Genomics Approaches to the Study of Diversity and Function of Aquatic Fungi

Inaugural-Dissertation
to obtain the academic degree
Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of Biology, Chemistry and Pharmacy
of Freie Universität Berlin

by

Felix Heeger

from Mettingen

2018

The PhD research presented here was conducted from 01.06.2014 to 31.05.2018 at the Leibniz-Institute of Freshwater Ecology and Inland Fisheries Berlin (IGB) and the Berlin Center for Genomics in Biodiversity Research (BeGenDiv). The work was financed by the SAW/PAKT program MycoLink of the Leibniz Association (WGL) and supervised by Dr. Michael T. Monaghan.

1 st Reviewer:	Dr. Michael T. Monaghan
2 nd Reviewer:	Prof. Dr. Anna Gorbushina

Date of Defense: 31. November 2018

Acknowledgements

I would like to thank everybody who has help to make this thesis reality.

First of all that is Michael Monaghan who was a wonderful supervisor, who gave me a lot of freedom to follow my many little site-projects that never made it into this thesis, but also always helped me to keep an eye in the overall goal.

I am thankful to Anna Gorbushina who agreed on short notice to be the second reviewer for this thesis.

None of this work would have been possible without the other participants of the MycoLink project. Liz Bourne who is, not by accident, an author on all three of the manuscripts in this thesis. Lars Ganzert who nearly fell into the Himmerlreichsee, Robert Taube who jumped into the Himmelreichsee and Christiane Baschien who taught me a lot about mycology.

Many thanks go to the people at the BeGenDiv, who let me make it “my” BeGenDiv. Camila Mazzoni who always asked the question that I hoped nobody would ask. Susan Mbedi who always had her office door open for me, when I needed to complain. Marcela Uliano who was always interested in every random thing I talked about. Paul Johnston who taught me a lot about RNA-Seq analysis and sarcasm. Max Driller who did bioinformatics support so that I did not have to. Christian Wurzbacher who taught me about metabarcoding and bottel gardens. And everybody else who made the BeGenDiv an awesome place.

And finally my family, my friends and all the people in my life that had my back when thing were not going well. Especially Cloelia, Mama, Papa, Leoni and Lukas.

Table of Contents

I Summary.....7

1 Summary in English.....7

2 Zusammenfassung in Deutsch.....8

II Introduction.....11

III 5.8S as a low variability complementary marker to ITS2 improves high level classifications of aquatic fungi.....23

1 Abstract.....23

2 Introduction.....24

3 Methods.....26

Classification Approaches.....26

Testing the effects of an incomplete reference database.....27

5.8S reference data set.....28

Description of the pipeline.....28

Test with reads from freshwater lake samples.....30

4 Results.....30

5 Discussion.....34

6 Author contribution.....36

7 References.....36

IV Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments.....41

1 Abstract.....41

2 Introduction.....41

3 Materials and Methods.....45

Isolates, Mock community, and Environmental samples.....45

DNA extraction.....46

PCR and chimera formation tests.....47

Library preparation and Sequencing.....48

Long-read metabarcoding pipeline.....49

Error rates based on isolate sequences....50
Evaluating chimera detection....51
Mock community classification....51
Environmental community classification....51
4 Results....52
Chimera formation and detection....53
Mock community classification....54
Environmental community classification....58
5 Discussion....60
Challenges of long reads....61
Classification....63
Classification conflicts and synergies....65
Conclusions....66
6 Acknowledgements....66
7 Author Contribution....67
8 References....67
V Identification of enzymes for lignocellulose degradation in *Clavariopsis aquatica*....73
1 Abstract....73
2 Introduction....73
3 Methods....75
Cultivation....75
Genome Sequencing....76
RNA Sequencing....76
Genome Assembly....77
Genome Annotation....77
Differential Expression and MGSA Analysis....78
4 Results....79
Genome Assembly and Annotation....79
RNA-Sequencing and Differential Expression....80
Gene Set Activation....82
5 Discussion....88

6 Acknowledgments..90
7 Author Contribution...90
8 References...91
VI Discussion...97
VII Appendix 1...105
1 Supplemental Figure 1...105
2 Supplemental Figure 2...105
3 Supplemental Figure 3...106
4 Supplemental Figure 4...106
5 Supplemental Figure 5...107
6 Supplemental Figure 6...107
7 Supplemental Figure 7...108
8 Supplemental Figure 8...108
9 Supplemental Figure 9...109
VIII Appendix 2...111
1 Supplemental Info 1: Pipeline steps..111
2 Supplemental Table 1...115
IX Appendix 3..117
1 Supplemental Info 1...117
Supplemental Table 1..118
2 Supplemental Table 2...121
3 Supplemental Table 3...125

I Summary

1 Summary in English

The kingdom of fungi comprises an enormous range of live styles and genetic variability. Different genomics approaches offer possibilities to investigate species diversity and ecological function of fungi. In this thesis I present improvements of metabarcoding methods for aquatic fungi and the application of whole genome sequencing and transcriptome sequencing to an exclusively aquatic fungus.

Beside the standard metabarcoding marker for fungi, the ITS (internal transcribed spacer) region, the eukaryotic rRNA operon contains two other markers, the SSU (small subunit) and LSU (large subunit), that are also often used for metabarcoding. When choosing a metabarcode there is a trade-off between high variability for fine grain species delineation and high conservation for good primer binding and high level classification of novel species, which are not represented in reference databases. In the work presented in chapter III, we investigated the possibility to use the information from the more conserved 5.8S sequence, that is part of many amplicons used for ITS2 sequencing. It is normally discarded, but we used it as a complementary marker to ITS2 and showed that it can improve classification of novel species with an incomplete reference database. In chapter IV this is taken one step further by using third generation sequencing to sequence the full ITS region together with the more conserved SSU and LSU in the same amplicon. This gives us the option to use different markers with different databases for classification in parallel and to circumvent the trade-off between high variability and high conservation.

Fungi are ecologically very important decomposers of lignocellulose from plant biomass. The occurrence and expression of gene families for the degradation of lignin from lignocellulose has been extensively studied with whole genome and transcriptome sequencing in terrestrial, but not in aquatic fungi. In the work presented in chapter V, we used whole genome and transcriptome sequencing to investigate differential gene expression in the exclusively aquatic fungus *Clavariopsis aquatica* when grown on media with more and less lignin rich carbon sources and investigated the expression patterns of peroxidases, laccases and other protein families involved in plant biomass degradation. This observed up-regulation of laccases, peroxidases and genes from the cytochrome P450 super-family, as well as other gene families involved in cellulose and

hemicellulose degradation, strongly suggests that *C. aquatica* is able to modify lignin to some extent; perhaps in order to facilitate the utilization of lignocellulose as a carbon and energy source.

2 Zusammenfassung in Deutsch

Das Königreich der Pilze enthält eine enorme Bandbreite von Lebensweisen und genetischer Variabilität. Verschiedene genomische Ansätze bieten die Möglichkeit die Artenvielfalt und ökologisch Funktion von Pilzen zu untersuchen. In dieser Doktorarbeit präsentiere ich die Verbesserung von Metabarcodingmethoden für aquatische Pilze und die Anwendung von Ganzgenomsequenzierung und Transkriptomsequenzierung eines exklusive aquatischen Pilzes.

Neben dem Standard-Marker für Pilze, der ITS-Region (internal transcribed spacer), enthält das eukaryotische rRNA-Operon zwei andere Marker, die SSU (smal subunit) und die LSU (large subunit), die auch als Metabarcodingmarker verwendet werden. Bei der Auswahl eines Metabarcodingmarkers gibt es einen Abwägung zwischen hoher Variabilität zu fein abgestuften Speziesunterscheidung und hoher Konservierung für gute Primerbindung und zur Klassifizierung von neuen Arten, die nicht in der Referenzdatenbank vertreten sind. In der Arbeit, die in Kapitel II vorgestellt wird, untersuchten wir die Möglichkeit die Information des stärker konservierten 5.8S Gens, das Teil vieler Amplikons ist, die zur ITS2-Sequenzierung verwendet werden, zu verwenden. Er wird wird normalerweise verworfen, aber wir verwendeten es als ergänzenden Marker zu ITS2 und konnten zeigen, dass es die Klassifizierung von neuen Arten verbessert, wenn die Referenzdatenbank unvollständig ist. In Kapitel III wird dies einen Schritte weiter getrieben indem Sequenzierung der dritten Genration genutzt wurde um die komplette ITS-Region zusammen mit den stärker konservierten SSU und LSU in einem Amplikon zu sequenzieren. Dies eröffnete uns die Möglichkeit verschiedene Marker mit verschiedenen Datenbanken parallel zu verwenden und die Abwägung zwischen hoher Variabilität und hoher Konservierung zu umgehen.

Pilze sind sind ökologisch sehr wichtig für den Abbau von Lignozellulose aus pflanzlicher Biomasse. Das Vorkommen und die Expression von Genfamilien für den Abbau von Lignin aus Lignozellulose wurde in terrestrischen Pilzen bereits umfangreich mit Genom- und Transkriptomsequenzierung untersucht, jedoch nicht in aquatischen Pilzen. In der Arbeit, die in Kapitel V vorgestellt wird, verwendeten wir Genom- und Transkriptomesequenzierung um in dem exklusiv aquatischen Pilz *Clavariopsis aquatica*, der auf Substraten mit mehr oder weniger Lignin kultiviert wurde, die Expressionsmuster von Peroxidasen, Laccasen und anderen Genfamilien, die

am Abbau von pflanzlicher Biomasse beteiligt sind, zu untersuchen. Die beobachtete Hochregulierung von Laccasen, Peroxidasen, Genen der Cytochrome P450 Super-Familie und weiter Genfamilien, die am Abbau von Zellulose und Hemizellulose beteiligt sind, deutet stark darauf hin, dass *C. aquatica* in der Lage ist Lignin zu einem gewissen Grad zu modifizieren; möglicherweise um die Verwendung von Zellulose und Hemizellulose als Energie- und Kohlenstoffquelle zu ermöglichen.

II Introduction

The kingdom of fungi comprises a enormous range of live styles and genetic variability. Their diversity in terms of existing species as well as function in ecosystems is far from fully explored. Different genomics approaches offer possibilities to investigate different aspects of fungal diversity. Metabarcoding can be used to identify different fungi from the environment, whole genome sequencing and transcriptome sequencing to investigate the enzymatic abilities of a fungus and its reaction to changes in the environment. In this thesis I present improvements of metabarcoding methods for fungi and the application of whole genome sequencing and transcriptome sequencing. The focus of the work is on fungi that appear in aquatic habitats, that are especially poorly characterized (Grossart and Rojas-Jimenez, 2016). Fungi that spent significant part of their life cycle submerged in water are called aquatic fungi, but this group is not monophyletic (Shearer et al., 2009) and in practice in many cases it will be not possible to distinguish them from fungi that were washed or blown into the water and do not actually grow there.

Fungi can be found all over the plant and often occur in harsh conditions like high radiation (Dadachova and Casadevall, 2008) or salinity (Vaupotic et al., 2008). Their live styles reach from parasitic and pathogenic, over degradation of dead biomass to symbiotic relationships with other organisms. Many fungi can opportunistically take more than one of these roles. About 8,000 fungi are known plant pathogens (Nature Microbiology Editorial, 2017) like rusts, smuts and rots and a number of human diseases, especially in immunocompromised patients, are caused by fungi. Fungi can also be found as parasites of insects, nematodes and even other fungi.

One of the most important form of fungal symbiosis is mycorrhiza, where fungi grow in symbiosis with plant roots and benefit the plants nutrient uptake, while getting energy in form of sugar from the plant. Another form of fungal symbiosis are lichens, that are symbiotic communities of fungi and algae or cyanobacteria.

Fungi also play an important role as decomposers of dead biomass. Especially of recalcitrant materials like wood.

The number of fungal species is topic of debate and estimations range from 1.5 to 5 million (Blackwell, 2011; Hawksworth, 1991). New species are discovered frequently even among the macroscopic mushrooms (e.g. Chakraborty et al., 2018; Tibpromma et al., 2018; Vizzini et al., 2018; Wang et al., 2018). Even the high level evolutionary relationships between fungi are not fully resolved and tree underlying the taxonomic classification of fungi is still very much in flux. For example over the course of the last twenty years the number of suggested phyla inside the kingdom of fungi has changed from seven (Hibbett et al., 2007) to twelve (Tedersoo et al., 2017a). This includes not only rearrangement of known orders into new phyla, but also the introduction of the phylum Cryptomycota (also known as Rozellomycota) (Jones et al., 2011a) in which few species have been named so far, but that is believed to contain substantial genetic diversity (Jones et al., 2011b).

Cells of vascular plants have a cell wall that protects them from the outside and allows them to keep up high osmotic gradients. These cell walls consist of a interwoven matrix of different carbohydrates. The main components are cellulose, hemicellulose and lignin. Cellulose is a linear polymer of D-glucose monomers and makes up the biggest proportion of the cell wall. Hemicellulose is the collective name for different polymers of different sugars like xylose, galactose, mannose and others. Unlike cellulose the structure of hemicellulose is much more random with shorter chain length and branching chains. Lignin is a polymer of different phenolic monomers, that is hydrophobic and very resistant to biodegradation.

Fungi are the only organisms that can degrade lignin to access cellulose and hemicellulose in wood and other plant material. Wood decay by fungi is often separated into to categories. “White rot” in which the lignin is fully degraded and white cellulose is exposed and “brown rot” in which the lignin is modified making it possible for the fungus to degrade the cellulose and hemicellulose and leaving the brown lignin behind. Lignin degradation in white rot is mostly facilitated by peroxidases like lignin peroxidase, manganese peroxidase and versatile peroxidase. In addition other lignin modifying enzymes like laccase play an important role.

Many fungi species are microscopic and hard to classify by morphological features. Traditionally fungi were identified and classified by their fruiting bodies and spores during sexual reproduction. Fungi that do not sexually reproduce or where sexual reproduction could not be observed were summarized in a separated phylum called Deuteromycota. In fact the International Code of Botanical Nomenclature (ICBN) that governs the naming of fungi specifically allowed to give different names to the sexual reproductive stage (teleomorph) and the asexual stage (anamorph) of the same fungus. With the development of the polymerase chain reaction (PCR) and DNA sequencing, studies of molecular phylogeny of fungi became possible and showed the Deuteromycota did not form monophyletic clade and that asexual species could be nested in sexually reproducing genera (Berbee and Taylor, 1992). This led to the change of the ICBN to no longer allow two names for one fungus in 2011 (International Association for Plant Taxonomy, 2012).

Besides being useful for computing molecular phylogenetic trees, DNA sequencing and amplification of defined stretches of DNA by PCR also allowed for identification of species by short, distinctive parts of the genome. These species markers, called barcodes, are often more informative than morphological or functional features and can be more easily tested in modern molecular labs. For barcoding to work one needs a barcode (or marker) sequence, that can be amplified from all species in question and is different for each species. For amplification there are two constraints. Firstly the barcode gene or region must exist in all species in question. Secondly to allow the design of primers that can bind for all studied species, the sequence of the barcode or at least the sequence of flanking regions need to be conserved to a certain degree. For the barcode to be different in every species one must choose a region that is not conserved between species and ideally is not under stabilizing selective pressure. Obviously these two requirements, conservation for primer binding and variability for species delineation, are in direct conflict. Either a trade off between them has to be made when choosing a barcode or a sequence region can be found that has a well conserved structure with regions of high sequence conservation flanking a region of high variability.

Like in bacteria the gene encoding the rRNA for the small ribosomal subunit (SSU or 18S) is used as a marker in fungi. It is well conserved in all cellular organisms, but contains multiple regions (called V1-V9) of higher variability. Although the variable regions show higher variability than the rest of the gene they can not always resolve differences between closely related species (Cole et al., 2014; Schoch et al., 2012). In most eukaryotes the SSU appears in an operon together with the two genes for the rRNAs in the large ribosomal subunit, the 5.8S and the 28S (or LSU). Between the SSU and the 5.8S, and between the 5.8S and the LSU genes, there are the two itergenic spacers ITS1 and ITS2. These two spacers together with the 5.8S form the ITS region, that separates the SSU and LSU. Since the two spacers are non-coding they show high variability. This fact together with the conserved flanking regions of SSU and LSU makes the ITS region a good marker for fungi. In 2012 (Schoch et al., 2012) the ITS region was chosen as the official barcode for fungi, but the SSU and the LSU are still often used (e.g. Davison et al., 2015; Jumpponen et al., 2015; Rojas-Jimenez et al., 2017; Roy et al., 2017).

