. Goh, J. Harel, and S. M. Hemmingsen. 2001. "Streptococcus Suis Serotypes Characterized by Analysis of Chaperonin 60 Gene Sequences." *Applied and Environmental Microbiology* 67 (10): 4828–33.

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.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson. 2014. "Fast and Sensitive Protein Alignment Using DIAMOND." *Nature Methods* 12 (1): 59–60. doi:10.1038/nmeth.3176.

Castillo, Dean M., and Teresa E. Pawlowska. 2010. "Molecular Evolution in Bacterial Endosymbionts of Fungi." *Molecular Biology and Evolution* 27 (3): 622–36. doi:10.1093/molbev/msp280.

Conesa, Ana, and Stefan Götz. 2008. "Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics." *International Journal of Plant Genomics* 2008: 1–12. doi:10.1155/2008/619832.

Desirò, Alessandro, Alessandra Salvioli, Eddy L Ngonkeu, Stephen J Mondo, Sara Epis, Antonella Faccio, Andres Kaech, Teresa E Pawlowska, and Paola Bonfante. 2014. "Detection of a Novel Intracellular Microbiome Hosted in Arbuscular Mycorrhizal Fungi." *The ISME Journal* 8 (2): 257–70. doi:10.1038/ismej.2013.151.

Deveau, A., B. Palin, C. Delaruelle, M. Peter, A. Kohler, J. C. Pierrat, A. Sarniguet, J. Garbaye, F. Martin, and P. Frey-Klett. 2007. "The Mycorrhiza Helper Pseudomonas Fluorescens BBc6R8 Has a Specific Priming Effect on the Growth, Morphology and Gene Expression

of the Ectomycorrhizal Fungus Laccaria Bicolor S238N." *The New Phytologist* 175 (4): 743–55. doi:10.1111/j.1469-8137.2007.02148.x.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21. doi:10.1093/bioinformatics/bts635.

Duponnois, R., and C. Plenchette. 2003. "A Mycorrhiza Helper Bacterium Enhances Ectomycorrhizal and Endomycorrhizal Symbiosis of Australian Acacia Species." *Mycorrhiza* 13 (2): 85–91. doi:10.1007/s00572-002-0204-7.

Gerlach, W., and J. Stoye. 2011. "Taxonomic Classification of Metagenomic Shotgun Sequences with CARMA3." *Nucleic Acids Research* 39 (14): e91–e91. doi:10.1093/nar/gkr225.

Ghignone, Stefano, Alessandra Salvioli, Iulia Anca, Erica Lumini, Giuseppe Ortu, Luca Petiti, Stéphane Cruveiller, et al. 2012. "The Genome of the Obligate Endobacterium of an AM Fungus Reveals an Interphylum Network of Nutritional Interactions." *The ISME Journal* 6 (1): 136–45. doi:10.1038/ismej.2011.110.

Ghosh, Tarini Shankar, M. Monzoorul Haque, and Sharmila S. Mande. 2010. "DiScRIBinATE: A Rapid Method for Accurate Taxonomic Classification of Metagenomic Sequences." *BMC Bioinformatics* 11 Suppl 7: S14. doi:10.1186/1471-2105-11-S7-S14.

Halverson, Larry J., Murray K. Clayton, and Jo Handelsman. 1993. "Population Biology of Bacillus Cereus UW85 in the Rhizosphere of Field-Grown Soybeans." *Soil Biology and Biochemistry* 25 (4): 485–93. doi:10.1016/0038-0717(93)90074-L.

Hill, J. E. 2004. "cpnDB: A Chaperonin Sequence Database." *Genome Research* 14 (8): 1669–75. doi:10.1101/gr.2649204.

Huson, D. H., A. F. Auch, J. Qi, and S. C. Schuster. 2007. "MEGAN Analysis of Metagenomic Data." *Genome Research* 17 (3): 377–86. doi:10.1101/gr.5969107.

Jargeat, P., C. Cosseau, B. Ola'h, A. Jauneau, P. Bonfante, J. Batut, and G. Becard. 2004. "Isolation, Free-Living Capacities, and Genome Structure of 'Candidatus Glomeribacter Gigasporarum,' the Endocellular Bacterium of the Mycorrhizal Fungus Gigaspora Margarita." *Journal of Bacteriology* 186 (20): 6876–84. doi:10.1128/JB.186.20.6876-6884.2004.

