Title: Effect of LSU and ITS genetic markers and reference databases on analyses of fungal communities

Chao Xue[1,2], Yuewen Hao[1], Xiaowei Pu[1], C. Ryan Penton[3,4], Qiong Wang[2], Mengxin Zhao[2,5], Bangzhou Zhang[2,6], Wei Ran[1], Qiwei Huang[1], Qirong Shen[1†], James M. Tiedje[2†]

[1] Jiangsu Provincial Key Lab for Solid Organic Waste Utilization, National Engineering Research Center for Organic-based Fertilizers, Jiangsu Collaborative Innovation Center for Solid Organic Waste Resource Utilization, Nanjing Agricultural University, Nanjing, 210095, China

[2] Center of Microbial Ecology, Michigan State University, East Lansing, 48824, MI, USA

[3] College of Integrative Sciences and Arts, Faculty of Science and Mathematics, Arizona State University, Mesa, AZ, 85212, USA

[4] Center for Functional and Applied Microbiomics, Biodesign Institute, Arizona State University, Tempe, AZ

[5] State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing, China

[6] State Key Laboratory of Marine Environmental Science and Key Laboratory of the Ministry of Education for Coast and Wetland Ecosystems, School of Life Sciences, Xiamen University, Xiamen, 360102, Fujian, China

†**Corresponding author:** Qirong Shen; Tel: (86)-025-84395366; Fax: (86)-025-84395212; E-mail: shenqirong@njau.edu.cn. James M. Tiedje; Tel: (01)-517-3539021; Fax: (01)-517-3532917; E-mail: tiedjej@msu.edu.

Key words: Fungal database; LSU; ITS; Warcup; UNITE

Abstract:

The effect of genetic markers and reference databases on analyses of fungal communities were estimated using fungal large subunit (LSU) and internal transcribed spaces (ITS) amplicon datasets in consecutive years of rhizosphere samples from three candidate biofuel crops, corn (*Zea maize*), switchgrass (*Panicum virgatum.*) and Miscanthus (*Miscanthus × giganteus)*. These two marker genes were selected to contrast possible differences in biological conclusions. In addition, two ITS schemes based on two ITS reference databases were used to assess differences due to reference database composition. A taxonomy-supervised method was invoked using the RDP naïve Bayesian classifier that accesses all three databases. The UNITE classification scheme had the highest number of classified taxa in the raw classification result, however it also had the highest proportion of unknown taxa (sequences that were classified to "unclassified", "unidentified", Incertae sedis or, in the case of Warcup, to matches containing two unique names). After removal of these unknown taxa, LSU had highest classification rate followed by Warcup and UNITE. As expected, the communities resolved using the two ITS databases (same sequences) were relatively more similar than those from the lower coverage LSU classification scheme. The choice of marker gene or even the same reads with different classification databases revealed different community patterns due to the

coverage of the database, e.g. the relative abundance of the most abundant groups changed or were only detected in one or two of the classification schemes, such as for *Mortierella*, *Fusarium*, and *Phoma*. However, no marked influence on fungal beta-diversity was found due to the three methods, yet all three methods distinguished the fungal communities of the three biofuel crops and between the drought and normal rainfall years. Classification rates, taxonomic conflicts, and coverage differences of high-abundance fungal groups were identified among classification schemes, but these analysis choices did not affect conclusions about beta diversity differences.

Introduction

Fungi are the most diverse and dominant eukaryotes in soil (Hawksworth 2001) and can increase plant production by increasing water uptake (Miransari et al. 2007), promoting growth (Rodríguez and Fraga 1999), and bio-control (Thangavelu et al. 2004) of plant pathogens. Profiling of fungal communities can be achieved by sequencing multiple regions of the fungal rRNA genes including the small subunit (SSU) and large subunit (LSU) rRNA genes and the internal transcribed spacer (ITS) region. Sequencing of the SSU rRNA is highly effective in bacteria, but this region does not evolve rapidly enough in eukaryotes for use in higher resolution taxonomic assignments (Vandenkoornhuyse et al. 2002). The LR3/LR0R primer combination that spans divergence region (D1/D2) of LSU (28S) rRNA gene is suitable for both classification accuracy and resolution to the genus level (Liu et al. 2012). The ITS region that separates the SSU and LSU rRNAs in eukaryotes evolves at a much faster rate and can be used to identify fungi to genus and often to species level (O'Brien et al. 2005), although resolution varies across taxonomic lineages.

A previous study by Tedersoo et al. (2015) compared nine different primer pairs targeting seven nuclear ribosomal DNA (rDNA) regions of fungi including SSU, LSU, and ITS fungal genetic markers, which are often used in fungal community analyses.

Their research indicated that primer choice biased fungal community studies causing different conclusions. The ITS1 and ITS2 amplicons provided greater taxonomic and functional resolution as well as coverage of the communities compared to SSU and LSU amplicons. Of these the use of ITS2 or the whole ITS region was recommended (Tedersoo et al. 2015). Also important is the choice of the reference database for sequence analysis to obtain the taxonomic annotation. Our study extends the evaluation of genetic marker analysis to the three fungal reference databases LSU, UNITE, and the new fungal ITS database Warcup (released in 2016) for fungal taxonomy assignment using RDP's naïve Bayesian classifier.

Taxonomy supervised methods (Sul et al. 2011) are independent from alignment/clustering and are critical when OTUs cannot be defined such as when comparing classifications originating from different genetic markers or different regions of the same genetic marker, or different sequencing platforms. RDP naïve Bayesian classifier is one of the most popular bioinformatics tools to assign taxonomy (Wang et al. 2007). Currently it supports 28S rRNA gene and ITS reads classification with the LSU (Liu et al. 2012), UNITE (Kõljalg et al. 2013), and Warcup (Deshpande et al. 2016) databases. The later two are ITS databases.

In this study, effect of genetic markers and reference databases on analyses of fungal communities were estimated using fungal LSU and ITS amplicon datasets in consecutive years of rhizosphere samples from three candidate biofuel crops, corn (*Zea maize*), switchgrass (*Panicum virgatum.*) and Miscanthus (*Miscanthus* × *giganteus*), in the long-term Great Lakes Biofuel Research Center (GLBRC) biofuel cropping system. We addressed differences in the assessment of fungal community composition between the two genetic markers and among the reference databases: LSU, Warcup and UNITE, using RDP's Bayesian classifier-based, taxonomy supervised method. Our overarching objective is to understand the effects of these biofuel crops on root-associated soil fungal community composition, particularly in regards to changes in the identities and/or relative abundances of potential pathogenic or system beneficial fungi.