When using a DNA barcode for identification of more than a few species the need for a reference database to compare to arises. General purpose sequence databases like GenBank can be used, but often lack a rigorous enough curation and thus may contain errors and especially sequences that are assigned to the wrong species. Because of its frequent use in bacteria several specialized databases with SSU sequences have been established. The Ribosomal Database Project (RDP, Cole et al., 2014) database and the Greengenes (DeSantis et al., 2006) database both contain SSU sequences from bacteria and archaea but not from fungi. The SILVA database (Quast et al., 2013) contains SSU sequences from bacteria, archaea and fungi. In addition the RDP also provides a specialized LSU database with fungal sequences and SILVA has a specialized LSU database for bacteria, archaea and fungi. For fungi there is the UNITE database (Kõljalg et al., 2013) with ITS region sequences. In this work UNITE was chosen as reference for ITS sequences and SILVA as a reference for SSU sequences, since they are the only database with fungal sequences for the respective marker. For LSU the RDP database was chosen as reference since they provide a better curated dataset with less length differences.

Originally barcoding was done with Sanger sequencing. It provides sequences of a length between 300-1000 bp. One of its limitations is that only one sequence can be present in the sample that is sequenced. A mix of DNA molecules with different sequences will lead to mixed base signals in sequencing and low quality sequence reads. Consequently investigated samples should only contain one species or DNA molecules from different species have to be separated. This means that barcoding all species in an environmental sample (so called metabarcoding) with sanger sequencing has to includes a time consuming cloning step and thus is not feasible for more than a few samples and few dozen species per sample.

Second generation sequencing (e.g. Illumina/Solexa and Roche 454) offers the possibility to sequence millions of molecules from the same sample. In addition by adding an index sequence to the DNA molecules of each sample during sequencing preparation, multiple samples can be sequenced at once and separated by the index sequence in post processing to be analyzed independently. This advance in sequencing technology made it possible to easily barcode thousands of species from one environmental sample. Especially Illumina sequencing with very low cost per base pair sequenced, allows for large scale ecological studies. The drawback is the shorter sequence read length (100 – 300 bp depending on the instrument). The sequence length that can be read can be improved by an overlapping paired end design. A DNA molecule is sequenced with a given read length from both sides and the length ratio between DNA molecule and reads is chosen such that the two reads overlap in middle. Given the maximum read length of 300 bp and a minimal overlap length to make sure that the two reads can be correctly joined, molecules up to 550 bp can be sequenced in this way. This reduced sequence length means that the full ITS region (300 to 1,200 bp) can not be used. Because most of the variability in the ITS region comes from the ITS1 and ITS2, while the 5.8S is more conserved, most studies use an amplicon containing the ITS1 or ITS2 . The other drawback of second generation sequencing is a slightly higher error rate (0.1%, (Goodwin et al., 2016)) compared to Sanger sequencing. Together with real intra-species variation, that is more visible because of the much higher read number, this means that a certain amount of sequence variability in the reads has be expected. To deal with this fact reads are normally clustered into so called operational taxonomic units (OTUs). Of course ideally one OTU would correspond to one species, but in practice this can not always be garantied.

With high throughput ecological studies the probability to encounter unknown species increases. This is especially true for fungi where many species are not formally described and even fewer are represented by sequences in the reference databases. Species that are unknown or not represented in the database can not be identified by barcoding. In most cases a classification is still attempted for these sequences by comparing them to known sequences and assigning them to a higher taxonomic rank than species. This classification can be based on sequence comparison to the database or on phylogenetic analysis.

One of the most common database search based approaches is the Naive Bayesian classifier implemented by th RDP project (Wang et al., 2007). It pre-processes the database by investigating the occurrence of subsequences of length k (kmers) in every reference sequence in each taxonomic group. When a query sequence is classified, the kmers in it, are compared to the ones in the database and from the result a likelihood for the sequence to come form a certain taxonomic group is computed. This process is repeated with different subsets of all possible kmers in the query sequence to give a bootstrap value for the most likely assignment.

Another database search based method is the lowest common ancestor (LCA) approach (Huson et al., 2007). First the query sequence is compared to the reference database by any alignment method (blast search in most applications). A set of “good” hits is determined by certain requirements to the alignment. From the taxonomic assignment of all these hits a classification is determined by finding the lowest common ancestor in the tree that underlies the taxonomy.

Phylogenetic classification approaches rely on a multiple sequence alignment of sequences with known taxonomic assignment (database sequences) and sequences with unknown taxonomic assignment (query sequences). From the alignment a phylogenetic tree is computed. Query sequences can then be assigned to the same taxonomic group as the database sequences they form a monophyletic group with.

In this work database search based approaches are applied, mainly because the standard barcode for fungi, the ITS region is too variable between all fungi to be properly aligned. Its high variability also causes problems for database search based classification approaches. If a sequence from a new species is to different from all the database sequences, it can be hard to determine to which of them it is the closest. Besides the problem of conservation for primer binding site, this adds another drawback to high variability barcodes like the ITS1 and ITS2. Barcodes with lower variability may offer the possibility to classify a sequence at least to a higher taxonomic rank, when an identification to species level is not possible.

There is trade off between alignable, more conserved barcodes, that can be used for phylogenetic analysis and higher level taxonomic assignment of new species and high variability barcodes that give the possibility to identify sequences to the species level. One idea to get around this is to sequence one marker from the first category and one from the second category at the same time. The easiest would be to amplify and sequence them as one sequence to keep the information that they are from the same individual. The organization of the rRNA operon offers the opportunity to sequence the ITS (as variable marker) together with parts of either the SSU or LSU (as more conserved marker). When using Illumina sequencing this is prevented by the short read length. Third generation sequencing technologies from Pacific Biosciences (PacBio) and Oxford Nanopore offer significantly longer reads (>30 kbp) with reasonably high throughput. Their drawback lies in higher error rates (>10%) that make them difficult to use for barcoding.

For PacBio sequencing hairpin adapters are added to the double stranded DNA molecules, so that they essentially form a single strand loop. If the original DNA molecule was short enough this loop might be read multiple times around. This multiple reads from the same molecule can be aligned and consensus sequence can be computed. The error rate in the resulting sequences is comparable to Illumina sequencing (Travers et al., 2010) and thus low enough to make barcoding viable. This approach has shown promising results for bacteria (Franzén et al., 2015; Schloss et al., 2016; Singer et al., 2016) where the full length 16S gave better results than the partial sequences normally used. For fungi Terdesoo et al. (Tedersoo et al., 2017b) sequenced parts of the rRNA operon and found increased taxonomic resolution compared to using only the ITS1 or ITS2.

Existing metabarcodes can not solve the trade-off between high variability and high conservation and thus can either not reliably classify sequences to the species level or not reliably classify novel species to any level. In the work presented in chapter III we investigated the possibility to use the information from the partial 5.8S sequence, that is part of many amplicons used for ITS2 sequencing. It is normally discarded, but we used it as a complementary marker to ITS2 and showed that it can improve classification of novel species with an incomplete reference database. In chapter IV this is taken one step further by using third generation sequencing to sequence the full ITS region together with the SSU and a big part of the LSU in the same amplicon. This gives us the option to use different markers with different databases for classification in parallel and to circumvent the trade-off between high variability and high conservation.

Whole genome sequencing and mRNA sequencing (RNA-Seq) are two other methods, that became only feasible with the development of second generation sequencing, that makes it possible to generate hundreds of millions of reads at a reasonable price.

For genome squencing the DNA of an organism is fragmented into smaller pieces of a few hundred base pairs and sequenced with second generation sequencing. The reads are then combine into so called contigs (continues sequences) by specialized assembly algorithms. The contigs can ideally be as long as the original DNA molecules, but are in practice often shorter. How successful the assembly is, depends on many factors, but mostly is a question of how well the genome is covered by the reads. The assembly algorithms rely on overlaps between the reads to combine them, and on redundancy in the reads to correct sequencing errors. This means that every base in the genome has to be covered by multiple reads (ideally >30).

From the genome sequence genes and proteins can be predicted with computational methods. These methods use statistical models of gene structures and alignments of known proteins and transcripts from the same or closely related organism to predict the positions of exons, introns, and start and stop codons. Although these predictions are not completely reliable, protein sequences can be inferred based on them. To get a insight into possible functions, predicted proteins are compared to databases of proteins and protein domains with known functions. Because the annotations produced in this way are not very reliable the analysis often focuses on groups of genes with similar functions or in a functional pathway, with the assumption that presence of multiple genes from one group, gives a better signal than single genes.

RNA-Seq is used to study expression of genes under certain conditions instead of only looking at the static genome of a species. RNA that is extracted from tissue or a cell culture that was grown in the condition, is reverse transcribed and fragmented. The fragments are than sequenced and the reads are aligned to a reference genome or transcriptome to estimate the number of transcripts that were present in the original sample. Because the absolute number of reads is dependent on many factors and difficult to compare between genes, normally RNA from a treatment and a control condition are compared and the differential expression is analyzed. Because the expression level between genes can vary strongly a high number (10 – 30 millions) of reads is necessary to also capture expression of genes with a low number of transcripts.

The interpretation of the results is difficult especially for non-model organisms, because the function of genes is mostly unknown and functional annotation is difficult (see above).

The proteins present in a genome and especially their up- or down-regulation under certain circumstances can be used to get insight into interaction of organisms with their environment and their role in the ecosystem. For example different research questions concerning the role of fungi as pathogens (e.g. Dobon et al., 2016; Galidevara et al., 2016), symbionts (Joneson et al., 2011; e.g. Perotto et al., 2014) have been studied with RNA-Seq. One question that has gotten considerable attention, because of its ecological importance as well as the potential commercial applications in biofuel productions, is the role of fungi as degraders of biomass (e.g. Ries et al., 2013; Yang et al., 2012).

In the chapter V we used whole genome sequencing and RNA-Seq to investigate the differential expression of the aquatic fungus *Clavariopsis aquatica* when grown on media with more and less recalcitrant carbon sources and investigated the expression patterns of peroxidases, laccases and other protein families involved in plant biomass degradation.

References

Berbee, M.L., and Taylor, J.W. (1992). 18S Ribosomal RNA gene sequence characters place the human pathogenSporothrix schenckii in the genusOphiostoma. Experimental Mycology *16*, 87–91.

Blackwell, M. (2011). The Fungi: 1, 2, 3 … 5.1 million species? American Journal of Botany *98*, 426–438.

Chakraborty, D., Vizzini, A., and Das, K. (2018). Two new species and one new record of the genus Tylopilus (Boletaceae) from Indian Himalaya with morphological details and phylogenetic estimations. MycoKeys *33*, 103–124.

Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., and Tiedje, J.M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res *42*, D633–D642.

Dadachova, E., and Casadevall, A. (2008). Ionizing Radiation: how fungi cope, adapt, and exploit with the help of melanin. Curr Opin Microbiol *11*, 525–531.

Davison, J., Moora, M., Öpik, M., Adholeya, A., Ainsaar, L., Bâ, A., Burla, S., Diedhiou, A.G., Hiiesalu, I., Jairus, T., et al. (2015). Global assessment of arbuscular mycorrhizal fungus diversity reveals very low endemism. Science *349*, 970–973.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G.L. (2006). Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl. Environ. Microbiol. *72*, 5069–5072.

Dobon, A., Bunting, D.C.E., Cabrera-Quio, L.E., Uauy, C., and Saunders, D.G.O. (2016). The host-pathogen interaction between wheat and yellow rust induces temporally coordinated waves of gene expression. BMC Genomics *17*.

Franzén, O., Hu, J., Bao, X., Itzkowitz, S.H., Peter, I., and Bashir, A. (2015). Improved OTU-picking using long-read 16S rRNA gene amplicon sequencing and generic hierarchical clustering. Microbiome *3*.

Galidevara, S., Reineke, A., and Koduru, U.D. (2016). In vivo expression of genes in the entomopathogenic fungus Beauveria bassiana during infection of lepidopteran larvae. J. Invertebr. Pathol. *136*, 32–34.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics *17*, 333–351.

Grossart, H.-P., and Rojas-Jimenez, K. (2016). Aquatic fungi: targeting the forgotten in microbial ecology. Current Opinion in Microbiology *31*, 140–145.

Hawksworth, D.L. (1991). The fungal dimension of biodiversity: magnitude, significance, and conservation. Mycological Research *95*, 641–655.

Hibbett, D.S., Binder, M., Bischoff, J.F., Blackwell, M., Cannon, P.F., Eriksson, O.E., Huhndorf, S., James, T., Kirk, P.M., Lücking, R., et al. (2007). A higher-level phylogenetic classification of the Fungi. Mycological Research *111*, 509–547.

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007). MEGAN analysis of metagenomic data. Genome Res. *17*, 377–386.

International Association for Plant Taxonomy (2012). International code of nomenclature for algae, fungi and plants (Melbourne code): adopted by the Eighteenth International Botanical Congress Melbourne, Australia, July 2011 (Königstein, Germany: Koeltz Scientific Books).

Jones, M.D.M., Richards, T.A., Hawksworth, D.L., and Bass, D. (2011a). Validation and justification of the phylum name Cryptomycota phyl. nov. IMA Fungus *2*, 173–175.

Jones, M.D.M., Forn, I., Gadelha, C., Egan, M.J., Bass, D., Massana, R., and Richards, T.A. (2011b). Discovery of novel intermediate forms redefines the fungal tree of life. Nature *474*, 200–203.

Joneson, S., Armaleo, D., and Lutzoni, F. (2011). Fungal and algal gene expression in early developmental stages of lichen-symbiosis. Mycologia *103*, 291–306.

Jumpponen, A., Brown, S.P., Trappe, J.M., Cázares, E., and Strömmer, R. (2015). Analyses of Sporocarps, Morphotyped Ectomycorrhizae, Environmental ITS and LSU Sequences Identify Common Genera that Occur at a Periglacial Site. Journal of Fungi *1*, 76–93.

Kõljalg, U., Nilsson R. Henrik, Abarenkov Kessy, Tedersoo Leho, Taylor Andy F. S., Bahram Mohammad, Bates Scott T., Bruns Thomas D., Bengtsson Palme Johan, Callaghan Tony M., et al. (2013). Towards a unified paradigm for sequence based identification of fungi. Molecular Ecology *22*, 5271–5277.

Nature Microbiology Editorial (2017). Stop neglecting fungi. Nature Microbiology *2*, 17120.

Perotto, S., Rodda, M., Benetti, A., Sillo, F., Ercole, E., Rodda, M., Girlanda, M., Murat, C., and Balestrini, R. (2014). Gene expression in mycorrhizal orchid protocorms suggests a friendly plant–fungus relationship. Planta *239*, 1337–1349.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res *41*, D590–D596.

Ries, L., Pullan, S.T., Delmas, S., Malla, S., Blythe, M.J., and Archer, D.B. (2013). Genome-wide transcriptional response of Trichoderma reesei to lignocellulose using RNA sequencing and comparison with Aspergillus niger. BMC Genomics *14*, 541.

Rojas-Jimenez, K., Wurzbacher, C., Bourne, E.C., Chiuchiolo, A., Priscu, J.C., and Grossart, H.-P. (2017). Early diverging lineages within Cryptomycota and Chytridiomycota dominate the fungal communities in ice-covered lakes of the McMurdo Dry Valleys, Antarctica. Sci Rep *7*.

Roy, J., Reichel, R., Brüggemann, N., Hempel, S., and Rillig, M.C. (2017). Succession of arbuscular mycorrhizal fungi along a 52-year agricultural recultivation chronosequence. FEMS Microbiol Ecol *93*.

Schloss, P.D., Jenior, M.L., Koumpouras, C.C., Westcott, S.L., and Highlander, S.K. (2016). Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. PeerJ *4*, e1869.

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., Chen, W., and Consortium, F.B. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. PNAS *109*, 6241–6246.