Jones, P., D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, et al. 2014. "InterProScan 5: Genome-Scale Protein Function Classification." *Bioinformatics* 30 (9): 1236–40. doi:10.1093/bioinformatics/btu031.

Koussounadis, Antonis, Simon P. Langdon, In Hwa Um, David J. Harrison, and V. Anne Smith. 2015. "Relationship between Differentially Expressed mRNA and mRNA-Protein Correlations in a Xenograft Model System." *Scientific Reports* 5 (June): 10775. doi:10.1038/srep10775.

Krüger, Manuela, Claudia Krüger, Christopher Walker, Herbert Stockinger, and Arthur Schüßler. 2012. "Phylogenetic Reference Data for Systematics and Phylotaxonomy of Arbuscular Mycorrhizal Fungi from Phylum to Species Level." *New Phytologist* 193 (4): 970–84. doi:10.1111/j.1469-8137.2011.03962.x.

Lan, Yemin, Gail Rosen, and Ruth Hershberg. 2016. "Marker Genes That Are Less Conserved in Their Sequences Are Useful for Predicting Genome-Wide Similarity Levels between Closely Related Prokaryotic Strains." *Microbiome* 4 (1). doi:10.1186/s40168-016-0162-5.

Larimer, Frank W, Patrick Chain, Loren Hauser, Jane Lamerdin, Stephanie Malfatti, Long Do, Miriam L Land, et al. 2004. "Complete Genome Sequence of the Metabolically Versatile

Photosynthetic Bacterium Rhodopseudomonas Palustris." *Nature Biotechnology* 22 (1): 55–61. doi:10.1038/nbt923.

Lee, Eun-Hwa, Ju-Kyeong Eo, Kang-Hyeon Ka, and Ahn-Heum Eom. 2013. "Diversity of Arbuscular Mycorrhizal Fungi and Their Roles in Ecosystems." *Mycobiology* 41 (3): 121. doi:10.5941/MYCO.2013.41.3.121.

Levy, A., B. J. Chang, L. K. Abbott, J. Kuo, G. Harnett, and T. J. J. Inglis. 2003. "Invasion of Spores of the Arbuscular Mycorrhizal Fungus Gigaspora Decipiens by Burkholderia Spp." *Applied and Environmental Microbiology* 69 (10): 6250–56. doi:10.1128/AEM.69.10.6250-6256.2003.

Lumini, Erica, Valeria Bianciotto, Patricia Jargeat, Mara Novero, Alessandra Salvioli, Antonella Faccio, Guillaume Bécard, and Paola Bonfante. 2007. "Presymbiotic Growth and Sporal Morphology Are Affected in the Arbuscular Mycorrhizal Fungus Gigaspora Margarita Cured of Its Endobacteria." *Cellular Microbiology* 9 (7): 1716–29. doi:10.1111/j.1462-5822.2007.00907.x.

Maier, Tobias, Marc Güell, and Luis Serrano. 2009. "Correlation of mRNA and Protein in Complex Biological Samples." *FEBS Letters* 583 (24): 3966–73. doi:10.1016/j.febslet.2009.10.036.

McBride, Mark J., Weifeng Liu, Xuemei Lu, Yongtao Zhu, and Weixin Zhang. 2014. "The Family Cytophagaceae." In *The Prokaryotes: Other Major Lineages of Bacteria and The Archaea*, edited by Eugene Rosenberg, Edward F. DeLong, Stephen Lory, Erko Stackebrandt, and Fabiano Thompson, 577–93. Berlin, Heidelberg: Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-38954-2_382.

Naito, Mizue, Joseph B. Morton, and Teresa E. Pawlowska. 2015. "Minimal Genomes of Mycoplasma-Related Endobacteria Are Plastic and Contain Host-Derived Genes for Sustained Life within Glomeromycota." *Proceedings of the National Academy of Sciences* 112 (25): 7791–96. doi:10.1073/pnas.1501676112.