Materials and methods

Field site and experiment design

The experimental site is located at the Kellogg Biological Station (KBS), Hickory Corners, MI, USA (42˚23'47"N, 85˚22'26"W), a GLBRC-initiated bioenergy cropping system site established in 2008. Based on a five-block design (Fig. S1), each block, (27 m x 43 m) contained randomly assigned plots for continuous corn, switchgrass and Miscanthus. Corn was planted at 76 cm row spacings, switchgrass was planted at an

average seeding rate of 7.6 kg seed ha^{-1}, and Miscanthus rhizomes were manually planted at a depth of 10 cm at a density of 17,200 rhizomes ha^{-1}. The previous cover crop was alfalfa and no-till practices were adopted site-wide after initial site preparation. Since this is a biofuel cropping system study, all mechanically harvestable plant tops (the biofuel) were removed after plant senesce each fall minimizing plant litter return to the soil. The corn plots were applied 19-17-0 liquid fertilizer at 15.1 gallons $acre^{-1}$. The Miscanthus and switchgrass plots were fertilized with 28-0-0 fertilizer at 16.7 gallons $acre^{-1}$. The fertilizer was sprayed at planting. The predominant soil series at KBS is Kalamazoo sandy loam (MSCO 2013). Further description of the site, experimental design and productivity data are reported in Sanford et al. (2016). Background bacterial and fungal data, including lipid derived biomass, from the second crop year (2009) are reported in da C. Jesus et al. (2010) and Liang et al. (2012). To determine the organic C content, soil samples were oven-dried and then were pulverized and combusted in a Costech Elemental Combustion System 4010 (Costech Analytical Technologies, Valencia CA). Soil pH was determined with a compound electrode (PE-10, Sartorious, Germany) in a 1:2.5 soil/water ratio solution. The organic matter and soil pH were summarized in table S1.

Soil sampling and DNA extraction

Samples were collected on October 13, 2012 (5^{th} crop year since establishment) and July 26, 2013 (6^{th} crop year) within blocks 2-4. For each crop, 7 replicates were collected with 3 replicates from block 2 and 2 replicates from each of the blocks 3-4. For each replicate, roots with soil attached from three randomly selected adjacent plants were collected. Roots with very small soil particles (<1mm from root) were cut and washed using sterilized water. The rhizosphere soil was collected by centrifugation and DNA extracted from 0.5 g soil using the Powersoil DNA Extraction kit (MoBio Laboratories, Carlsbad, CA, USA).

The 2012 crop year was a severe drought with summer (from June to August) rainfall 152 mm, which severely affected crop yield, e.g. 6 Mg DM ha^{-1} for corn biomass (grain plus vegetative biomass), versus 2013 which was a more normal year, 310 mm summer (from June to August) rainfall and 16 Mg DM ha^{-1} corn biomass yield (Sanford et al. 2016).

Amplification and sequencing

Primer sets ITS9 (adapter-mid-TATGGTAATT-GG-GAACGCAGCRAAIIGYGA) and ITS4 (adapter-mid-AGTCAGTCAG-GG-TCCTCCGCTTATTGATATGC) were used to amplify the ITS2 region and primers LR3 (mid-CCGTGTTTCAAGACGGG) and LR0R (mid-ACCCGCTGAACTTAAGC) were used to amplify the partial fungal LSU genes for

all samples according to previously published protocols (White et al. 1990; Menkis et al. 2012; Penton et al. 2013), where "mid" refers to a unique barcode sequence used for sample sorting. For each sample, 3 replicated 20 µl PCRs were performed, with each mixture containing 1µl of DNA (around 20 ng/µl), 17 µl of AccuPrime Pfx SuperMix (Invitrogen, CA, USA), 1µl (10 pmol) each forward and reverse primer. Amplification was performed with an initial denaturation of 4 min at 95 °C; followed by 30 cycles of 45 sec at 94 °C, 30 sec at 55 °C and 1 min at 72 °C, and a final extension at 72 °C for 6 min. The LSU amplicons, were adapter ligated and bi-directionally sequenced on the 454 Life Sciences Titanium platform, using Lib-L kits. Sequencing was performed at Center for Integrated BioSystems (Utah State University, USA). Sequencing of ITS fragments was conducted by Joint Genome Institute (JGI) on an Illumina MiSeq sequencing platform. The longer LSU sequences required the longer read sequencing technology, which gives less coverage than the higher coverage, shorter read technology suitable for the ITS sequences.

Sequence processing

Sequence data were processed using the RDPipeline (http://pyro.cme.msu.edu) with the removal of low quality (Qscore < 20 for LSU and < 26 for ITS reads) and short reads (length < 220 bp). Chimeric reads were removed through Uchime (Edgar 2010). All

samples were randomly resampled to the same number of reads as the sample with the lowest reads (4,217 reads per sample). Taxonomy was assigned using RDP (Ribosomal Database Project) classifier (version 2.7) with 50% confidence (Wang et al. 2007). LSU reads were classified based on the fungal LSU training set released by RDP in 2011 (Liu et al. 2012) containing 11,442 sequences in 1,895 genera. ITS reads were classified based on ITS training sets UNITE (July, 2014) (Kõljalg et al. 2013) and Warcup (Deshpande et al. 2016), released by RDP in 2016. The Warcup training set contains 17,923 unique sequences including 1,461 genera, while UNITE has 145,019 unique sequences covering 2,137 genera. The classification rate was calculated after sequences classified to "unclassified", "unidentified", Incertae sedis, or, in the case of Warcup, to matches containing two unique names were filtered (Tables 1 and 2). All sequences were deposited in the NCBI Sequence Read Archive (SRA) database (Accession numbers: SRX483129 and SRX483122).

Data analysis

Non-metric multidimensional scaling (NMDS) analyses based on classification results from class to species levels of LSU, UNITE, and Warcup were performed to illustrate the beta-diversity (Bray-Curtis distances) between individual samples. Permutational multivariate analyses of variance (PERMANOVA) (Anderson 2001) was performed to

determine the significance of community composition differences between treatments. A multiple regression tree (MRT) was generated to illustrate plant type (annual/perennial), temporal (2012/2013) and crop (corn/Miscanthus/switchgrass) effects on fungal community composition (De'ath 2002). Shared taxa that were present in more than one database were calculated and illustrated in venn diagrams, from the phylum to species level. Genera were sorted by the maximal abundance of each genus among the three classification schemes and the top 30 most abundant groups were illustrated in a heatmap graph.

Results

Classification rate

In total, 177,114 LSU (4,217 reads per sample) and 4,344,975 ITS (105,975 reads per sample) reads were obtained after processing. Following RDP Naïve Bayesian classification, phyla to species level lineages were extracted. The LSU sequences exhibited the highest classification rate, followed by Warcup and UNITE (Fig. 1). From the phylum to genus, the LSU classification scheme contained 18.7% to 37.8% unclassified reads. From phylum to species levels, the unclassified reads accounted for 14.8% to 65% of the total reads in the Warcup classification scheme and 22.5% to 83.6% for UNITE, respectively.

Shared and unique taxa

Four phyla (Ascomycota, Basidiomycota, Chytridiomycota and Glomeromycota) were shared by all classification schemes with similar abundances of unique taxa from phylum to order taxonomic levels (Fig. 2). At the family and genus level, the LSU classification scheme exhibited the highest number of unique taxonomies, while UNITE and Warcup were similar. At the species level, Warcup classification resulted in almost double the number of unique taxonomies, compared to the UNITE classification scheme. Overall, 19 classes, 50 orders, 96 families, and 183 genera were shared among classification schemes with 505 species shared by the UNITE and Warcup classification schemes.