Shearer, C.A., Raja, H.A., Miller, A.N., Nelson, P., Tanaka, K., Hirayama, K., Marvanová, L., Hyde, K.D., and Zhang, Y. (2009). The molecular phylogeny of freshwater Dothideomycetes. Studies in Mycology *64*, 145–153.

Singer, E., Bushnell, B., Coleman-Derr, D., Bowman, B., Bowers, R.M., Levy, A., Gies, E.A., Cheng, J.-F., Copeland, A., Klenk, H.-P., et al. (2016). High-resolution phylogenetic microbial community profiling. The ISME Journal *10*, 2020–2032.

Tedersoo, L., Bahram, M., Puusepp, R., Nilsson, R.H., and James, T.Y. (2017a). Novel soil-inhabiting clades fill gaps in the fungal tree of life. Microbiome *5*, 42.

Tedersoo, L., Ave, T.-K., and Anslan Sten (2017b). PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. New Phytologist *217*, 1370–1385.

Tibpromma, S., Hyde, K.D., Bhat, J.D., Mortimer, P.E., Xu, J., Promputtha, I., Doilom, M., Yang, J.-B., Tang, A.M.C., and Karunarathna, S.C. (2018). Identification of endophytic fungi from leaves of Pandanaceae based on their morphotypes and DNA sequence data from southern Thailand. MycoKeys *33*, 25–67.

Travers, K.J., Chin, C.-S., Rank, D.R., Eid, J.S., and Turner, S.W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res *38*, e159.

Vaupotic, T., Veranic, P., Jenoe, P., and Plemenitas, A. (2008). Mitochondrial mediation of environmental osmolytes discrimination during osmoadaptation in the extremely halotolerant black yeast Hortaea werneckii. Fungal Genetics and Biology *45*, 994–1007.

Vizzini, A., Angelini, C., Losi, C., and Ercole, E. (2018). Diversity of polypores in the Dominican Republic: Pseudowrightoporia dominicana sp. nov. (Hericiaceae, Russulales). MycoKeys *34*, 35–45.

Wang, M., Tan, X.-M., Liu, F., and Cai, L. (2018). Eight new Arthrinium species from China. MycoKeys *34*, 1–24.

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

Yang, Y., Fan, F., Zhuo, R., Ma, F., Gong, Y., Wan, X., Jiang, M., and Zhang, X. (2012). Expression of the laccase gene from a white rot fungus in Pichia pastoris can enhance the resistance of this yeast to H2O2-mediated oxidative stress by stimulating the glutathione-based antioxidative system. Appl. Environ. Microbiol. *78*, 5845–5854.

III 5.8S as a low variability complementary marker to ITS2 improves high level classifications of aquatic fungi

Felix Heeger, Christian Wurzbacher, Elizabeth C. Bourne, Camila J. Mazzoni, Michael T. Monaghan

1 Abstract

The kingdom Fungi comprises an enormous amount of evolutionary diversity. Current estimates range from 1.5 – 6 million species within 12 phyla. The large majority of species are not described and those that are often require specialist identification. The internal transcribed spacer (ITS) region of the rRNA operon is widely used as a DNA barcode for fungi in metabarcoding studies. However in the absence of a sufficiently similar reference sequence, query sequences may be classified simply as fungi. Many DNA metabarcoding studies sequence a part of the 5.8S region located between ITS1 and ITS2, when sequencing the ITS2. We performed an *in silico* analysis of 5.8S and ITS sequences from the UNITE database and found that while the 5.8S region was too conserved for species-level identification, it outperformed ITS for producing higher level classifications, even in the absence of closely related reference data. We then developed an automated pipeline for the combined analysis of 5.8S and ITS2, whereby data from both regions, derived from a single DNA metabarcode sequence that is widely used in fungal diversity studies, were used to classify fungi. To evaluate the pipeline, we amplified part of the 5.8S gene together with ITS2 from sediment and water samples from freshwater lakes. 86% of the OTUs from these samples could be classified at least to the class level with the 5.8S while with the ITS2 only 46% could be classified to this level. In many studies the part of the 5.8S is sequenced to provide a conserved primer binding site, but it is discarded before the analysis. We show that it can be used to complement ITS2 data and help with high level taxonomic classification for sequences where ITS2 is failing to give any classification. This is especially helpful in understudied environments like freshwater lakes, where database coverage is poor.

2 Introduction

The kingdom of Fungi contains an enormous diversity of species and life styles. Estimations of the number of species range from 1.5 to 6 million (Hawksworth, 1991; Taylor D. Lee et al., 2014) of which only a small fraction (<144,000, http://www.speciesfungorum.org/Names/Names.asp, accessed May 2018) have been formally described. The evolutionary relationships between fungal species are far from resolved even at higher taxonomic ranks. Even giving the number of phyla inside the kingdom of fungi is therefore difficult. The classification of fungi by Hibbett et al. (Hibbett et al., 2007) in 2007 names seven phyla. In 2011 Blackwell gave the number of phyla as “about 10” (Blackwell, 2011). And after the recent definition of the new phylum Cryptomycota (or Rozellomycota) (Corsaro et al., 2014; Jones et al., 2011; Lara et al., 2010), a study by Tedersoo et al. (2017a) speaks of 12 phyla in the introduction, but also indicates that there may be more phyla to find and shows that “all fungal phyla accommodate previously unrecognized fungal groups”. The UNITE database (Kõljalg et al., 2013) currently (version 7.2, 2017-12-01) lists 18 phyla, including preliminary named phyla GS01 and GS19.

Schoch et al. (Schoch et al., 2012) proposed the internal transcribed spacer (ITS) region of the eukaryotic rRNA operon as a universal fungal DNA barcode. The ITS region is ca. 300 - 1,200 bp and is located between the 18S (SSU) and 28S (LSU) rRNA genes. It contains the two highly variable spacers, ITS1 and ITS2, that are separated by the less variable 5.8S gene (Nilsson et al., 2008). Subsequently, a community-curated reference database (UNITE , Kõljalg et al., 2013) was established for ITS sequences of fungi.

Advances in sequencing technologies have enabled a shift to DNA metabarcoding surveys of environmental samples, whereby sample throughput is much higher and whole communities can be studied without the need for isolation and culture of single species. A trade-off is that sequences from high-throughput methods are shorter than those produced by Sanger sequencing, traditionally used for DNA barcoding. Illumina sequencers are most commonly used and have a read length < 300 bp. Even with an overlapping paired-end design, the maximum length for a continuously read sequence is approximately 550 bp. As a result, it is not feasible to sequence the whole ITS region and most studies focus on either the ITS1 or ITS2 (Miller et al., 2016; Tedersoo et al., 2014; Wurzbacher et al., 2017).

The ability of short DNA metabarcodes to identify fungal taxa in mixed samples varies among studies. An in silico test with 8,967 ITS sequences from a range of fungal phyla (Porras-Alfaro et al., 2014) reported that > 90% of test data (ITS1 91%; ITS2 93%) were identified to the correct genus. In a mock community of 24 Dikarya species, both ITS1 and ITS2 sequences of different species could be clustered into operational taxonomic units (OTUs) and classified correctly (Tedersoo et al., 2015). Classification of ITS sequences obtained from environmental samples has proven more challenging in many studies. Rime et al. (2015) report that 5% of the ITS2 OTUs from soil samples could not be classified to Phylum (i.e. only to kingdom fungi). Wurzbacher et al. (2017) found that with ITS2, 25% of fungal OTUs in permafrost thaw ponds could not be assigned to phylum. In a study of fungi in decaying wood Yang et al. (2016) found that 19 - 25% of OTUs could not be classified to phylum.

A potential reason for the inability of ITS DNA metabarcodes to classify a proportion of fungi from environmental samples, even to higher taxonomic levels, is the high variability of the marker sequence. While high variability among taxa is an important criterion for any marker to be able to distinguish groups, the variability may hinder classification of evolutionarily more distant taxa because the high divergence can make it difficult to establish homology and thereby identify a closest match. This may be especially problematic in less studied habitats such as freshwater, where a high variety of early diverging fungal lineages thrive (Grossart et al., 2016) and for which sequences from closely related species are often not available in reference databases. In this case, classification to any taxonomic level becomes impossible.

Interestingly, many fungal DNA metabarcoding studies amplify the ITS2 region using the primer pair ITS3/ITS4 (White et al., 1990), which includes a region of the 5.8S rRNA gene that is normally discarded before analysis (Bálint et al., 2014; Lindahl et al., 2013). The 5.8S rRNA gene has a much lower substitution rate compared to either ITS (Nilsson et al., 2008) and here we tested whether this more conserved region could provide higher level classification in cases where ITS2 could not. The 5.8S rRNA gene has been used for phylogenetic classification in cases where the ITS1 or ITS2 did not give any classification in low throughput studies of fungi (Neubert et al., 2006; Roose-Amsaleg et al., 2004). The fact that the 5.8S gene is included in the ITS reference database UNITE, gives a taxonomy that can be used in analysis of the 5.8S, and makes comparison with the ITS1 and ITS2 uncomplicated.

We investigated the use of 5.8S as complementary marker for higher taxonomic ranks using in situ environmental samples and by performing an in silico analysis of sequences in the UNITE database. We classified query sequences at different taxonomic ranks using the 5.8S, ITS1 and ITS2 and examined how classification worsened as the reference database was less complete. Specifically, we excluded all other sequences from individuals of the same species, genus, or family. We observed that ITS1 and ITS2 are clearly superior for species-level classifications when the reference database is complete, but that 5.8S outperforms both at higher level taxonomic assignments with a incomplete database. We develop and implement an automated pipeline to analyze amplicons that contain both 5.8S and ITS2 rRNA gene regions, typical of most fungal DNA metabarcoding studies. The two markers are independently analyzed while keeping track of which two marker sequences come from the same molecule to combine the result into a final classification. A test on sediment and water samples from 20 freshwater lakes showed that the 5.8S sequence added phylum level classifications for most (74%) of the 64% of our ITS2 OTUs that were unclassified at that level with ITS2 alone. The current version of the pipeline can be found at www.github.com/f-heeger/two_marker_metabarcoding.

3 Methods

Classification Approaches

For the in silico as well as lake community analyses, we used a lowest common ancestor (LCA) classification based on database search results similar to the one employed in

MEGAN (Huson et al., 2007). First a database search of each sequence is performed against the UNITE database. For each sequence hits with an e-value below a minimum value are considered. Any hit with an identity or query coverage below a certain threshold or a bitscore lower than a certain fraction of the best score for that sequence is excluded. For the remaining hits the lowest common ancestor in the taxonomic tree that underlies UNITE is determined in the following way: For each level in the taxonomic tree, starting from kingdom, classifications of all hits are compared. If the classification at this level of 90% or more of the hits are the same, it will be accepted as the classification on this level for the query sequence. Otherwise the lowest common ancestor is found and the query will only be classified to the last level, where a 90% majority was achieved. During this process any classifications of “undetermined” or “unclassified” are ignored.

ITS2 sequences were additionaly classified with the RDP (Wang et al., 2007) classifier to make sure the LCA approach we implemented here gives results comparable to widely applied tools. We used the classifier trained for use in the PIPITS pipeline (Gweon et al., 2015) on ITS sequences from the current version (7.2, 2017-12-01) of UNITE.

Testing the effects of an incomplete reference database

For the *in silico* evaluation of how an incomplete reference database affects classification with different rRNA markers, we created a dataset whereby the correct assignment of each query sequence was known, and where a sequence from the same species, genus and family was also available. This allowed us to test whether classifications at a given rank were correct, even when all other sequences for the species, genus, or family were removed. An additional criterion was that ITS1, ITS2, and 5.8S had to be available to allow for comparison between the markers. We created such a dataset in the following way: Fungal ITS1, 5.8S and ITS2 sequences were extracted from sequences in the UNITE database (version 7.2, 2017-12-01) using ITSx with default parameters (Bengtsson-Palme et al., 2013). Sequences that satisfied the following three criteria were selected: i) all three markers could be detected by ITSx, ii) a species-level classification was available in UNITE, and iii) at least one other sequence was available for the same species, genus, and family. There were 5,802 sequences that satisfied these criteria and from these we chose a random subset of 100 sequences for our evaluation.

Marker sequences (ITS1, ITS2, 5.8S) were classified independently with the LCA approach using the UNITE database as reference. For 5.8S and ITS2, the classification was run with range of parameter values for minimum identity, minimum coverage, top bit score fraction cutoff, as well as LCA majority stringency. This was done to investigate the parameter stability of the approach. The effect of missing database coverage was tested by first classifying query sequences using the complete reference database, and in subsequent iterations classifying the same query after removal of sequences from the same species, same genus, same family as the query. To asses the necessity of classifying the 5.8S and ITS2 independently the combined fragment of 5.8S and ITS2 was also classified with the LCA approach. The resulting classifications were compared with the classifications given in the UNITE database to determine correct and wrong classifications at each taxonomic level.

5.8S reference data set

As reference dataset for classification of 5.8S sequences we used the 5.8S sequences that were extracted from UNITE with ITSx (above) and complemented them with (non-fungal) 5.8S sequences from the 5.8S rRNA family (RF00002) of the Rfam database (Kalvari et al., 2018). Identical sequences were reduced to one representative with vsearch (Rognes et al., 2016). For each representative a taxonomic assignment was determined by generating a LCA from the classifications of all sequences it represents. For RFAM sequences classified as fungi any classification at lower rank was ignored and priority was given to the taxonomy information from the UNITE database.

Description of the pipeline

The pipeline was implemented as a workflow with snakemake (Köster and Rahmann, 2012) and has four main stages: 1) initial read processing, 2) 5.8S classification, 3) ITS2 classification and 4) final classification (Fig. 1).

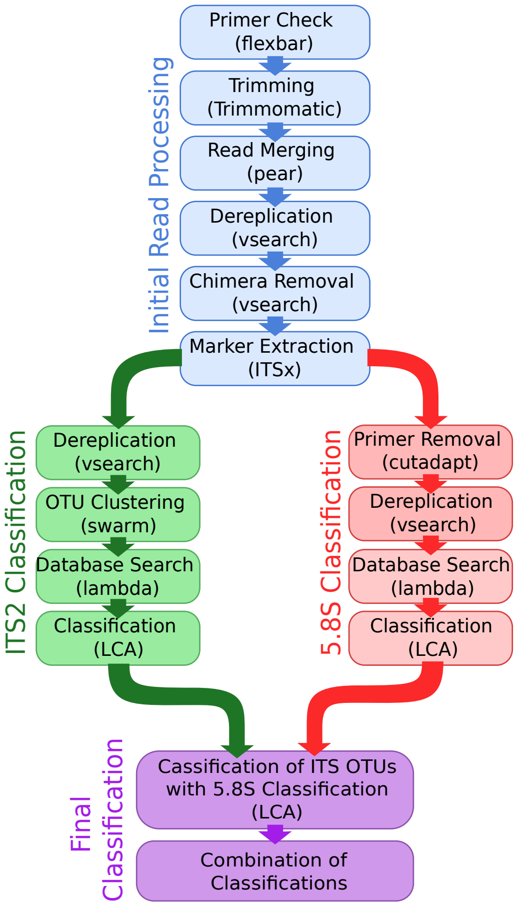


Figure 1: Overview of the steps in the automated snakemake pipeline for parallel classification with ITS2 and 5.8S. External tools and used approaches are given in parenthesis.

(1) Initial read processing starts by producing quality plots with FastQC (version 0.11.2, Andrews). The presence of the forward or reverse primer in the first 25 bp of the respective read is checked with flexbar (version v2.5_beta, Roehr et al., 2017). Quality trimming with Trimmomatic (version 0.35, Bolger et al., 2014) consists of a sliding window trimming with a window size of 8 and a minimum Phred score of 20 and removal of trailing bases with a Phred quality < 20, followed by the removal of sequences with a length < 200 and an average Phred score (after trimming) < 30. Next forward and reverse read of each pair are joined with Pear (version 0.9.6, Zhang et al., 2014). By default the minimum overlap for merging is set to 10. Pairs that can not be merged or are shorter than 150bp or longer than 550bp after merging are discarded. Merged sequences are dereplicated with vsearch. Potential chimeras (including “suspicious” sequences) are removed with vsearch in *de novo* chimera detection mode with default parameters. The 5.8S and ITS2 sequences are extracted with ITSx with default parameters. Partially recognized 5.8S sequences are accepted. The 5.8S and the ITS2 sequences are independently classified in stage 2 and 3 respectively.