Naumann, Maria, Arthur Schüßler, and Paola Bonfante. 2010. "The Obligate Endobacteria of Arbuscular Mycorrhizal Fungi Are Ancient Heritable Components Related to the Mollicutes." *The ISME Journal* 4 (7): 862–71. doi:10.1038/ismej.2010.21.

Ounit, Rachid, Steve Wanamaker, Timothy J Close, and Stefano Lonardi. 2015. "CLARK: Fast and Accurate Classification of Metagenomic and Genomic Sequences Using Discriminative K-Mers." *BMC Genomics* 16 (1). doi:10.1186/s12864-015-1419-2.

Park, Sang-Hyun, and Ahn-Heum Eom. 2007. "Effects of Mycorrhizal and Endophytic Fungi on Plant Community: A Microcosm Study." *Mycobiology* 35 (4): 186. doi:10.4489/MYCO.2007.35.4.186.

Peabody, Michael A., Thea Van Rossum, Raymond Lo, and Fiona S. L. Brinkman. 2015. "Evaluation of Shotgun Metagenomics Sequence Classification Methods Using in Silico and in Vitro Simulated Communities." *BMC Bioinformatics* 16 (1). doi:10.1186/s12859-015-0788-5.

Roberts, Adam, and Lior Pachter. 2012. "Streaming Fragment Assignment for Real-Time Analysis of Sequencing Experiments." *Nature Methods* 10 (1): 71–73. doi:10.1038/nmeth.2251.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010. "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1): 139–40. doi:10.1093/bioinformatics/btp616.

Robinson, Mark D., and Alicia Oshlack. 2010. “A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data.” *Genome Biology* 11 (3): R25. doi:10.1186/gb-2010-11-3-r25.

Salvioli, Alessandra, Marco Chiapello, Joel Fontaine, Anissa Lounes Hadj-Sahraoui, Anne Grandmougin-Ferjani, Luisa Lanfranco, and Paola Bonfante. 2010. “Endobacteria Affect the Metabolic Profile of Their Host Gigaspora Margarita, an Arbuscular Mycorrhizal Fungus.” *Environmental Microbiology* 12 (8): 2083–95. doi:10.1111/j.1462-2920.2010.02246.x.

Traganos, F., Z. Darzynkiewicz, and M. R. Melamed. 1982. “The Ratio of RNA to Total Nucleic Acid Content as a Quantitative Measure of Unbalanced Cell Growth.” *Cytometry* 2 (4): 212–18. doi:10.1002/cyto.990020403.

Tuomi, Juha, Minna-Maarit Kytoviita, and Roger Hardling. 2001. “Cost Efficiency of Nutrient Acquisition and the Advantage of Mycorrhizal Symbiosis for the Host Plant.” *Oikos* 92 (1): 62–70. doi:10.1034/j.1600-0706.2001.920108.x.

Whiteside, Matthew D., Michelle A. Digman, Enrico Gratton, and Kathleen K. Treseder. 2012. “Organic Nitrogen Uptake by Arbuscular Mycorrhizal Fungi in a Boreal Forest.” *Soil Biology and Biochemistry* 55 (December): 7–13. doi:10.1016/j.soilbio.2012.06.001.

Wood, Derrick E, and Steven L Salzberg. 2014. “Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments.” *Genome Biology* 15 (3): R46. doi:10.1186/gb-2014-15-3-r46.

Appendix A: Molecular Functions Represented by *Dh*MRE Contigs

The Gene Ontology map of molecular functions predicted for the contigs assigned to the *Dh*MRE clade by the Blast2GO software package may be found in a supplemental file named **contigs_DhMRE_GO_Molecular_function.png.**

Appendix B: Biological Processes Represented by *Dh*MRE Contigs

The Gene Ontology map of biological processes predicted for the contigs assigned to the *Dh*MRE clade by the Blast2GO software package may be found in a supplemental file named **contigs_DhMRE_GO_Biological_process.png.**

Appendix C: Differential Transcriptome Abundance Tests

Full results of the tests for differential transcriptome abundance for bacterial taxa between fungal inoculations and the control treatment may be found in a supplemental file named **differential_transcriptome_abundances.xlsx**. This list excludes those bacterial taxa with a higher transcriptome abundance in shoots compared to roots.