Community pattern

Overall, the combination of genetic marker and reference database influenced the sample clustering patterns. In addition, the separation was more apparent when the classification went down to the order level (Fig. 3). Dispersion among replicates was significantly higher ($p<0.05$) in LSU data (0.58±0.12) than for the UNITE (0.51±0.12) or Warcup (0.51±0.11) classification schemes. Based on different phylogenetic levels of the three classification schemes separately, NMDS ordinations showed consistent patterns from class to species levels even at family level the separation is clearly visible and does not significantly change down to the species level indicating that genetic markers and

databases had little influence on the overall beta-diversity comparisons between treatments (Fig. 4). The fungal community composition of the corn samples (annual) was most different from Miscanthus and switchgrass (perennials) that grouped closely to each other. Overall, PERMANOVA results for all databases indicated significant differences by plant, time and crop*time interactions (Table S2). Sum of squares indicated plant type (annual vs. perennial) as the primary driver of community differences. This was supported by MRT analyses, based on genus level classifications, where plant type drove the first branch split (Fig. 5). After the initial split, the Warcup and LSU classification schemes were further split by crop type (Miscanthus and switchgrass) then by sampling time. A contrary result was identified in the UNITE classification scheme, where the MRT tree was split by time then by crop type.

Community composition

In order to identify if the different genetic markers or training sets influenced relative abundances, the top 30 most abundant groups are illustrated in a heatmap (Fig. 6) Among these, *Mortierella*, *Fusarium, Gibellulopsis, Talaromyces,* and *Phoma* were detected only using the ITS genetic marker; *Orbicula* was detected only in LSU classification scheme; *Eupenicillium* and *Didymella* were absent in UNITE classification scheme; *Plectosphaerella*, *Metarhizium, Gibellulopsis*, *Aspergillus*, *Cladosporium*, *Cercophora*,

Monascus, *Edenia*, and *Hydropisphaera* were detected in a higher proportion in one or two of the three classification schemes but were in lower abundances in another.

Discussion

Influence of genetic marker and database choice on apparent rhizo-fungal community composition

In order to understand the differential effects of targeting the LSU and ITS genetic markers on perceived fungal community compositions, we utilized a taxonomy-supervised methodology (Sul et al. 2011) using the fungal LSU and two new ITS training sets (Warcup and UNITE) that have been recently integrated into the RDP naïve Bayesian classifier. Although the ITS region is capable of species-level classification, only a small proportion of sequences (35.1%-Warcup, 16.3%-UNITE) were classified to that level. The highest proportion of reads was classified using the smaller LSU reference classification scheme while the ITS training sets contained more taxonomic conflicts. This is due to the higher number of the sequences used to create the ITS training sets and may be the underlying cause of the resulting conflicts (Porras-Alfaro et al. 2014).

Less than 50% of the classified genera were shared by all classification schemes, supporting the notion that database composition strongly influences fungal classification (Porras-Alfaro et al. 2014). However, regardless of the genetic marker and database

used, ordinations showed similar patterns, indicating that both genetic marker and database had little influence in revealing overall differences in fungal community composition (beta-diversity) between treatments (Fig. 4). Further, community differences between years as well as crops were revealed by all three methods. The soil conditions for the two years, July vs October sampling and a severe drought and normal rainfall year likely drove the year differences. Interestingly, the fungal community pattern did not change with the increase of sequencing depth between 454 and Illumina MiSeq, indicating that broad biological conclusions remained constant, regardless of community coverage. Lastly, the variation in classification success (from 16.4% to 85.2%) among databases also had little influence on the conclusions.

Specific differences due to fungal genetic marker selection

The goal of any gene-targeted metagenomic study is to obtain the highest possible coverage in concert with the most accurate and diverse taxonomic information. These results have demonstrated that fungal community classification can be problematic when different genetic markers and reference databases are utilized. For example, due to differences in fungal ITS and LSU reference database coverage, a low number of shared genera in concert with a high number of unique genera was identified. Furthermore, *Mortierella*, *Fusarium*, *Gibellulopsis*, *Talaromyces*, and *Phoma* were detected as the top

abundant fungi in the ITS classification schemes, although they were absent in the LSU classification scheme, due to the lack of matching reference sequences. For example, there is only one *Mortierella* reference sequence in LSU, but 142 and 18 sequences in UNITE and WARCUP, respectively. *Fusarium*, *Gibellulopsis*, *Talaromyces*, and *Phoma* are absent in LSU reference database resulting in their failure detection using LSU database. This ultimately may be due either to the lack of LSU primer coverage or possibly from variations in taxonomic identification among classification schemes (e.g. anamorphic vs. telemorphic nomenclature). For example, the anamorph of *Fusarium* is *Gibberella* which is present in all classification schemes. *Penicillium* is anamorph of *Eupenicillium*. Teleomorph states of *Phoma* have been described in the genera *Didymella*, *Leptosphaeria*, *Pleospora* and *Mycosphaerella*, and these genera are all present in the LSU classification scheme (de Gruyter et al. 2009). In addition, due to their high abundance in the ITS classification scheme, sequencing depth (Illumina vs. 454) was an unlikely culprit in explaining their absence in the LSU data. *Orbicula* was detected only in LSU classification scheme because *Orbicula* reference sequence is included in neither UNITE nor WARCUP ITS databases. Ultimately, we cannot rule out the possibility that amplification primers demonstrate primer bias towards these groups. For example, in other ITS-based studies, *Mortierella*, *Fusarium*, and *Phoma* have also been reported as

the predominant fungal groups in various environments (Xu et al. 2012; Voříšková and Baldrian 2013; Xiong et al. 2016). Notably, of these abundant groups, *Fusarium* contains species that cause wilt disease on numerous plants (Gilbert and Tekauz 2000; Flood 2006), and many *Phoma* species are plant pathogens that cause rot disease (Hollingsworth et al. 2005; Cullen et al. 2007).

Together, these data suggest that fungal genetic marker and classification database selection is important when targeting fungal pathogens, especially in the context of disease-focused studies. While these biases ultimately did not affect biological conclusions in the context of plant and temporal effects, it can impact, for instance, diversity measures when taxonomy-supervised methods are utilized. This bias is in addition to those previously identified in gene-targeted metagenomic studies, as well as due to primer choice, soil sample size used for DNA extraction, number of replicates, and inherent aggregate to field scale spatial heterogeneity, among others (Tedersoo et al. 2015; Penton et al. 2016; Schöler et al. 2017; Vestergaard et al. 2017).

Conclusion

The choice of genetic markers and reference databases, when used with a naïve Bayesian classifier, did not affect the ability to detect plant fungal community differences (beta diversity) due to crop and to sampling year. The choice of genetic marker and

classification database did, however, result in different relative abundances among the 30 most abundant groups, so at the level of particular taxa or for defining potential indicator species, these choices do matter. Classification rates, taxonomic conflicts, and coverage differences of high-abundance fungal groups were identified among classification schemes, but these analysis choices had only minor effects.

Acknowledgements

This work was funded by the DOE Great Lakes Bioenergy Research Center, DOE BER Office of Science (DE-FC02-07ER64494 and DE-FG02-99ER62848), the Jiangsu Science and Technology Department (BK20160730), the China Postdoctoral Science Foundation (2017M621761 and 2018T110510), and the Fundamental Research Funds for the Central Universities (KYZ201720). We also acknowledge the support of the GLBRC field staff who maintains the field experiment.