(2) 5.8S classification starts with removal of the forward primer and sequences with ambiguous bases are discarded using cutadapt (version 1.9.1, Martin, 2011). Sequences are dereplicated with vsearch. After that sequences are classified by a similarity search against our combined 5.8S reference dataset with lambda (version 0.9.3, Hauswedell et al., 2014)

followed by a LCA classification as described in the *Classification Approaches* section (above).

(3) ITS2 classification starts with dereplication of ITS2 sequences with vearch. Clustering into OTUs is done with swarm2 (version 2.1.6, Mahé et al., 2015). OTUs are classified by similarity search and LCA the same way as 5.8S sequences (above).

(4) The final classification combines the classifications from stage 2 and 3. For each read in an ITS2 OTU cluster all 5.8S sequences with their classifications are collected. The 5.8S classifications are combined with the same LCA approach explained above. The resulting classification is compared to the ITS2 classification. If 5.8S and ITS2 classification are concordant, but the ITS2 is classified to a lower taxonomic rank, the ITS2 classification is accepted. Sequences that are unclassified with ITS2 will automatically take the 5.8S classifications. All conflicting classifications can either be marked (default) or resolved by the user by giving priority to one of the markers.

Test with reads from freshwater lake samples

We tested the pipeline on an unpublished data set (Bourne E.C. et al. unpublished) of water and sediment samples, taken in October and November 2014 from the littoral zone of 20 freshwater lakes in North-West Germany. In six lakes additional sediment and water samples were taken from the pelagic zone. The standard primer pair ITS3/ITS4 (White et al., 1990) was used to amplify a 350-500 bp amplicon consisting of the full ITS2 and ca. 130 bp of the 5'-end of the 5.8S gene. Amplicons were sequenced with overlapping 300 bp paired-end reads on an Illumina MiSeq.

4 Results

Analysis of the classification of query sequences with an increasingly incomplete reference database (Fig. 2) showed a clear difference among markers. When no sequences were removed from the reference database, ITS1 classified 90% of queries to species and ITS2 classified 88%. There were no wrong classifications in either marker (Fig. 2). In contrast, 5.8S classified 5% of queries to species rank and < 60% of sequences were classified to Order. However, the removal of all sequences from the same species, genus, or family had an increasingly detrimental effect on the classification success of both ITS1 and ITS2 sequences (Fig. 2). Wrong classifications become more frequent and classifications at higher ranks were less successful; even the removal of only the species (i.e. other species in the genus still present in the database) cause a distinct drop in successful classification of ITS1 and ITS2 at the Kingdom, Phylum, and Class ranks (Fig. 2). In contrast, the kingdom and phylum rank classifications of 5.8S sequences were not notably affected by the removal of reference sequences, with classification at the class rank only dropping from 80% to 70% (Fig. 2).

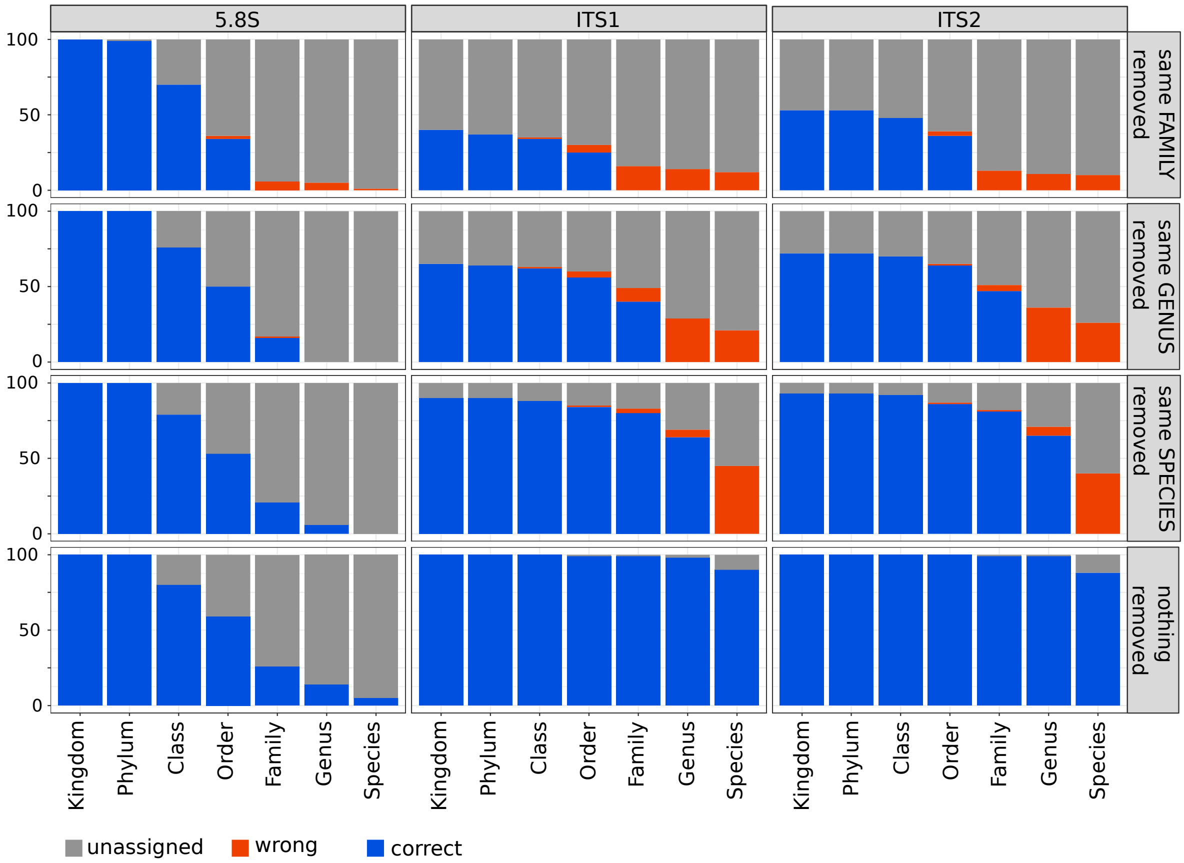


Figure 2: Results of classification of the in silico test set (100 sequences). LCA classification was performed with different markers (panels from left to right) and different completeness of the reference database (panels from top to bottom). Number of correct (blue), wrong (red) and unassigned (grey) classifications are given compared to the original classification in UNITE.

The LCA classification was performed with different parameters for ITS2 and 5.8S to test parameter stability. The stringency parameter had no strong influence on ITS2 classifications (Appendix 1, Suppl. Fig. 1). Lowering the minimum identity and minimum coverage parameters both increase correct as well as wrong classifications (Appendix 1, Suppl. Fig. 2 and 3). Lower values for the top bitscore fraction parameter caused more wrong ITS2 classifications without increasing the number of correct classifications (Appendix 1, Suppl. Fig. 4). Minimum identity and minimum coverage had no strong influence on 5.8S classifications (Appendix 1, Suppl. Fig. 5 and 6), although for the latter a very high value (100%) resulted in more wrong classifications. The top bitscore fraction parameter gave more correct 5.8S classifications for values <= 5%, but at the cost of an increased number of wrong classifications (Appendix 1, Suppl. Fig. 7). Finally a low value (<=85%) for the stringency parameter gave higher number of wrong 5.8S assignments, while a to high value (100%) caused a decrease in correct assignments (Appendix 1, Suppl. Fig. 8).

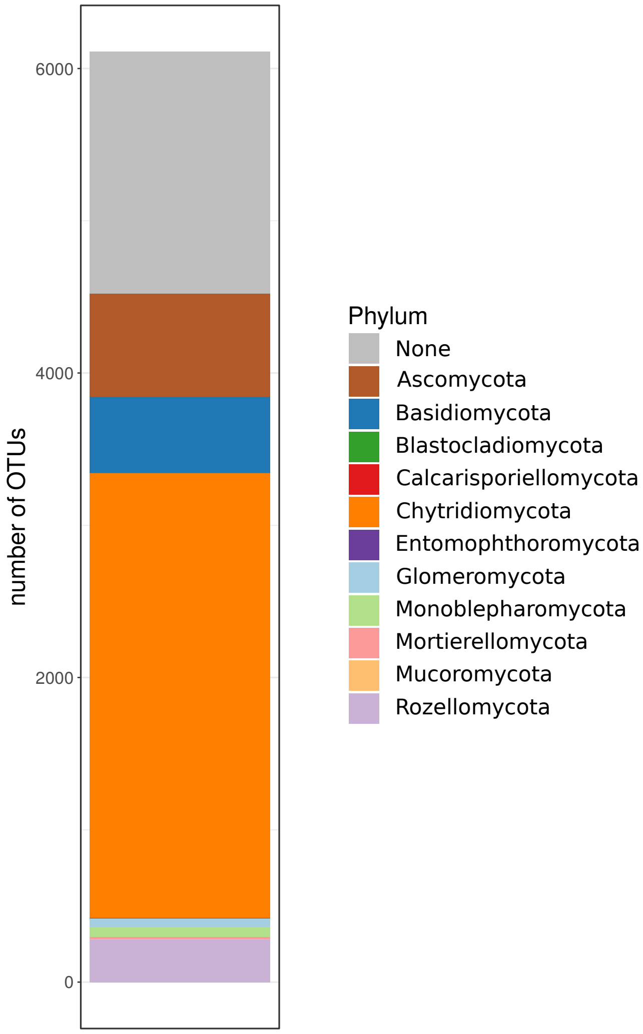


Figure 3: Number of additionally OTUs classified to the phylum level when using the 5.8S as complementary marker in addition to the ITS2.

Comparison with RDP classifications (Appendix 1, Suppl. Fig. 9) showed that the LCA approach gives comparable results to the RDP classifier (trained on the UNITE database) for our data. The comparison between independent classification of ITS2 and 5.8S with the classification of a combined fragment of both regions showed, that in the combined fragment the addition of the 5.8S improved classification at higher ranks (kingdom and phylum), but not same level as for independent 5.8S classification.

The environmental data set from 20 freshwater lakes (water and sediment samples) consisted of 13.6 million read pairs. Our analysis pipeline generated 17,514 non-singleton OTUs, 11,278 of which were classified as fungi. Of the fungal OTUs, 46% (ca. 37% of reads) were classified to the phylum rank or lower using only ITS2 for classification. Using 5.8S for classification in addition provided phylum-level classifications for 86% of OTUs (ca. 81% of reads), or 74% of OTUs with unknown phylum according to ITS2 (Fig. 3). Of these fungi that were only classified with 5.8S, nearly half (48%) were from the phylum Chytridiomycota (Fig. 3). Furthermore the additional data from the RFAM database also allows for a broad overview of the non-fungal classes amplified in the experiment.There was a classification conflict for only one OTU. The 5.8S classification was arthropoda, whereas the ITS2 classification was ascomycota. This was caused by a miss-classification of SH200261.07FU in the UNITE database.

5 Discussion

We implemented a modular pipeline for the processing of fungal DNA metabarcoding data that uses the taxonomic information from the 5.8S gene to complement the more standard ITS2 region. This allowed us to classify a substantially greater number of OTUs than ITS2 alone, in particular for less well studied, basal fungi.

As soon as the species is missing from the database, the ability of ITS to identify the query to any level decreases, with even Kingdom or Phylum being better identified with 5.8S.

The *in silico* analysis of the ITS1, ITS2 and 5.8S sequences in the UNITE database indicated that both ITS1 and ITS2 are very good marker sequences given a database containing the exact same sequence. In our test cases no sequences were assigned to the wrong species and very few were unclassified. This result is somewhat biased because we used only species with clear species identification in UNITE and available sequences from the same species, genus and family. When we removed all sequences assigned to the same species as the query from the database, not only was the algorithm obviously not able to assign the correct species, but also the ability to classify the genus correctly dropped to 65%, although sequences from other species in the same genus were in the database. Even for higher taxonomic ranks (phylum, class) the removal caused assignment problems. Simulating novel genera or families by removing the respective sequences from the database increased the effect even more. This is most likely the reason that many fungal OTUs remain unclassified in environmental studies that focus on poorly studied environments like freshwater (Grossart et al., 2016). New species, genera or families that do not any reference sequences available could even be unidentified at the kingdom rank, leading to fungal diversity being underestimated. The result from our environmental test data set showed that many Chytridiomycota could not be identified to the phylum level by ITS2. They are a group that is not well represented in the UNITE database (Frenken et al., 2017) and thus it is possible that we sequenced some species which are not represented in the database with a sequence from the same genus or family causing the classification to fail completely as shown in the *in silico* analysis. This could lead to a severe bias if we look at the proportion of fungal phyla in our data. Based on the ITS2 alone we would have estimated (based on proportion of reads) the percentage of Chytridoimycota to be 3% while the 5.8S classifications show that the actual proportion is an order of magniture higher (32%). Similarly, the percentage of Rozellomycota (also know as Cryptomycota) would change from 0.1% to 3%.

Although the ITS2 metabarcode allows for the high identification accuracy when perfect reference data is available, it also causes problems to find high enough similarity to any sequence when no closely related species is represented in the database. This is were the 5.8S sequence can help to classify OTUs at least to a higher taxonomic rank. In our environmental data, the 5.8S was especially helpful in splitting the results into fungal and non-fungal sequences when it comes to early diverging lineages or lineages that belong to the Top 50 unknown fungal lineages (Nilsson et al., 2016). However, it should be made clear that the 5.8S would be of limited use as a DNA barcode on its own, or to delineate OTUs, but it should rather be seen as a complementary information, that can be obtained together with ITS2 data. In this respect it is important to note that the employed primer pair has been used for over 20 years now (White et al., 1990) and is generally one of the most frequently employed primer set for fungal surveys in the environment. We hope this will make the suggested approach, highly interesting for the whole fungal scientific community. Our proof of concept implementation of a LCA based classification and combination of ITS2 and 5.8S classification performs comparable to the commonly used RDP classifier on our test dataset and was not very sensitive to parameter choice. Unlike using a single “best” (e.g. lowest e-value) blast hit for identification which can easily lead to wrong assignments if the query species is missing from the database, our approach uses a certain proportion of top blast hits to try and quantify the uncertainty of our assignment by choosing a higher taxonomic rank. Nevertheless we found substantial amount of wrong assignments in the *in silico* analysis, when the database was not complete (Fig. 2).

Although the short read sequencers are currently most efficient in producing a massive amount of data, new technologies are now available that allow to extent the length of the investigated barcode or amplicon. The potential of long read sequencers made researchers already switch to longer fragments such as full length 16S sequences for bacteria (Mosher et al., 2014; Schloss et al., 2016; Singer et al., 2016) a fragment spanning the full ITS region (Schlaeppi et al., 2016; Tedersoo et al., 2017b). Longer amplicons with multiple gene regions can be analyzed in a similar way as shown here. A combination of markers possibly each with their own advantages can be set up with respective reference databases and classification priority rules, gaining an even higher confidence level with each incorporated marker region. In chapter two of this thesis, this idea was applied to the full eukaryotic rRNA operon using the SSU, ITS region, and LSU as markers, which were independently classified.

6 Author contribution

F.H., E.C.B., C.W., C.J.M. and M.T.M. conceived and designed the overall study. **F.H.** designed and implemented the analysis pipeline and carried out analysis. **F.H.**, C.W. and M.T.M. wrote the manuscript. Research was partially funded by the Leibniz Association Pakt/SAW project “MycoLink” (SAW-2014-IGB-1).

7 References

Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data.

Bálint, M., Schmidt, P.-A., Sharma, R., Thines, M., and Schmitt, I. (2014). An Illumina metabarcoding pipeline for fungi. Ecol. Evol. *4*, 2642–2653.

Bengtsson-Palme, J., Ryberg Martin, Hartmann Martin, Branco Sara, Wang Zheng, Godhe Anna, Wit Pierre, Sánchez García Marisol, Ebersberger Ingo, Sousa Filipe, et al. (2013). Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. Methods Ecol. Evol. *4*, 914–919.