References

Anderson MJ (2001) A new method for non-parametric multivariate analysis of variance. Austral Ecol 26:32–46. doi: 10.1111/j.1442-9993.2001.01070.pp.x

Cullen DW, Toth IK, Boonham N, Walsh K, Barker I, Lees AK (2007) Development and validation of conventional and quantitative polymerase chain reaction assays for the detection of storage rot potato

pathogens, *Phytophthora erythroseptica*, *Pythium ultimum* and *Phoma foveata*. J Phytopathol 155:309–315. doi: 10.1111/j.1439-0434.2007.01233.x

da C Jesus E, Susilawati E, Smith SL, Wang Q, Chai B, Farris R, Rodrigues JLM, Thelen KD, Tiedje JM (2010) Bacterial communities in the rhizosphere of biofuel crops grown on marginal lands as evaluated by 16S rRNA gene pyrosequences. BioEnergy Res 3:20–27. doi: 10.1007/s12155-009-9073-7

De'ath G (2002) Multivariate regression trees: A new techique for modeling species-environment relationships. Ecology 83:1105–1117. doi: 10.1890/0012-9658(2002)083[1105:MRTANT]2.0.CO;2

de Gruyter J, Aveskamp MM, Woudenberg JHC, Verkley GJM, Groenewald JZ, Crous PW (2009) Molecular phylogeny of *Phoma* and allied anamorph genera: towards a reclassification of the *Phoma* complex. Mycol Res 113:508–19. doi: 10.1016/j.mycres.2009.01.002

Deshpande V, Wang Q, Greenfield P, Charleston M, Porras-Alfaro A, Kuske CR, Cole JR, Midgley DJ, Tran-Dinh N (2016) Fungal identification using a Bayesian classifier and the Warcup training set of internal transcribed spacer sequences. Mycologia 108:1–5. doi: 10.3852/14-293

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461. doi: 10.1093/bioinformatics/btq461

Flood J (2006) A review of fusarium wilt of oil palm caused by *Fusarium oxysporum* f. sp. *elaeidis*. Phytopathology 96:660–2. doi: 10.1094/PHYTO-96-0660

Gilbert J, Tekauz A (2000) Review: Recent developments in research on fusarium head blight of wheat in Canada. Can J Plant Pathol 22:1–8. doi: 10.1080/07060660009501155

Hawksworth DL (2001) The magnitude of fungal diversity: the 1.5 million species estimate revisited. Mycol Res 105:1422–1432. doi: 10.1017/S0953756201004725

Hollingsworth CR, Gray FA, Groose RW (2005) Evidence for the heritability of resistance to brown root rot of alfalfa, caused by *Phoma sclerotioides*. Can J Plant Pathol 27:64–70. doi: 10.1080/07060660509507195

Kõljalg U, Nilsson RH, Abarenkov K, Tedersoo L, Taylor AFS, Bahram M, Bates ST, Bruns TD, Bengtsson-Palme J, Callaghan TM, Douglas B, Drenkhan T, Eberhardt U, Dueñas M, Grebenc T, Griffith GW, Hartmann M, Kirk PM, Kohout P, Larsson E, Lindahl BD, Lücking R, Martín MP, Matheny PB, Nguyen NH, Niskanen T, Oja J, Peay KG, Peintner U, Peterson M, Põldmaa K, Saag L, Saar I, Schüßler A, Scott JA, Senés C, Smith ME, Suija A, Taylor DL, Telleria MT, Weiss M, Larsson K-H (2013) Towards a unified paradigm for sequence-based identification of fungi. Mol Ecol 22:5271–7. doi: 10.1111/mec.12481

Liang C, da C Jesus E, Duncan DS, Jackson RD, Tiedje JM, Balser TC (2012) Soil microbial communities under model biofuel cropping systems in southern Wisconsin, USA: Impact of crop species and soil properties. Appl Soil Ecol 54:24–31. doi: 10.1016/j.apsoil.2011.11.015

Liu K-L, Porras-Alfaro A, Kuske CR, Eichorst SA, Xie G (2012) Accurate, rapid taxonomic classification of fungal large-subunit rRNA genes. Appl Environ Microbiol 78:1523–33. doi: 10.1128/AEM.06826-11

Menkis A, Burokienė D, Gaitnieks T, Uotila A, Johannesson H, Rosling A, Finlay RD, Stenlid J, Vasaitis R (2012) Occurrence and impact of the root-rot biocontrol agent *Phlebiopsis gigantea* on soil fungal communities in *Picea abies* forests of northern Europe. FEMS Microbiol Ecol 81:438–45. doi: 10.1111/j.1574-6941.2012.01366.x

Miransari M, Bahrami HA, Rejali F, Malakouti MJ, Torabi H (2007) Using arbuscular mycorrhiza to reduce the stressful effects of soil compaction on corn (*Zea mays* L.) growth. Soil Biol Biochem 39:2014–2026. doi: 10.1016/j.soilbio.2007.02.017

MSCO (2013) 27 year summary of annual values for Gull Lake (3504) 1981-2010

O'Brien HE, Parrent JL, Jackson JA, Moncalvo J-M, Vilgalys R (2005) Fungal community analysis by large-scale sequencing of environmental samples. Appl Environ Microbiol 71:5544–50. doi: 10.1128/AEM.71.9.5544-5550.2005

Penton CR, Gupta VVSR, Yu J, Tiedje JM (2016) Size matters: assessing optimum soil sample size for fungal and bacterial community structure analyses using high throughput sequencing of rRNA gene amplicons. Front Microbiol 7:824. doi: 10.3389/fmicb.2016.00824

Penton CR, St Louis D, Cole JR, Luo Y, Wu L, Schuur EAG, Zhou J, Tiedje JM (2013) Fungal diversity in permafrost and tallgrass prairie soils under experimental warming conditions. Appl Environ Microbiol 79:7063–7072. doi: 10.1128/AEM.01702-13

Porras-Alfaro A, Liu K-L, Kuske CR, Xie G (2014) From genus to phylum: large-subunit and internal transcribed spacer rRNA operon regions show similar classification accuracies influenced by database composition. Appl Environ Microbiol 80:829–40. doi: 10.1128/AEM.02894-13

Rodríguez H, Fraga R (1999) Phosphate solubilizing bacteria and their role in plant growth promotion. Biotechnol Adv 17:319–339. doi: 10.1016/S0734-9750(99)00014-2

Sanford GR, Oates LG, Jasrotia P, Thelen KD, Robertson GP, Jackson RD (2016) Comparative productivity of alternative cellulosic bioenergy cropping systems in the North Central USA. Agric Ecosyst Environ 216:344–355. doi: 10.1016/j.agee.2015.10.018

Schöler A, Jacquiod S, Vestergaard G, Schulz S, Schloter M (2017) Analysis of soil microbial communities based on amplicon sequencing of marker genes. Biol Fertil Soils 53: 485–489. https://doi.org/10.1007/s00374-017-1205-1

Sul WJ, Cole JR, da C Jesus E, Wang Q, Farris RJ, Fish JA, Tiedje JM (2011) Bacterial community comparisons by taxonomy-supervised analysis independent of sequence alignment and clustering. Proc Natl Acad Sci U S A 108:14637–14642. doi: 10.1073/pnas.1111435108

Tedersoo L, Anslan S, Bahram M, Põlme S, Riit T, Liiv I, Kõljalg U, Kisand V, Nilsson H, Hildebrand F, Bork P, Abarenkov K (2015) Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. MycoKeys 10:1–43. doi: 10.3897/mycokeys.10.4852

Thangavelu R, Palaniswami A, Velazhahan R (2004) Mass production of *Trichoderma harzianum* for managing fusarium wilt of banana. Agric Ecosyst Environ 103:259–263. doi: 10.1016/j.agee.2003.09.026

Vandenkoornhuyse P, Baldauf SL, Leyval C, Straczek J, Young JPW (2002) Extensive fungal diversity in plant roots. Science 295:2051. doi: 10.1126/science.295.5562.2051

Voříšková J, Baldrian P (2013) Fungal community on decomposing leaf litter undergoes rapid successional changes. ISME J 7:477–86. doi: 10.1038/ismej.2012.116

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73:5261–5267. doi: 10.1128/AEM.00062-07

White T, Bruns T, Lee S, Taylor J (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis, M.A., Gelfand, D.H., Sninsky, J.J., White, T.J. (Eds.), PCR Protocols – A Guide to Methods and Applications. Academic Press, New York, pp. 315–322.

Xiong W, Zhao Q, Xue C, Xun W, Zhao J, Wu H, Li R, Shen Q (2016) Comparison of fungal community in black pepper-vanilla and vanilla monoculture systems associated with vanilla fusarium wilt disease. Front Microbiol 7:117. doi: 10.3389/fmicb.2016.00117

Xu L, Ravnskov S, Larsen J, Nilsson RH, Nicolaisen M (2012) Soil fungal community structure along a soil health gradient in pea fields examined using deep amplicon sequencing. Soil Biol Biochem 46:26–32. doi: 10.1016/j.soilbio.2011.11.010

Figures

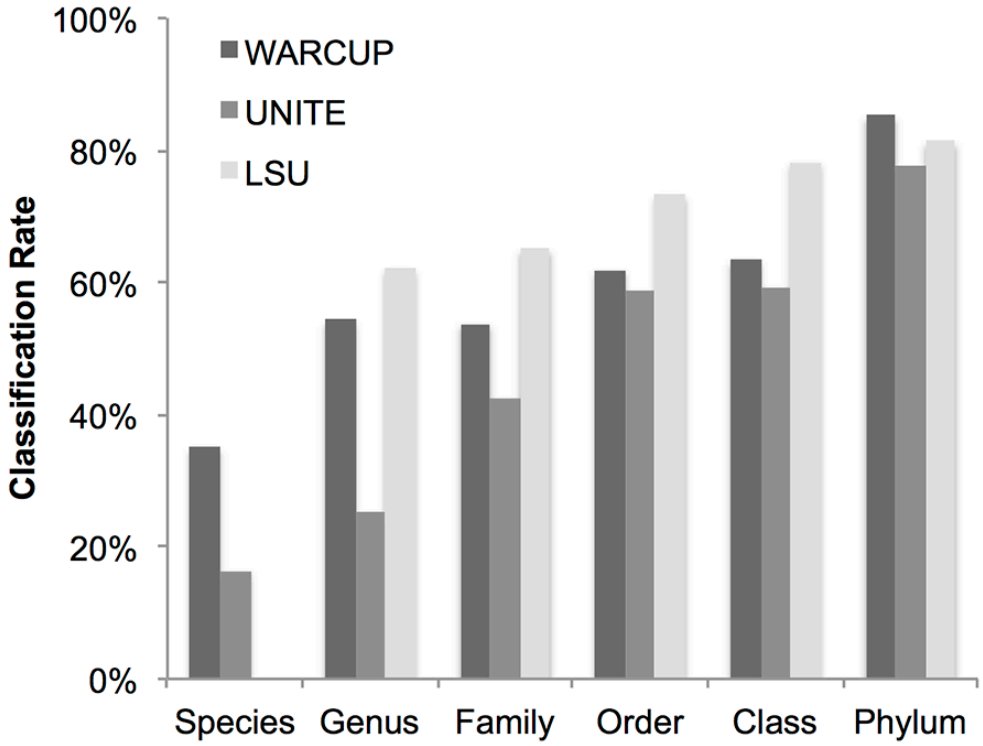


Fig. 1 Proportion of classified reads of LSU, UNITE, and Warcup classification schemes from phylum to species