Blackwell, M. (2011). The Fungi: 1, 2, 3 … 5.1 million species? Am. J. Bot. *98*, 426–438.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Corsaro, D., Walochnik, J., Venditti, D., Steinmann, J., Müller, K.-D., and Michel, R. (2014). Microsporidia-like parasites of amoebae belong to the early fungal lineage Rozellomycota. Parasitol. Res. *113*, 1909–1918.

Frenken, T., Alacid Elisabet, Berger Stella A., Bourne Elizabeth C., Gerphagnon Mélanie, Grossart Hans Peter, Gsell Alena S., Ibelings Bas W., Kagami Maiko, Küpper Frithjof C., et al. (2017). Integrating chytrid fungal parasites into plankton ecology: research gaps and needs. Environ. Microbiol. *19*, 3802–3822.

Grossart, H.-P., Wurzbacher, C., James, T.Y., and Kagami, M. (2016). Discovery of dark matter fungi in aquatic ecosystems demands a reappraisal of the phylogeny and ecology of zoosporic fungi. Fungal Ecol. *19*, 28–38.

Gweon, H.S., Oliver Anna, Taylor Joanne, Booth Tim, Gibbs Melanie, Read Daniel S., Griffiths Robert I., Schonrogge Karsten, and Bunce Michael (2015). PIPITS: an automated pipeline for analyses of fungal internal transcribed spacer sequences from the Illumina sequencing platform. Methods Ecol. Evol. *6*, 973–980.

Hauswedell, H., Singer, J., and Reinert, K. (2014). Lambda: the local aligner for massive biological data. Bioinforma. Oxf. Engl. *30*, i349-355.

Hawksworth, D.L. (1991). The fungal dimension of biodiversity: magnitude, significance, and conservation. Mycol. Res. *95*, 641–655.

Hibbett, D.S., Binder, M., Bischoff, J.F., Blackwell, M., Cannon, P.F., Eriksson, O.E., Huhndorf, S., James, T., Kirk, P.M., Lücking, R., et al. (2007). A higher-level phylogenetic classification of the Fungi. Mycol. Res. *111*, 509–547.

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007). MEGAN analysis of metagenomic data. Genome Res. *17*, 377–386.

Jones, M.D.M., Richards, T.A., Hawksworth, D.L., and Bass, D. (2011). Validation and justification of the phylum name Cryptomycota phyl. nov. IMA Fungus *2*, 173–175.

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A., Finn, R.D., and Petrov, A.I. (2018). Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. *46*, D335–D342.

Kõljalg, U., Nilsson R. Henrik, Abarenkov Kessy, Tedersoo Leho, Taylor Andy F. S., Bahram Mohammad, Bates Scott T., Bruns Thomas D., Bengtsson Palme Johan, Callaghan Tony M., et al. (2013). Towards a unified paradigm for sequence based identification of fungi. Mol. Ecol. *22*, 5271–5277.

Köster, J., and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. Bioinformatics *28*, 2520–2522.

Lara, E., Moreira, D., and López-García, P. (2010). The Environmental Clade LKM11 and Rozella Form the Deepest Branching Clade of Fungi. Protist *161*, 116–121.

Lindahl, B.D., Nilsson, R.H., Tedersoo, L., Abarenkov, K., Carlsen, T., Kjøller, R., Kõljalg, U., Pennanen, T., Rosendahl, S., Stenlid, J., et al. (2013). Fungal community analysis by high-throughput sequencing of amplified markers--a user's guide. New Phytol. *199*, 288–299.

Mahé, F., Rognes, T., Quince, C., Vargas, C. de, and Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ *3*, e1420.

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

Miller, K.E., Hopkins, K., Inward, D.J.G., and Vogler, A.P. (2016). Metabarcoding of fungal communities associated with bark beetles. Ecol. Evol. *6*, 1590–1600.

Mosher, J.J., Bowman, B., Bernberg, E.L., Shevchenko, O., Kan, J., Korlach, J., and Kaplan, L.A. (2014). Improved performance of the PacBio SMRT technology for 16S rDNA sequencing. J. Microbiol. Methods *104*, 59–60.

Neubert, K., Mendgen, K., Brinkmann, H., and Wirsel, S.G.R. (2006). Only a Few Fungal Species Dominate Highly Diverse Mycofloras Associated with the Common Reed. Appl. Environ. Microbiol. *72*, 1118–1128.

Nilsson, R.H., Kristiansson, E., Ryberg, M., Hallenberg, N., and Larsson, K.-H. (2008). Intraspecific ITS Variability in the Kingdom Fungi as Expressed in the International Sequence Databases and Its Implications for Molecular Species Identification. Evol. Bioinforma. Online *4*, 193–201.

Nilsson, R.H., Wurzbacher, C., Bahram, M., Coimbra, V.R.M., Larsson, E., Tedersoo, L., Eriksson, J., Duarte, C., Svantesson, S., Sánchez-García, M., et al. (2016). Top 50 most wanted fungi. MycoKeys *12*, 29–40.

Porras-Alfaro, A., Liu, K.-L., Kuske, C.R., and 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*, 829–840.

Rime, T., Hartmann Martin, Brunner Ivano, Widmer Franco, Zeyer Josef, and Frey Beat (2015). Vertical distribution of the soil microbiota along a successional gradient in a glacier forefield. Mol. Ecol. *24*, 1091–1108.

Roehr, J.T., Dieterich, C., and Reinert, K. (2017). Flexbar 3.0 – SIMD and multicore parallelization. Bioinformatics *33*, 2941–2942.

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

Roose-Amsaleg, C., Brygoo Yves, and Harry Myriam (2004). Ascomycete diversity in soil feeding termite nests and soils from a tropical rainforest. Environ. Microbiol. *6*, 462–469.

Schlaeppi, K., Bender, S.F., Mascher, F., Russo, G., Patrignani, A., Camenzind, T., Hempel, S., Rillig, M.C., and van der Heijden, M.G.A. (2016). High-resolution community profiling of arbuscular mycorrhizal fungi. New Phytol. *212*, 780–791.

Schloss, P.D., Jenior, M.L., Koumpouras, C.C., Westcott, S.L., and Highlander, S.K. (2016). Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. PeerJ *4*, e1869.

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., Chen, W., and Consortium, F.B. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc. Natl. Acad. Sci. *109*, 6241–6246.

Singer, E., Bushnell, B., Coleman-Derr, D., Bowman, B., Bowers, R.M., Levy, A., Gies, E.A., Cheng, J.-F., Copeland, A., Klenk, H.-P., et al. (2016). High-resolution phylogenetic microbial community profiling. ISME J. *10*, 2020–2032.

Taylor D. Lee, Hollingsworth Teresa N., McFarland Jack W., Lennon Niall J., Nusbaum Chad, and Ruess Roger W. (2014). A first comprehensive census of fungi in soil reveals both hyperdiversity and fine scale niche partitioning. Ecol. Monogr. *84*, 3–20.

Tedersoo, L., Bahram, M., Põlme, S., Kõljalg, U., Yorou, N.S., Wijesundera, R., Ruiz, L.V., Vasco-Palacios, A.M., Thu, P.Q., Suija, A., et al. (2014). Global diversity and geography of soil fungi. Science *346*, 1256688.

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

Tedersoo, L., Bahram, M., Puusepp, R., Nilsson, R.H., and James, T.Y. (2017a). Novel soil-inhabiting clades fill gaps in the fungal tree of life. Microbiome *5*, 42.

Tedersoo, L., Ave, T.-K., and Anslan Sten (2017b). PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. New Phytol. *217*, 1370–1385.

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

White, T.J., Bruns, T., Lee, S., and Taylor, J. (1990). Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In PCR Protocols, (San Diego: Academic Press), pp. 315–322.

Wurzbacher, C., Nilsson, R.H., Rautio, M., and Peura, S. (2017). Poorly known microbial taxa dominate the microbiome of permafrost thaw ponds. ISME J. *11*, 1938–1941.

Yang, C., Schaefer, D.A., Liu, W., Popescu, V.D., Yang, C., Wang, X., Wu, C., and Yu, D.W. (2016). Higher fungal diversity is correlated with lower CO_2 emissions from dead wood in a natural forest. Sci. Rep. *6*, 31066.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics *30*, 614–620.

IV Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments

Felix Heeger, Elizabeth C. Bourne, Christiane Baschien, Andrey Yurkov, Boyke Bunk, Cathrin Spröer, Jörg Overmann, Camila J. Mazzoni, Michael T. Monaghan

1 Abstract

DNA metabarcoding is widely used to study prokaryotic and eukaryotic microbial diversity. Technological constraints limit most studies to marker lengths below 600 bp. Longer sequencing reads of several thousand bp are now possible with third-generation sequencing. Increased marker lengths provide greater taxonomic resolution and allow for phylogenetic methods of classification, but longer reads may be subject to higher rates of sequencing error and chimera formation. In addition, most bioinformatics tools for DNA metabarcoding were designed for short reads and are therefore unsuitable. Here we used Pacific Biosciences circular consensus sequencing (CCS) to DNA-metabarcode environmental samples using a ca. 4,500 bp marker that included most of the eukaryote SSU and LSU rRNA genes and the complete ITS spacer region. We developed an analysis pipeline that reduced error rates to levels comparable to short-read platforms. Validation using a mock community indicated that our pipeline detected 98% of chimeras de novo. We recovered 947 OTUs from water and sediment samples in a natural lake, 848 of which could be classified to phylum, 397 to genus, and 330 to species. By allowing for the simultaneous use of three global databases (Unite, SILVA, RDP LSU), long-read DNA metabarcoding provided better taxonomic resolution than any single marker. We foresee the use of long reads enabling the cross-validation of reference sequences and the synthesis of ribosomal rRNA gene databases. The universal nature of the rRNA operon and our recovery of >100 non-fungal OTUs indicate that long-read DNA metabarcoding holds promise for studies of eukaryotic diversity more broadly.

2 Introduction

DNA-metabarcoding is widely used in the study of microbial communities from all three major domains of life (Wurzbacher et al., 2016), whereby one or more marker regions in the genome are

PCR-amplified and sequenced using a next-generation sequencing (NGS) platform. Reads are quality-filtered and sequences are clustered according to sequence similarity into putative taxa (Operational Taxonomic Units = OTUs). OTUs are then classified using marker-specific, and sometimes taxon-specific databases. DNA metabarcoding has become a commonly used tool because it provides an estimate of biodiversity, including that of taxa that cannot be cultured, and identification relies on relatively stable genetic information rather than often variable and subtle phenotypic characters. Limitations of the method include the fact that marker regions and PCR primers must be selected a priori to detect the taxa of interest, and that the variability of the marker region, and how well the taxa are represented within a given reference database, determine how well the members of an assemblage can be identified (Nilsson et al., 2018).

There is a fundamental trade-off between using a marker that is conserved enough to be amplified across a broad range of taxa, but variable enough to distinguish among closely related species. Marker length also has consequences for how many OTUs can be identified, and to what taxonomic resolution (Porras-Alfaro et al., 2014). Shorter markers within a given locus may include less genetic variation than longer markers, reducing the ability to distinguish closely related species (Singer et al., 2016). One consequence is that highly variable regions are often used as DNA metabarcoding markers. While variable regions may increase taxonomic resolution in groups for which reference sequences are available, sequence homology can be difficult or impossible to establish. This precludes phylogeny-based analyses and can result in the complete failure of classifying OTUs at any taxonomic level (Lindahl et al., 2013).

More recent (i.e. third-generation sequencing) technologies can provide much longer (several kbp) sequencing reads (Goodwin et al., 2016); however, their use in studies of environmental samples remains limited. The few existing studies, using full-length (~1.5 kbp) bacterial 16S (Franzén et al., 2015; Schloss et al., 2016; Singer et al., 2016) and parts of the eukaryotic rRNA operon including ITS (up to 2.6 kbp) (Schlaeppi et al., 2016; Tedersoo et al., 2017), have reported increased taxonomic resolution. The Pacific Biosciences (PacBio) RSII platform generates reads of >50 kbp by Single Molecule Real Time (SMRT) Sequencing. Single pass error rates of 13-15% (Goodwin et al., 2016) limit their value in DNA metabarcoding because species identification is unreliable at those levels of uncertainty. However, the circular consensus sequencing (CCS) version of SMRT sequencing greatly reduces the error rate. In CCS, double stranded DNA amplicon molecules are circularized by the ligation of hairpin adapters. The sequencing polymerase is then able to pass around the molecule and read the same insert multiple times (Travers et al., 2010). The repeated reads of the same amplicon molecule, together with the random nature of sequencing error, can then be used to reduce the final error rate to <1% (Goodwin 2016) by generating consensus sequences.

Beside the higher per base cost a primary reason why long-read approaches have not been applied to DNA metabarcoding is the fact that most of the existing bioinformatic tools have been optimized for the analysis of data from short-read technologies (e.g., Illumina). It is thus unclear how well they will perform on PacBio CCS reads. Longer sequences have more errors because even high-quality reads with low error rates will accumulate more total errors as a function of length. The types of errors in PacBio reads also differ from that of short-read technologies, with CCS reads tending to have more insertions and deletions, compared to substitutions more common in short-read data. Schloss et al. (2016) explored the error profile and steps that can be taken when targeting the 16S for a bacterial mock community, and environmental samples. They found that the error rate of CCS reads of their longest amplicon (V1-V9) was only 0.68% and could be further reduced to 0.027% by pre-clustering at 99% similarity. Chimera formation rate may also be increased in longer markers since longer amplicons may suffer premature elongation terminations, leading to more possibilities for the resulting incomplete amplicons to act as primers in the next PCR cycle and thus more chimeras to be formed (see also Laver et al., 2016). Existing algorithms commonly used to detect chimeras are not optimized for long reads and may therefore fail to detect chimeras.

Fungi are ecologically important eukaryotes, having diverse roles in carbon and nutrient cycling, occupying a range of niches, including as decomposers, parasites and endophytes, and are ubiquitous in terrestrial and aquatic habitats alike (e.g. Tedersoo et al., 2014; Wurzbacher et al., 2016). Microbial fungal communities are increasingly studied with DNA metabarcoding (e.g. Roy et al., 2017), taking advantage of the increased detection of taxa without the need to culture and the reduced cost of sequencing that has permitted ever deeper read depth The broad phylogenetic diversity of fungi has the consequence that fungal DNA metabarcoding studies typically use markers that vary depending on the taxonomic group of interest and the resolution desired. Different regions of the eukaryotic rRNA operon have been widely utilized for barcoding fungi due to its universality, and the fact that short stretches have been able to provide reasonable power for fungal identification. Within this region, the most commonly applied barcode is the internal transcribed spacer (ITS) (Schoch et al., 2012). This comprises the ITS1, the 5.8S rRNA gene and the ITS2, and depending on the lineage, varies from 300 to 1,200 bp in length. In fungal DNA metabarcoding, the ITS2 region is widely used to assess fungal diversity in environmental samples (Blaalid et al., 2013; Kõljalg et al., 2013); however, it is not as successful in identifying taxa as the full length ITS

(Tedersoo et al., 2017). For early diverging fungal lineages, such as those found in many aquatic habitats (Monchy et al., 2011; Rojas-Jimenez et al., 2017; Wurzbacher et al., 2016), sequences from the small subunit (SSU) rRNA gene (18S) can provide affiliation of higher taxonomic ranks, but are often not variable enough to distinguish among species (Cole et al., 2014). The LSU region has higher variability, and therefore resolution, than the SSU, and is often used for identification of specific fungal groups (e.g. Glomeromycota and Chytridiomycota) lacking ITS reference sequences. Databases have been established for all three different markers, e.g. UNITE for ITS (Kõljalg et al., 2013), SILVA for SSU (Quast et al., 2013), and RDP for LSU (Cole et al., 2014). Nevertheless, database coverage remains poor for several fungal lineages, for example Glomeromycota (Ohsowski et al., 2014), Chytridridiomycota (Frenken et al., 2017), and Cryptomycota, and for species from less-studied habitats such as aquatic, indoor, and marine environments.