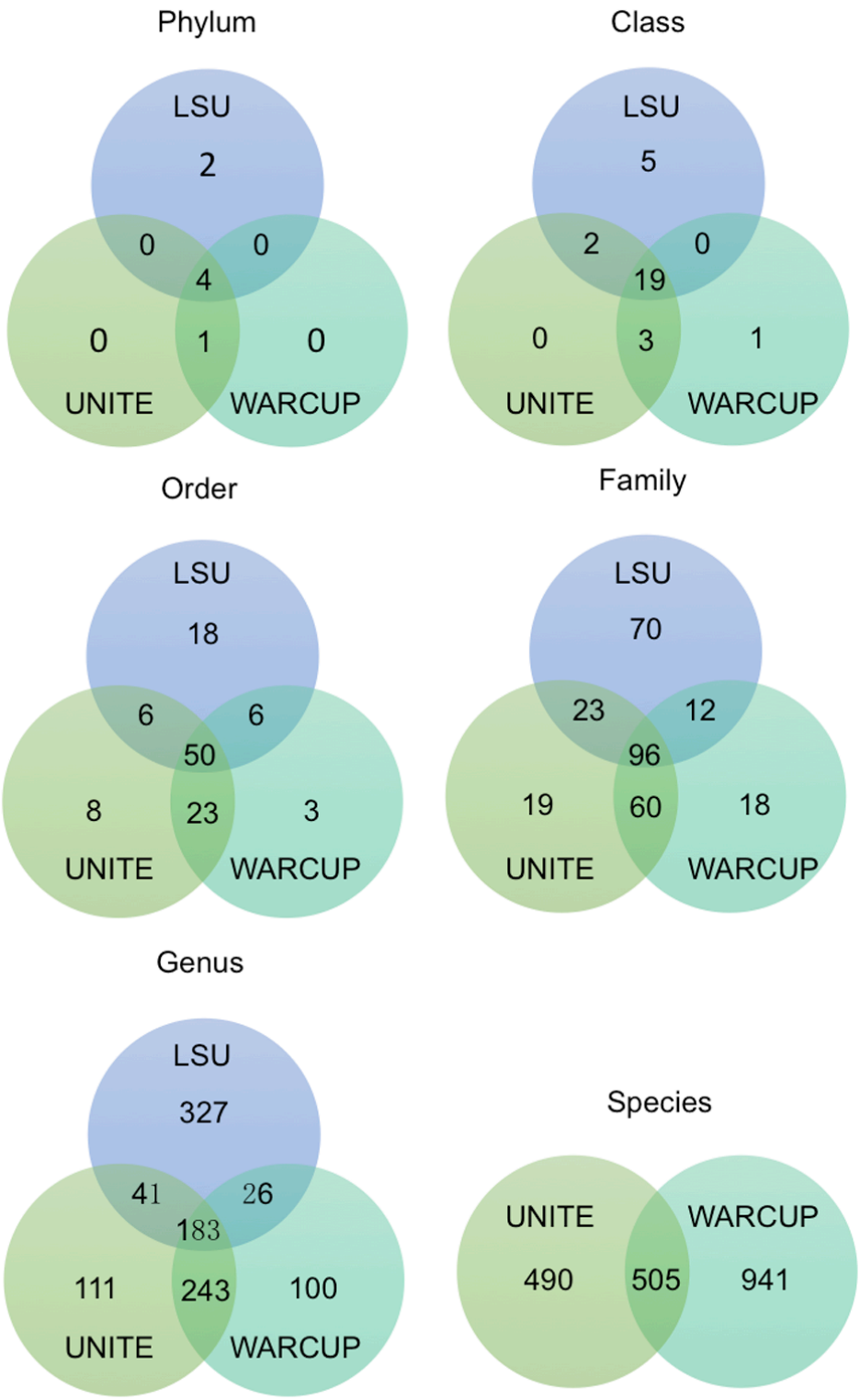


Fig. 2 Venn diagram showing shared taxa (unknown taxa removed) of UNITE, WARCUP, and LSU classification schemes from phylum to species

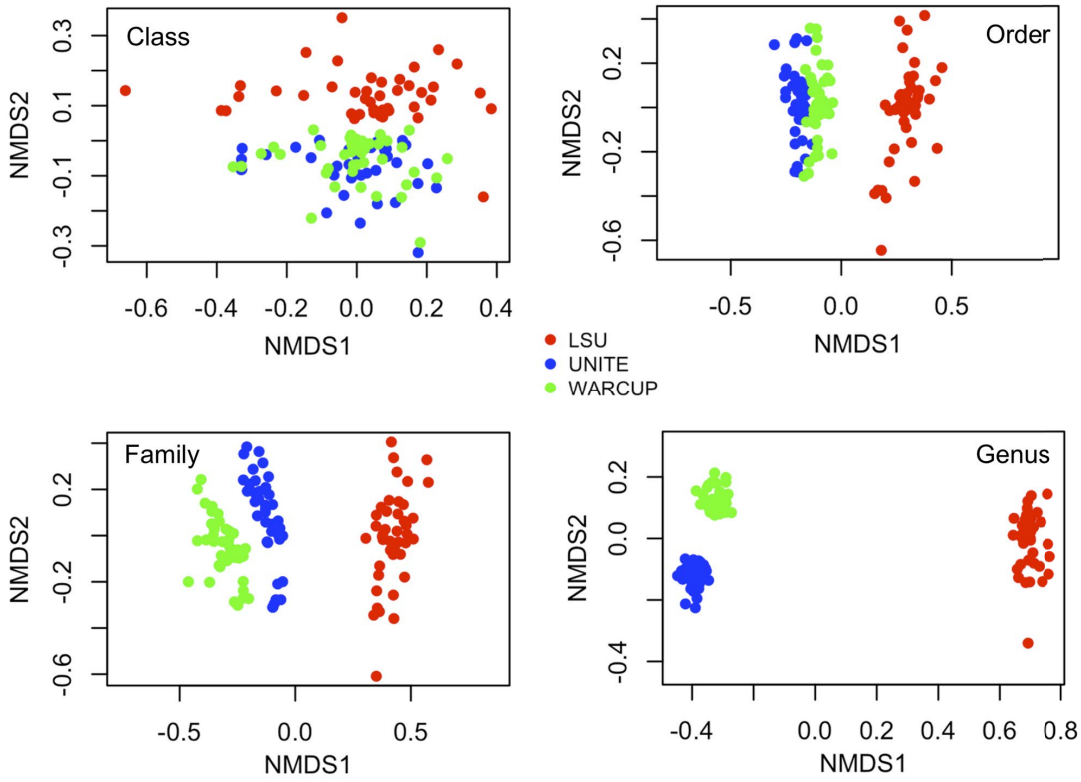


Fig. 3 Ordinations (NMDS) based on Bray-Curtis distance between samples of LSU, UNITE, and Warcup classification schemes together from class to genus levels