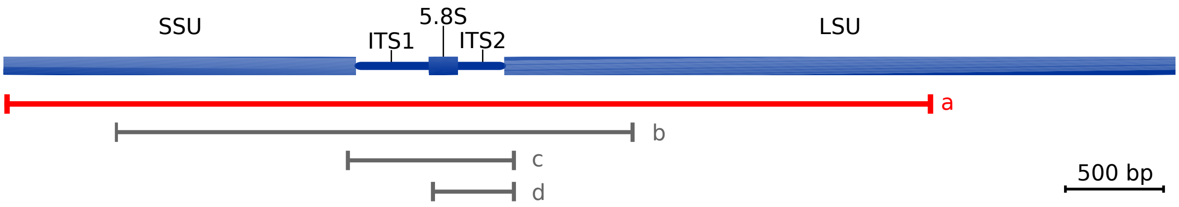


Figure 4: Region of the eukaryotic rRNA operon covered by the primer pair used in this studied (a) compared to the primer pair SSU515Fngs-TW13 used by Tedersoo et al. 2017 (b), the widely used (e.g. Schoch 2012) primer pairs ITS5-ITS4 (c) and ITS3-ITS4 (d)

We examined fungal diversity of field-collected samples from a temperate lake using SMRT CCS of a long (ca. 4,500 bp) DNA metabarcode that included the three major regions of the eukaryotic rRNA operon (SSU, ITS, LSU) in a single sequencing read (Fig. 4). We first sequenced cultured isolates comprising a broad phylogenetic range and a mock community to derive rates of sequencing error and chimera formation. We then developed a new bioinformatics pipeline designed for full length rRNA operon amplicons. We found error rates to be comparable to short-read approaches after filtering with our pipeline, and chimera-formation rates to be comparable to those found in studies with shorter amplicons. We identified 947 OTUs from environmental samples, 848 of which could be classified to phylum, 486 to family, 397 to genus and 330 to species. By allowing for the simultaneous use of three databases, long-read DNA metabarcoding provided much better taxonomic resolution than possible with a single-marker, single-database approach. The universal nature of the rRNA operon and our recovery of >100 non-fungal OTUs indicate that long-read DNA metabarcoding holds promise for future studies of eukaryotic diversity in general.

3 Materials and Methods

Isolates, Mock community, and Environmental samples

Isolates of sixteen fungal species (Table 1) were combined to form a mock community. This community was used to test PCR and library preparation protocols that were later applied to environmental samples, and to quantify the efficiency of *de novo* and reference-based chimera detection in our long-read bioinformatics pipeline described below. Environmental samples were collected from Lake Stechlin, an oligo-mesotrophic lake in North-East Germany (53.143° N 13.027° E) in October 2014. Littoral water samples (30 L total) were collected and pooled from surface water in the shallow zone along three 10 m transects, located within 5 m of the lake shore or reed belt. Pelagic water samples (30 L total) were collected from the deeper zone of the lake by pooling samples taken at multiple depths (0-65 m) at one point, using a Niskin-bottle (Hydro-Bios, Kiel, Germany). A subsample (2 L) of each (littoral and pelagic) was filtered through 0.22-µm Sterivex filters (Merck Millipore, Darmstadt, Germany) using a peristaltic pump (GT-EL2 Easy Load II, UGT, Müncheberg, Germany). Excess water was expelled using a sterile syringe and parafilm used to seal the ends. Sediment samples were collected from four locations in each zone (littoral, pelagic) using a PVC sediment corer (63 mm diameter) on a telescopic bar (Uwitec, Mondsee, Austria). The uppermost 2 cm from each sediment core were pooled in the field and divided into 2 ml subsamples for storage. Sterivex filters and sediment subsamples were frozen in liquid Nitrogen in the field and returned to the laboratory for long-term storage at -80°C.

Table 1: Isolates used and their contribution to the mock community.

Taxon	Code	Isolate	DNA pooled (ng)	% of mock community
Clavariopsis aquatica	CA	DSM 29862†	60	7.6
Chytridiomycota	CHY1	CHY1‡	60	7.6
Cladosporium sp.	Csp1	KR4‡	20	2.5
Clonostachys rosea	CR	DSM 29765§	60	7.6
Cystobasidium laryngis	CL	CBML 151a§	5	0.6
Cladosporium herbarum	CH	KR13‡	20	2.5
Exobasidium vaccinii	EV	DSM 5498§	60	7.6
Leucosporidium scottii	LS	CBML 203§	60	7.6
Metschnikowia reukaufii	MR	DSM 29087§	60	7.6
Mortierella elongata	ME	CBML 271§	60	7.6
Penicillium brevicompactum	PB	KR5‡	80	10.2
Phanerochaete chrysosporium	PC	DSM 1547§	60	7.6
Phoma sp.	Psp1	KR1‡	3	0.4
Saccharomyces cerevisiae	SC	DSM 70449§	60	7.6
Trichoderma reesei	TR	DSM 768†	60	7.6
Ustilago maydis	UM	DSM 14603†	60	7.6

† extracted using Qiagen Dneasy Plant Mini Kit
‡ extracted using peqGOLD Tissue DNA Mini Kit
§ extracted using MasterPure Yeast DNA Purification kit

DNA extraction

Genomic DNA was extracted from fungal isolates using three different methods (Table 1). Environmental DNA was extracted from water and sediment samples using a modified phenol-chloroform method (after Nercessian et al., 2005). Frozen Sterivex cartridges were broken open and sterilized forceps were used to transfer half of the fragmented filter into each of two 2-ml tubes. Sediment samples were thawed and aliquoted into two 2-ml tubes, each containing 200 mg. Beads (0.1 and 1.0 mm zirconium, and 3x 2.5mm glass beads, Biospec, Bartlesville, USA) were added to 0.3 volume of the tube. For cell lysis and extraction, the following reagents were added: 0.6 ml CTAB extraction buffer (5% CTAB-120 mM phosphate buffer), 60 µl 10% sodium dodecyl sulfate, 60 µl 10% N-lauroyl sarcosine, followed by 0.6 ml of phenol:chloroform-isoamyl alcohol (25:24:1). Samples were vortexed immediately to homogenise and then ground for 1.5 min at 30 Hz (Retsch mill, Retsch GmbH, Haan, Germany) with short breaks for cooling on ice. Samples were incubated for 1 hr at 65 °C, with occasional mixing, and then centrifuged at 17,000 g for 10 minutes. The upper aqueous phase was transferred to a new tube and mixed with an equal volume of chloroform-isoamyl alcohol (24:1), centrifuged at 17,000 g for 10 min and the upper aqueous phase transferred to a new tube. Nucleic acids were precipitated with 2 volumes of PEG/NaCl (30% PEG 6000 in 1.6 M NaCl) for 2 h. Samples were centrifuged at 16,000 g for 45 min, and the supernatant discarded. The nucleic acid pellet was washed twice by the addition of 1 ml ice-cold 70% ethanol, centrifuged at 17,000 g for 15 min, and the supernatant discarded and following removal of ethanol traces, eluted in 50 µl nuclease-free water. Subsamples were pooled to give 100 µl nucleic extract per sample. RNA was removed by the addition of 0.5 µl (5 µg) RNase A (10 mg/ml DNase and protease free, ThermoFisher Scientific, Waltham, US) to 80 µl of the pooled sample, incubated at 37 °C for 30 min, and cleaned using the PowerClean Pro DNA Clean-Up kit (MoBio Laboratories, Carlsbad, USA). DNA was quantified in triplicate using a Qubit HS dsDNA Assay (Invitrogen, Carlsbad, USA) and gel-checked for quality.

PCR and chimera formation tests

Approximately 4,500 bp of the eukaryotic rRNA operon (Fig. 4), including SSU, ITS1, 5.8S, ITS2, and LSU (partial) regions, was PCR-amplified using the primers NS1_short (5’-CAGTAGTCATATGCTTGTC-3’) and RCA95m (5’-ACCTATGTTTTAATTAGACAGTCAG-3’) (Wurzbacher et al., 2018). Symmetric (reverse complement) 16-mer barcodes were added to the 5’ ends of primers following the PacBio manufacturer's guidelines on multiplexing SMRT sequencing. We aimed to minimize chimera formation by minimizing the number of PCR cycles performed per sample. Cycle numbers were chosen after amplifying all samples with a variable number of cycles (13-30) and identifying the exponential phase of PCR (Lindahl et al., 2013) according to band visibility on an agarose gel. Based on these results, we used 15-20 cycles to amplify isolates (3-8 ng template DNA), 13-30 cycles for mock community samples (2-20 ng), and 22-26 cycles for environmental samples (10 ng). Barcodes were allocated to the different PCR conditions tested as shown in supplemental table 1 (Appendix 2). All standard PCRs were conducted in 25 µl reactions using 0.5 µl Herculase II Fusion enzyme (Agilent Technologies, Cedar Creek, USA), 5 µl of 5x PCR buffer, 0.62 µl each primer (10 uM), 0.25 µl dNTPs (250 mM each), 0.3 µl BSA (20mg/ml BSA, ThermoFisher Scientific, Waltham, US) on a SensoQuest labcycler (SensoQuest Gmbh, Göttingen, Germany) with 2 min denaturation at 95 °C, 13-30 cycles (see above) of 94 °C for 30 sec, 55 °C for 30 sec and 70 °C for 4 min, and a final elongation at 70 °C for 10 min. Multiple PCR reactions (up to 50) were required for each environmental sample to ensure sufficient product for library preparation (1 µg purified PCR product). We also included a two-step emulsion PCR (emPCR) of the mock community in order to test whether emPCR could reduce chimera formation rate by the physical isolation of DNA template molecules (Boers et al., 2015). The Micellula DNA Emulsion kit (Roboklon GmbH, Berlin) was used for a two-step PCR: a first amplification of 25 cycles, with 2µl of the cleaned template used in a second 25 cycle PCR.

Library preparation and Sequencing

Replicate PCRs were pooled back to sample level, and products were cleaned with 0.45 x CleanPCR SPRI beads (CleanNa, Waddinxveen, Netherlands), pre-cleaned according to PacBio specifications (C. Koenig, pers. comm.), quantified twice using a Qubit HS dsDNA Assay, and quality-checked on an Agilent® 2100 Bioanalyzer System (Agilent Technologies, Santa Clara, USA). Samples were then pooled into libraries before being quality-checked on an Agilent® 2100 Bioanalyzer following PacBio guidelines (Pacific Biosciences, Inc., Menlo Park, CA, USA) for amplicon template library preparation and sequencing.

SMRTbell™ template libraries were prepared according to the manufacturer’s instructions following the Procedure & Checklist – Amplicon Template Preparation and Sequencing (Pacific Biosciences). Briefly, amplicons were end-repaired and ligated overnight to hairpin adapters applying components from the DNA/Polymerase Binding Kit P6 (Pacific Biosciences). We included enough DNA from each sample to obtain the required library concentration (37 ng µl^{-1}) for end-repair. Reactions were carried out according to the manufacturer´s instructions. Conditions for annealing of sequencing primers and binding of polymerase to purified SMRTbell™ template were assessed with the Calculator in RS Remote (Pacific Biosciences). SMRT sequencing was carried out on the PacBio *RSII* (Pacific Biosciences) taking one 240-minutes movie.

In total, we ran 8 libraries and 27 SMRT cells. Three of the isolates (*Trichoderma reesei, Clonostachys rosea,* and a species belonging to the phylum Chytridiomycota) were sequenced on one SMRT cell to test the protocol for CCS. The remaining 13 isolates and one of the mock community conditions (30 PCR cycles) were prepared as part of the libraries containing the environmental samples (Appendix 1, Table 1), which were each run on three SMRT cells. Mock community samples and the emPCR sample were pooled in equimolar ratio and sequenced using two SMRT cells.

Demultiplexing and extraction of subreads from SMRT cell data was performed applying the RS_ReadsOfInsert.1 protocol included in SMRTPortal 2.3.0 with minimum 2 full passes and minimum predicted accuracy of 90%. Barcodes were provided as FASTA files and barcode extraction was performed in a symmetric manner with a minimum barcode score of 23 within the same protocol. Mean amplicon lengths of 3800 – 4500 kbp were confirmed. Demultiplexed reads were downloaded from the SMRT Portal as fastq files for further analysis.

Long-read metabarcoding pipeline

We developed an analysis pipeline for PacBio CCS reads using the python workflow engine snakemake (version 3.5.5, Köster and Rahmann, 2012). Our pipeline combines steps directly implemented in python with steps that use external tools. The implementation is available on github (https://github.com/f-heeger/long_read_metabarcoding) and parameters used for the external tools can be found in the supplemental methods (Appendix 2, supplemental info 1).

Read Processing stage – Reads longer than 6,500 bp were excluded to remove chimeric reads formed during adapter ligation and reads containing double-inserts due to failed adapter recognition during the CCS generation. Reads shorter than 3,000 bp were removed to exclude incompletely amplified sequences and other artifacts. Reads were then filtered by a maximum mean predicted error rate of 0.004 that was computed from the Phred scores. Reads with local areas of low quality were removed if predicted mean error rate was > 0.1 in any sliding window of 8 bp. cutadapt (version 1.9.1, Martin, 2011) was used to remove forward and reverse amplification primers and discard sequences in which primers could not be detected. Random errors were reduced by pre-clustering reads from each sample at 99% similarity using the cluster_smallmem command in vsearch (version 2.4.3, Rognes et al., 2016). Reads were sorted by decreasing mean quality prior to clustering to ensure that high quality reads were used as cluster seeds. vsearch was configured to produce a consensus sequence for each cluster.

OTU clustering and classification stage – Chimeras were detected and removed with the uchime_denovo command in vsearch. Based on tests using mock community samples (see below), we determined this was a suitable method of chimera detection following the read processing stage (above). Only sequences that were classified as non-chimeric were used for further analysis. The rRNA genes (SSU, LSU, 5.8S) and internal transcribed spacers (ITS1, ITS2) in each read were detected using ITSx (version 1.0.11, Bengtsson-Palme et al., 2013). To generate OTUs, the ITS region (ITS1, 5.8S, ITS2) was clustered using vsearch at 97% similarity. SSU and LSU sequences were then placed into clusters according to how their corresponding ITS was clustered. OTUs were taxonomically classified using the most complete available database for each marker. For the ITS we used the general FASTA release of the UNITE database (version 7.1, 20.11.2016, only including singletons set as RefS, Kõljalg et al., 2013); for the SSU we used the truncated SSU release of the SILVA database (version 128, Quast et al., 2013), excluding database sequences with quality scores below 85 or Pintail chimera quality below 50; and for the LSU we used the RDP LSU data set (version 11.5, Cole et al., 2014). The ITS, SSU and LSU regions of the representative sequence of each OTU were locally aligned to the database using lambda (version 1.9.2, Hauswedell et al., 2014). For LSU and SSU the alignment parameters had to be modified to allow for longer alignments (see Appendix 2, supplemental info 1). From the alignment results, a classification was determined by filtering the best matches and generating a lowest common ancestor (LCA) from their classifications as follows. For each query sequence, matches were filtered by a maximum e-value (10^{-6}), a minimum identity (80%) and a minimum coverage of the shorter of the query or database sequence (85%). For the SSU and LSU, non-overlapping matches between each query and database sequence were combined. For each query sequence, a cutoff for the bit score was established representing 95% of the value for the best match, above which all matches for that given sequence were considered. For the SSU and LSU, bit scores were normalized by the minimum length of query and database sequences to account for the varying lengths of database sequences. To determine the LCA from the remaining matches, their classifications were compared at all levels of the taxonomic hierarchy starting at kingdom (highest) and ending at species (lowest) level. For each OTU, the classifications of all matches at a given taxonomic rank were compared and if >90% of them were the same then this was accepted. If <90% were the same then the OTU remained unclassified at this and all lower ranks.

Error rates based on isolate sequences

Isolate sequences were processed using the Read Processing stage of the pipeline (described above) in order to generate error-corrected consensus sequences from pre-clusters. The consensus sequences of the largest pre-cluster for each isolate were > 99 % identical to the Sanger sequencing data obtained from the same isolate (not shown), with most differences found in bases that were of low quality in the Sanger sequence data. We therefore used the consensus sequence of the largest cluster for each isolate as a reference for that species in all further analysis. CCS reads from each isolate were then aligned with the respective consensus sequence using blasr (github comit 16b158d, Chaisson and Tesler, 2012) to estimate error rates of CCS reads. Sequences after filtering steps were also compared in order to estimate remaining errors.

Evaluating chimera detection

De novo and reference-based chimera classifications were compared as a way of estimating the reliability of *de novo* chimera calls. The CCS reads from the mock community samples were tested for chimeras with vsearch once in *de novo* mode (uchime_denovo) and once with a reference-based approach (uchime_ref). For the *de novo* approach, reads were processed with the Read Processing stage of the pipeline (above) to generate error-corrected sequences from pre-clusters. Cluster sizes resulting from the pre-clustering step were used as sequence abundances. For the reference-based approach, a reference file was created from the consensus sequence of the largest cluster for each isolate sample. A random subset of reads (100 sequences, 1.3% of the data) was generated from the mock community sample with the highest chimera rate and the most reads (30 PCR cycles). The subset of reads was aligned to the consensus sequences from the isolate samples and visually inspected for chimeras in Geneious (version 7.1.9, Kearse et al., 2012). These “manual” chimera calls were then used to verify reference-based chimera classifications for these reads. Chimeras identified by the reference-based approach were used to compute the chimera formation rate under different PCR conditions.

Mock community classification

We tested classification with the DNA metabarcoding pipeline using the mock community sample with the most reads (30 PCR cycles). In the pipeline, chimeras were classified *de novo* and OTU classification was performed using the public databases. We manually classified the same OTUs using consensus sequences from our isolate samples as reference. For each read, chimeras were detected with a reference-based approach using vsearch and the classification of the read was determined by mapping reads to the isolate sample sequences with blasr. To better understand the resolution that can be expected from the different regions of the rRNA operon, each region (SSU, ITS1, 5.8S, ITS2, LSU) was clustered independently. Chimeras were first removed using the reference-based approach with our isolate sequences as references. The different regions in each read were separated with ITSx, dereplicated and clustered at 97%.

Environmental community classification

Sequences from the environmental samples from Lake Stechlin were processed with the full rRNA metabarcoding pipeline described above. Chimeras were detected using the *de novo* approach, which we conclude provides a very good diagnosis of chimeras based on our validation using the mock community to compare *de novo* and reference-based approaches (see Results). The resulting classifications obtained with SSU, ITS, and LSU markers were then compared at each taxonomic level. OTUs with only one read (singletons) were excluded from this comparison.

4 Results

Sequencing resulted in a total number of 233,176 CCS reads, which were submitted to the NCBI Sequence Read Archive (SRR6825218 - SRR6825222). 215,720 of these reads were within the targeted size range of 3,000 – 6,500 bp (Table 2). After stringent filtering using average- and window- quality criteria, 69,342 reads remained that contained an identifiable amplification primer sequence (Table 2). Pre-clustering of isolate samples with the metabarcoding pipeline resulted in one large (> 80 reads) pre-cluster for each sample. Besides these big clusters, six samples had additional very small (< 3 reads) clusters. For isolates sequenced on two different SMRT-cells, consensus sequences of the large pre-clusters were identical across cells except for *S. cerevisiae* where a T homopolymer in the ITS2 was 6 bases long in one consensus and 7 in the other and *U. maydis* which shows a SNP in the LSU. Consensus sequences of large clusters were used as reference for further analysis and submitted to gene bank (MH047187 - MH047202). The mean sequencing error rate of quality-filtered CCS reads, based on comparison to the consensus sequences of the large clusters (taken to be our reference for each isolate), was 0.2216% (SD 0.1621%). Deletions were by far the most common error (0.1756%), with insertions and substitutions much lower (Table 3).

Table 2: Number of sequencing reads remaining after each step in the bioinformatics pipeline for each sample type.

Analysis step	Isolates	Mock community	Environmental samples	Total
Raw CCS	46,740	60,448	125,988	233,176
Length-filtered	44,595	53,730	117,395	215,720
Average quality-filtered	18,532	16,054	48,778	83,364
Window quality-filtered	16,353	11,263	43,385	71,001
Primer-filtered	16,082	10,891	42,369	69,342

Table 3: Error rates in CCS reads computed by mapping to consensus sequences of isolates.

Analysis step	Substitutions mean (SD)	Insertions mean (SD)	Deletions mean (SD)	Total mean (SD)
Raw CCS	0.0453% (0.1277%)	0.3140% (0.6108%)	0.8650% (1.1960%)	1.2243% (1.5575%)
Filtered	0.0080% (0.0273%)	0.0380% (0.0476%)	0.1756% (0.1550%)	0.2216% (0.1621%)

Chimera formation and detection

Using reference-based chimera detection in the mock community, chimera formation rate (i.e. sequences classified as chimeras or as unsure) rose from <2% of sequences at 13-18 PCR cycles to 16.3% at 30 cycles (Fig. 5). The emPCR (25 cycles) resulted in 4.4% of sequences classified as chimeric (Fig. 5), compared to 14.1% for 25 cycles under standard PCR conditions. Template DNA amounts played no measurable role in chimera formation rate, with 2, 8 and 20 ng of DNA all resulting in <2% chimeric sequences (18 cycles). Manual inspection of 100 randomly chosen isolate sequences classified 16 of these as chimeras. Reference-based detection identified 15 of these as chimeric and one as “suspicious”. Of the 84 confirmed as non-chimeric by manual inspection, the reference-based algorithm classified 82 (97.6%) as non-chimeric and 2 as “suspicious”. *De novo* chimera detection (i.e., in the absence of a reference) classified 98.6% of the reads in the sample in the same way as using the reference-based approach.

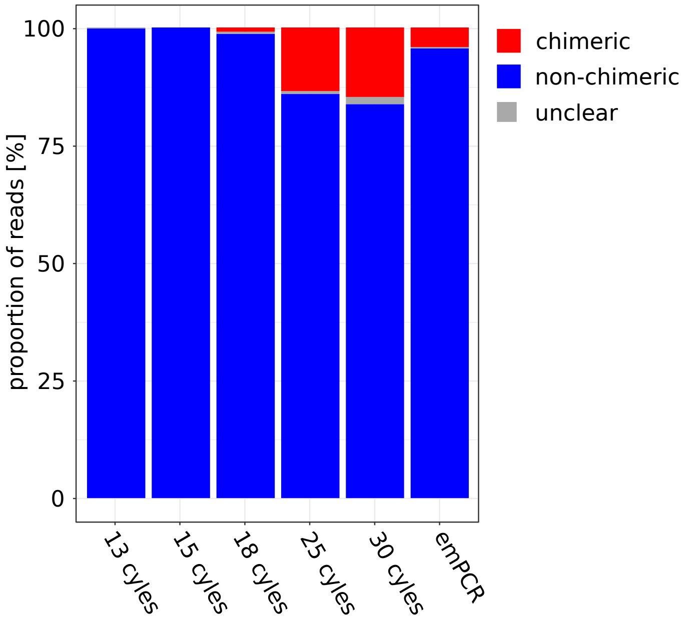


Figure 5: Chimera calls by vsearch with reference-based approach for different PCR conditions. Reads are classified as “chimeric” (red), “non-chimeric“ (blue), or in edge cases as “unclear” (gray).

Mock community classification

The five marker regions (SSU, ITS1, 5.8S, ITS2, LSU) clearly distinguished 8 of the 14 isolates we could recover within the mock community, but revealed cases of intra-specific variation as well as overlap among recognized species (Fig. 6). Seven species were clearly distinguished at all five markers, i.e. formed a single cluster for each region (Fig. 6). *Metschnikowia reukaufii* produced multiple clusters for ITS1 and ITS2, as expected based on previous reports of extraordinarily high rRNA operon variation in this genus (Lachance et al., 2002; Sipiczki et al., 2013). *Clavariopsis aquatica* and *Phoma* sp. were separated by all regions except SSU. *Trichoderma reesei* and

Clonostachys rosea were separated by ITS1, ITS2, and LSU but not with SSU and 5.8S genes. *Cladosporium herbarum* and *Cladosporium* sp. were differentiated only with the ITS2, although one of the two clusters was mixed (Fig. 6). OTU clustering resulted in 16 non-singleton OTUs. Twelve OTUs consisted of sequences from one species as well as a few chimeric sequences, one contained sequences from *Cladosporium herbarum* and the other *Cladosporium* sp., and three smaller OTUs were entirely made up of chimeric sequences (Table 4). *Mortierella elongata* and *Cystobasidium laryngis* did not appear in any OTUs, although we did observe low read abundance (<10 reads) of these species prior to quality filtering in some of the mock community samples.

OTUs were classified to varying taxonomic ranks by the three different genetic markers (Table 4). The SSU gene provided mainly order- and family-level classifications, the ITS region provided family- to species-level classifications, and the LSU gene provided genus-level classifications in some cases and higher level classifications in others. The *Metschnikowia reukaufii* OTU was classified to different species by ITS (*M. cibodasensis*) and LSU (*M. bicuspidata*). Different genus-level classifications by ITS and LSU for the Chytrid species were the result of different taxonomies used in the UNITE and the RDP databases. The best match in both databases was *Globomyces pollinis-pini*, but the higher classification at higher ranks differs among the databases. Similar discrepancies caused by differences in database taxonomy also occurred for some of the other species. Other than that classifications by all three markers were consistent with each other and with the manual classification.

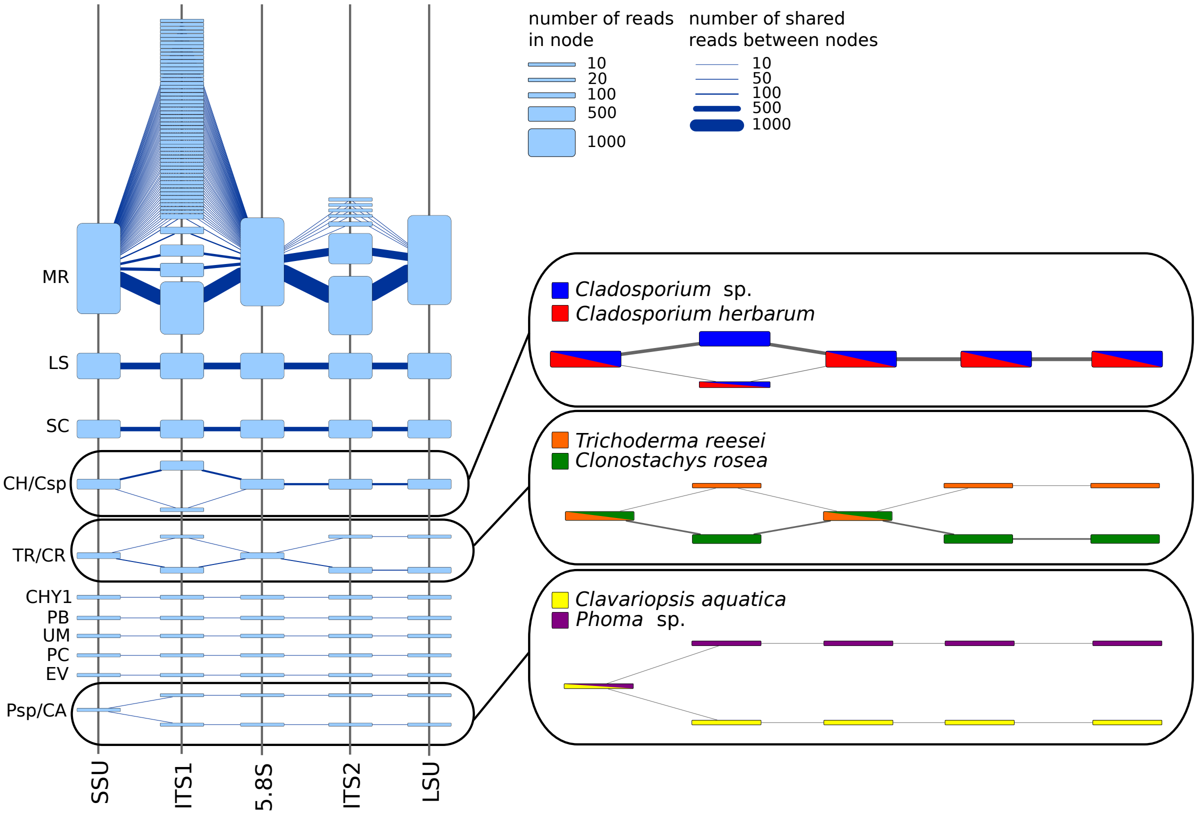


Figure 6: Resolution of different regions of the rRNA operon for our mock community. Each node represents a cluster and each edge between two clusters represents shared reads between the clusters. Node height and edge thickness is proportional to read number. Nodes and edges with less than 3 reads are not shown. Identification codes are given in Table 1. Components with multiple species are shown in detail on the right. Nodes are colored by species appearing in them. The graph was initially created with Cytoscape (version 3.2.1, Shannon et al., 2003) and manually adapted for better readability.

		Classification method			
OTU	Size	Manual	SSU	ITS	LSU
11	6	*Clavariopsis aquatica*	Pleosporales (Order)	*Clavariopsis aquatica*	Pleosporales (Order)
6	44	Chytridiomycota	Chytridiomycetes (Class)	Globomyces (Genus)	Rhizophydium (Genus)
4	344	*Cladosporium* sp. + *Cladosporium herbarum*	Cladosporium (Genus)	Cladosporium (Genus)	Davidiella (Genus)
5	140	*Clonostachys rosea*	Hypocreales (Order)	Bionectriaceae (Family)	Hypocreales (Order)
1	4165	*Metschnikowia reukaufii*	Saccharomycetales (Order)	*Metschnikowia cibodasensis*	*Metschnikowia bicuspidata*
2	1096	*Leucosporidium scottii*	Basidiomycota (Phylum)	Leucosporidiaceae (Family)	*Leucosporidium* (Genus)
3	719	*Saccharomyces cerevisiae*	Saccharomycetaceae (Family)	Saccharomyces *(Genus)*	*Saccharomyces* (Genus)
7	37	*Penicillium brevicompactum*	Trichocomaceae (Family)	Penicillium *(Genus)*	Fungi (Kingdom)
8	34	*Ustilago maydis*	Ustilaginaceae (Family)	Ustilaginaceae (Family)	*Ustilago maydis*
9	20	*Exobasidium vaccinii*	Exobasidiales (Order)	*Exobasidium vaccinii*	Exobasidium (Genus)
10	21	*Phanerochaete chrysosporium*	Agaricomycetes (Class)	*Phanerochaete* sp.	Agaricomycetes (Class)
12	5	*Phoma* sp.	Pleosporales (Order)	Pleosporales Incertae sedis (Family)	Didymellaceae (Family)
13	5	*Trichoderma reesei*	Hypocreaceae (Family)	Trichoderma (Genus)	Hypocreaceae (Family)
14	3	chimeric	Saccharomycetales (Order)	Nectriaceae (Family)	*Metschnikowia bicuspidata*
16	3	chimeric	Saccharomycetales (Order)	*Metschnikowia cibodasensis*	unknown
17	9	chimeric	Saccharomycetales (Order)	*Metschnikowia cibodasensis*	Bionectria (Genus)

Table 4: Mock-community OTU classification with our analytical pipeline. Manual classifications were made by comparison to full-length reference sequences. rRNA gene region classifications were made based on reference sequences in SILVA (SSU), UNITE (ITS) and RDP (LSU) databases. Size indicates the number of reads

Environmental community classification

OTU clustering of the environmental samples produced 947 non-singleton OTUs, of which 799 (84%) were classified as fungi by at least one of the three markers (SSU, ITS, LSU). The SSU database also allowed identification of non-fungal sequences, and 112 OTUs were assigned to Metazoa, 10 to Discicristoidea, 2 to Stramenopiles, 2 to Alveolata and 1 to Chloroplastida. The 200 most abundant fungal OTUs (91% of fungal reads; 61% of total reads) were consistently classified to phylum level by all three markers except for 9 cases in which SSU and LSU gave different classifications for the same OTU (Fig. 7). There were no conflicts between SSU and ITS, although the SILVA and UNITE databases use different names for the phylum Cryptomycota/Rozellomycota (Fig. 7). Classification at the phylum level was most successful with SSU (188 reads, i.e., 94% of the 200 most abundant fungal OTUs). Fewer OTUs were classified to phylum with LSU (126, 63%) and ITS regions (36, 18%). Classification to the species level was most successful with LSU (55, 27.5%) and less successful for ITS (20, 10%) and SSU (13, 6.5%) (Fig. 7).