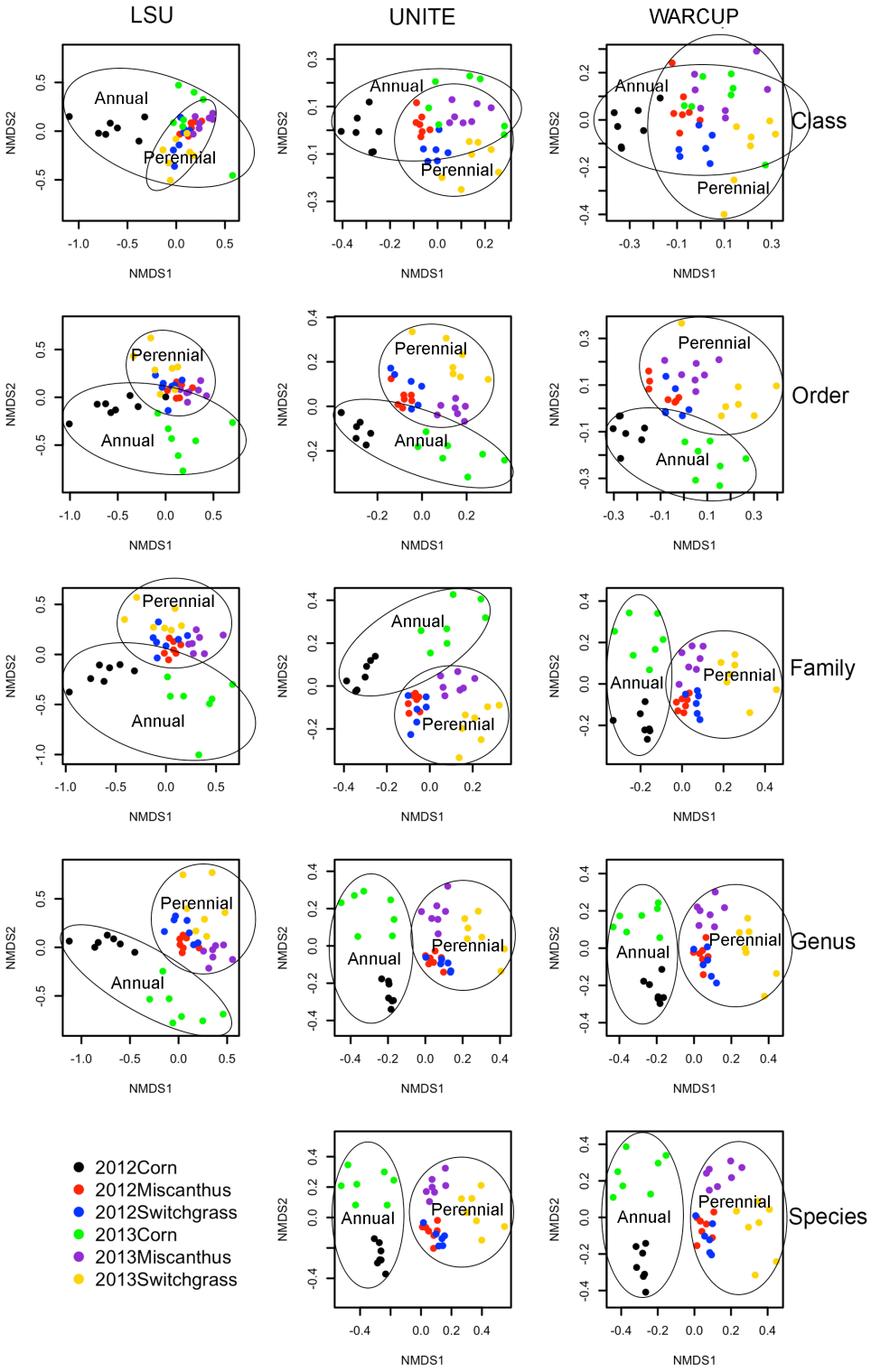
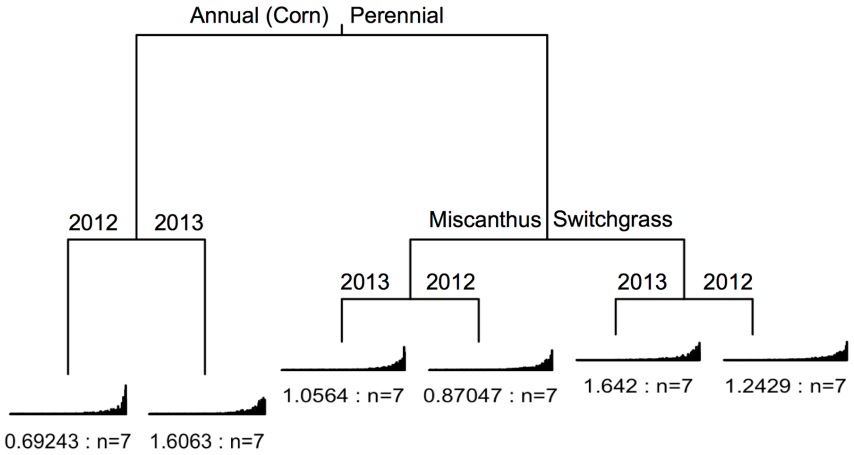
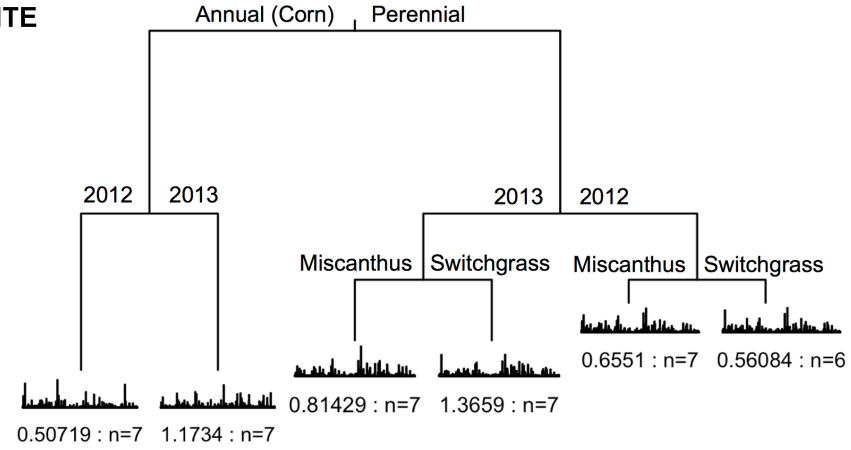


Fig 4 Ordinations (NMDS) based on Bray-Curtis distance between samples of LSU, UNITE, and Warcup classification schemes from class to species levels

LSU



UNITE



WARCUP

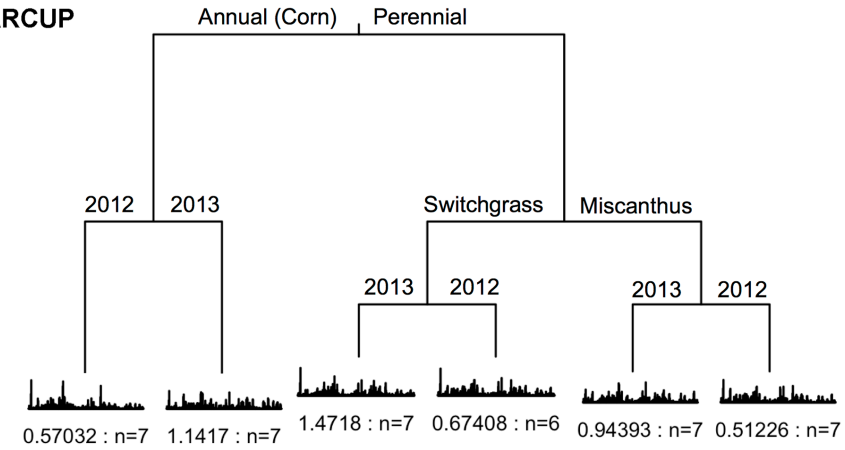


Fig 5 Multiple regression tree (MRT) of LSU, UNITE, and Warcup classification schemes at the genus level. The histograms show the distributions of the genera and the decimals are the deviances of the community calculated by MRT analysis

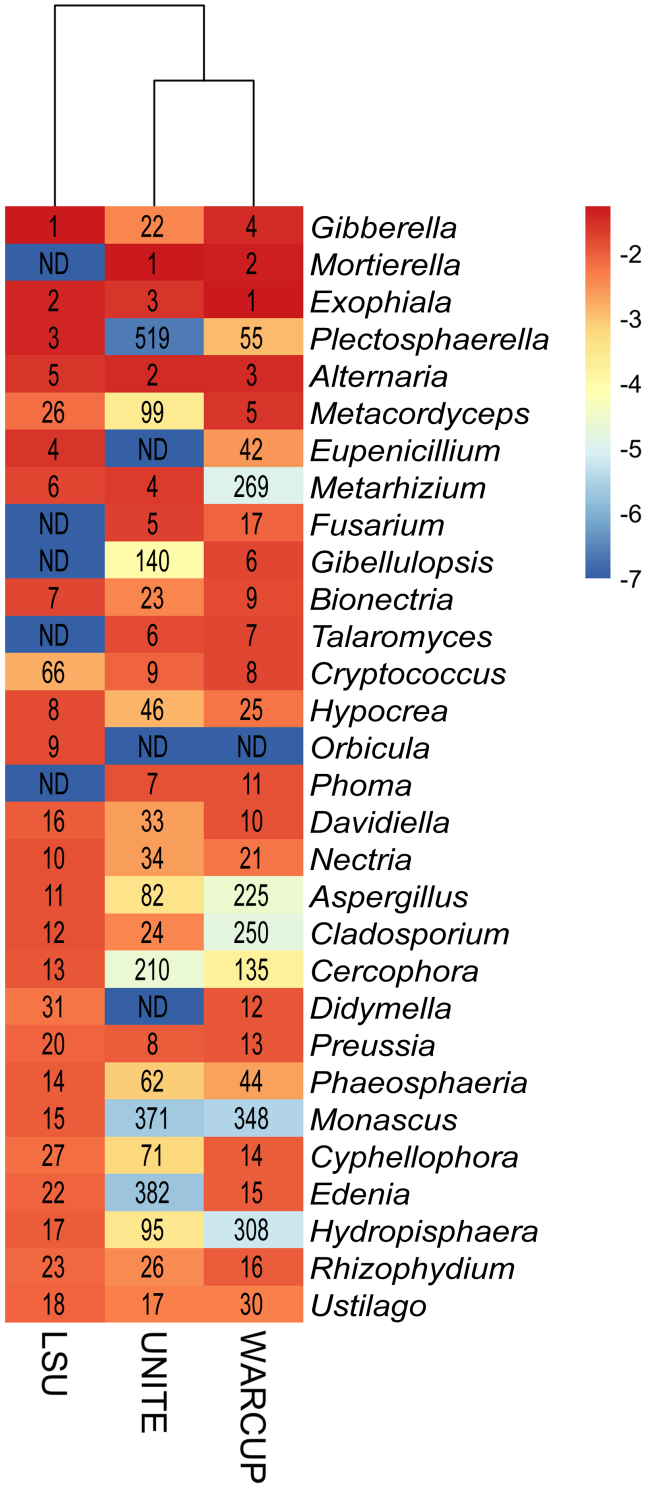


Fig. 6 Heatmap showing top 30 abundant fungal genera in LSU, UNITE, and Warcup classification schemes. Key on the right from blue to red represents the least abundant to most abundant genera. Numbers in cells represent the rank of the genus in the corresponding classification scheme. ND represent undetected in the corresponding classification scheme. The tree on the top is based on hclust

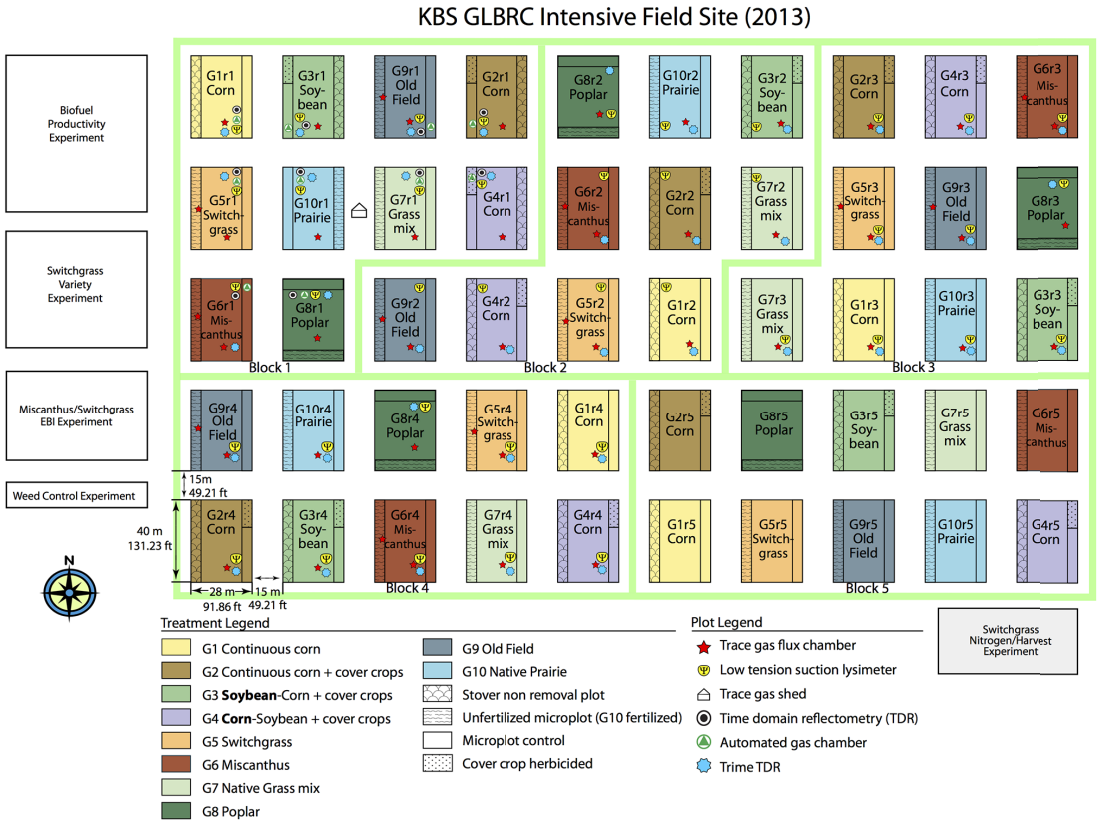


Fig. S1 Experiment design of long-term Great Lakes Biofuel Research Center (GLBRC) biofuel cropping system experiment at Kellogg Biological Station (KBS). http://lter.kbs.msu.edu/research/long-term-experiments/glbrc-intensive-experiment.

Tables

Table 1 Number of classified taxa from phylum to species levels

Classification scheme	Phylum	Class	Order	Family	Genus	Species
LSU	6 (8)	26 (29)	80 (90)	201 (230)	577 (586)	NA
UNITE	5 (6)	24 (33)	87 (116)	198 (292)	578 (732)	995 (2597)
Warcup	5 (5)	23 (29)	82 (93)	186 (211)	552 (552)	1446 (1551)

NA represents data not applicable.

Numbers represent number of classified taxa after removal of unknown taxa.

The numbers in brackets represent number of classified taxa with unknown taxa.

Table 2 Proportion of "unknown" taxa from phylum to species level of LSU, Warcup, and UNITE classification schemes

Classification scheme	Phylum (%)	Class (%)	Order (%)	Family (%)	Genus (%)	Species (%)
LSU	25.0 (0.6)	10.3 (NA)	11.11 (2.3)	12.6 (4.7)	1.88 (0.09)	NA
UNITE	16.7 (7.8)	27.3 (19.0)	25.9 (17.0)	32.2 (29.5)	21.0 (18.5)	61.8 (39.5)
Warcup	0 (0)	20.7 (5.3)	11.8 (3.57)	11.8 (6.5)	0 (0)	6.77 (2.7)

Numbers represent the proportion of unknown taxa among all identified, classified taxons.

The numbers in brackets represent the proportion of sequences of the unknown taxa.

0<NA<0.01 or NA represents data not applicable.

Table S1 Soil organic matter and pH in different treatments.

Treatment	Organic Matter(%)	pH
Corn-2012	3.39±0.22	6.09±0.3
Miscanthus-2012	3.76±0.48	6.11±0.48
Switchgrass-2012	3.87±0.36	6.24±0.09
Corn-2013	3.2±0.41	5.67±0.18
Miscanthus-2013	3.95±0.29	5.82±0.22
Switchgrass-2013	3.47±0.5	5.81±0.15

Table S2 PERMANOVA results for all databases indicated significant plant, time and crop*time interactions.

Database	Plant	Time	Plant*Time
LSU	F=8.24	F=8.02	F=3.27
	p<0.01	p<0.01	p<0.01
WARCUP	F=5.95	F=6.92	F=2.58
	p<0.01	p<0.01	p<0.01
UNITE	F=8.07	F=10.17	F=3.03
	p<0.01	p<0.01	p<0.01