Extended to all 947 OTUs, the results were similar. SSU provided the most classifications, especially for higher taxonomic ranks, and ca. 20% of these were classified the same using the ITS (Fig. 8 A) and ca. 66% were classified the same by LSU (Fig. 8 B). ITS classifications matched those of SSU (Fig. 8 C) and LSU (Fig. 8 D) at ranks from kingdom to class. At family, genus and species rank, most OTUs that were classified by ITS were not classified by SSU (Fig. 8 C) and many were classified differently by LSU (Fig. 8 D). At higher taxonomic rank (kingdom to class), OTUs classified by LSU were classified the same way as by SSU. But more than 50% were either not assigned to any taxon or were classified differently by SSU at lower ranks (order to species; Fig. 8 E). Most OTUs classified by the LSU were not classified by ITS at kingdom to class ranks (> 60%), although those that were, were classified the same. At the order to species rank, OTUs classified by both LSU and ITS were rare and differences between the markers were more common (Fig. 8 F).

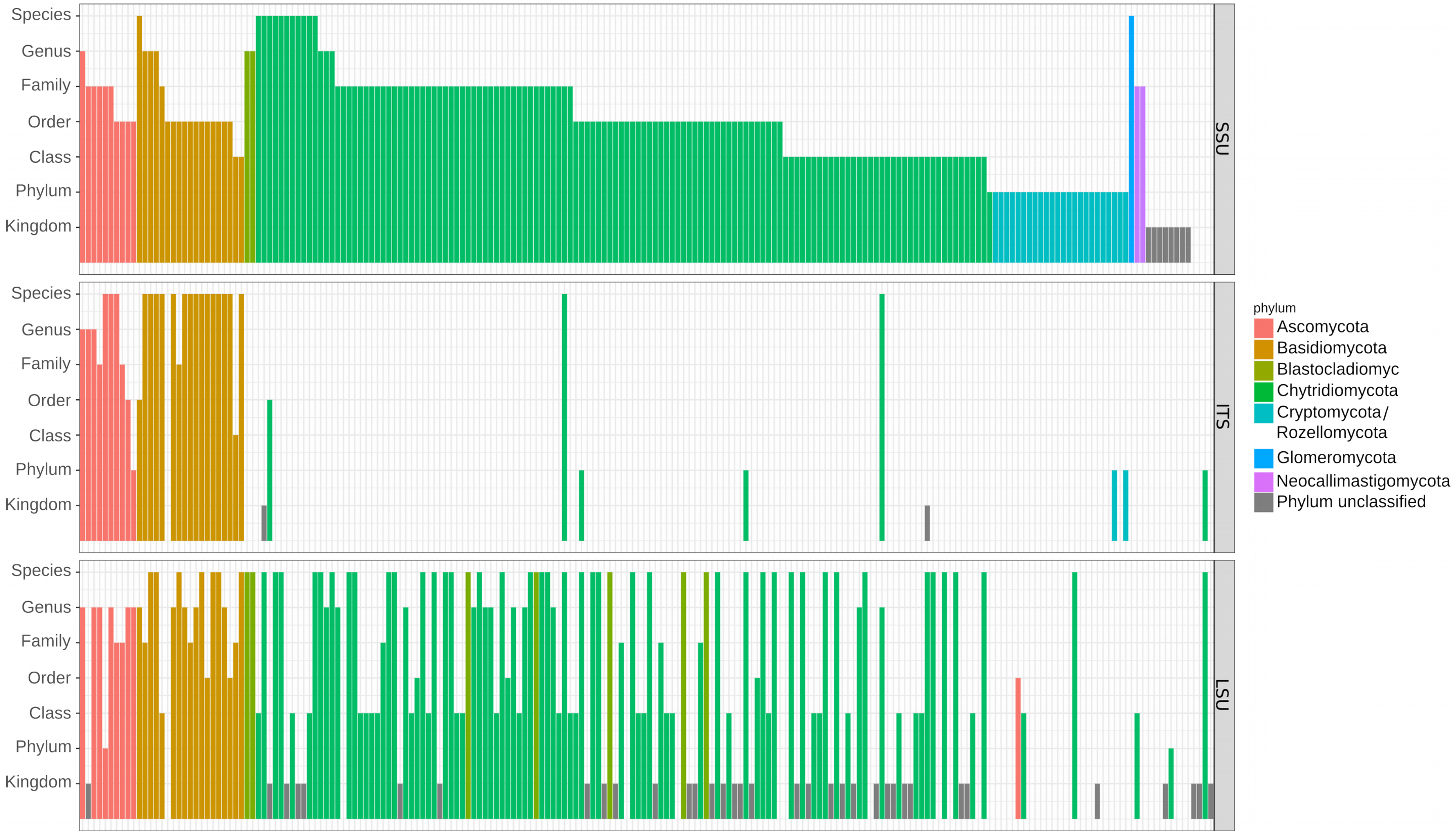


Figure 7: Classification specificity of the 200 most abundant fungal OTUs for the three different regions (SSU, ITS, LSU). The three rows give classifications by the three different regions. Each OTUs classification is given by a bar in each row. The height of the bar represents level of classification. Bars are colored by phylum.

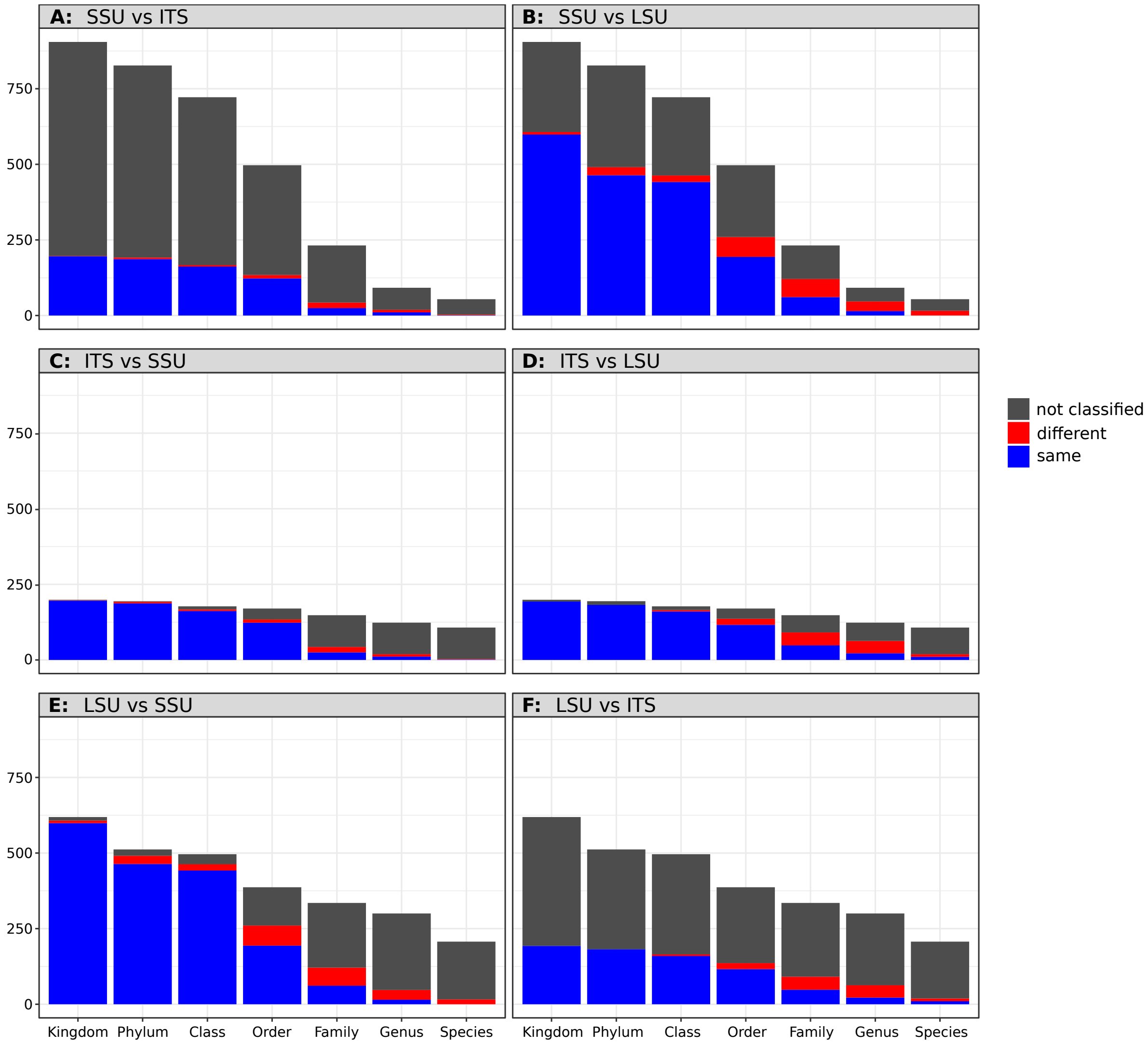


Figure 8: Agreement of classifications of all OTUs by the different regions. Each panel represents a comparison between two regions. Each set of stacked bars shows numbers of agreeing (blue), disagreeing (red) and unknown (gray) OTU classifications in the second region of the comparison compared to the first, at each taxonomic level.

5 Discussion

Long sequencing reads have the potential to provide many benefits for DNA metabarcoding. These include taxonomic assignment of OTUs at lower taxonomic levels (Franzén et al., 2015; Porter and Golding, 2011), the use of homology-based classification and phylogenetic reconstruction (e.g. Tedersoo et al., 2017), and higher sequencing quality for standard-length DNA barcodes in reference databases (Hebert et al., 2018). Disadvantages of long reads include lower sequence quality (D’Amore et al., 2016; Glenn, 2011), a possible increase in the rate of chimera formation, and the fact that most bioinformatics tools are optimized for shorter reads. Here we produced DNA metabarcodes nearly twice as long as any used to date (ca. 4,500 bp), comprising the whole eukaryotic rRNA operon (SSU, ITS, LSU). We combined circular consensus sequencing with our newly developed bioinformatics pipeline and obtained error rates comparable to short-read Illumina sequencing (D'Amore et al., 2016; Glenn, 2011). The use of multiple markers allowed us to use the ITS region for OTU delineation (clustering) and automated species-level taxonomic classifications for environmental OTUs with both ITS and LSU sequences. Finally, the inclusion of the SSU rRNA gene into the analyses allowed us to classify OTUs that were not represented in ITS and LSU databases, including many fungi that belong to basal lineages and are common in freshwater habitats (Rojas-Jimenez et al., 2017; Wurzbacher et al., 2016).

Challenges of long reads

A significant challenge in using longer reads for DNA-metabarcoding of mixed samples is the fact that most bioinformatics tools have been designed for the analysis of short sequences (typically 200-600 bp). Although we obtained very high-quality CCS reads, the higher indel rate and accumulation of errors in long reads requires analyses that differ from that of more commonly used sequencing platforms like Illumina. For example the clustering algorithm applied by swarm (Mahé et al., 2015) relies on a low total number of errors per sequence (ideally 1 error). In long sequences, even with low error rates, the total number of errors are higher, makeing it unfeasible to use this algorithm. Other widely used clustering tools like uclust (Edgar, 2010) or vsearch use heuristics to choose starting points for clustering. Reads are first de-replicated and those with the most identical copies are used as cluster starting points. This could not be applied to our data set because the comparably high nucleotide deletion rate and the long read length made almost all reads unique.

In the future it might be beneficial to develop specialized software for clustering and correcting PacBio long range amplicons. Here we used heuristic clustering starting with high quality reads and with a high similarity threshold (99%), and a consecutive consensus calling for correction of random sequencing errors. This also gave us clusters of highly similar sequences, that we could use for chimera detection and OTU clustering instead of the groups of identical reads resulting from de-replication, that are normally used for these steps.

One of the problems in any study applying PCR to mixed samples is chimera formation. Our comparison of *de novo* and reference-based chimera detection found them to produce the same classifications in > 98% of cases. This indicates that *de novo* chimera classification in our long-read pipeline provided a good estimate of chimera formation rate and is suitable for data sets where no complete reference database is available. We can therefore be confident in our results for the environmental samples, even where no reference sequences were available in databases. Interestingly, a manual inspection of conflicting read assignments in the independent clustering of the different regions (data not shown) found a few cases (9 of 6,585 reads in the one mock community sample) of chimeras that could not be detected. Neither reference-based nor *de novo* approaches detected these chimeras because 3’ and 5’ ends were both from the same species, and only the central section originated from a second species. Most chimera detection software, including vsearch, model chimeras from two origins i.e., different 3' and 5' ends, but not more. These methods would then fail to identify chimeras if the 3' and 5' ends are from the same species and a second species is in the middle, as we observed. Although this was very rare in our data (0.1% of reads investigated), it created small OTUs made up almost entirely of these complex chimeras in our mock community (OTU 14, 16 and 17, see Table 4). As a general rule, chimeras are most likely to be found associated with the most frequent sequences in a PCR sample (e.g. Sommer et al., 2013) and this is also true for the complex chimeras we observed here. In fact, all three chimeric OTUs found in our mock community involved the species with the most read abundance, *Metschnikowia reukaufii*. DECIPHER (Wright et al., 2012) is one tool that may detect these chimeras, but requires a complete reference database of possible parent sequences and is therefore unsuitable for use with environmental samples (for which reference sequences are difficult to obtain) and long reads.

We also attempted to minimize chimera formation in the laboratory, by exploring the influence of reduced PCR cycle numbers, emulsion PCR, and template concentration. Although we were initially concerned that our *ca.* 4,500 bp amplicon length would lead to higher chimera formation rates during PCR, the mock community sample that was amplified with the highest cycle number (30) formed chimeras at a rate within the range reported by short-read studies (ca. 4-36%, Ahn et al., 2012; Qiu et al., 2001). We observed reduced chimera formation with fewer cycles which is also consistent with short-read studies (D’Amore et al., 2016; Lahr and Katz, 2009; Qiu et al., 2001). Unlike other studies (D’Amore et al., 2016; Lahr and Katz, 2009) we did not find a notable influence of DNA template concentration in our samples, possibly because at 18 cycles all reactions were still in the exponential phase, before depletion of reagents (see below). Chimera formation rates in our mock community may underestimate rates in environmental samples because the lower species richness in the mock community may have led to reduced chimera formation (Fonseca et al., 2012). However, the chimera rate detected by *de novo* chimera detection in our environmental data was < 1%, i.e., even lower than the *de novo* detection rate in the less diverse mock community samples. Chimera formation occurs primarily during the saturation phase of a PCR, when a large amount of PCR product has accumulated and the template:primer ratio increases (Judo et al., 1998). For a given cycle number, the amount of accumulated product may differ between the environmental and mock community samples, because although a similar amount of template DNA was used in mock community (8 ng) and environmental (10 ng) samples, the amount of template available for primer binding might be lower in the latter because they also contain non-fungal DNA. Environmental samples may also contain more PCR inhibitors (Schrader et al., 2012), which would reduce PCR efficiency and delay the saturation phase to a higher cycle number in environmental samples compared to the mock community. Optimization of DNA extraction and amplification could make lower PCR cycle numbers feasible and thus further reduce the problem of chimera formation. Our emPCR results also indicate that this might be a promising way of reducing chimera formation when more PCR cycles are required.

Classification

Although the ITS region has been proposed as a standard barcode for fungi (Schoch et al., 2012) other regions of the rRNA operon remain popular choices as fungal barcodes (Roy et al., 2017; Stielow et al., 2015; Wurzbacher et al., 2016). Compared to rRNA genes, ITS1 and ITS2 often exhibit higher interspecific variability and thus can provide greater species delineation power (i.e., more OTUs) than SSU and (in most fungal groups) LSU (Schoch et al., 2012). Indeed we found that isolate species of the same genus (Cladosporium) and even from the same order (Hypocreales) and sub-division (Pezizomycotina) could not be separated by the SSU (Fig. 6), and that the use of ITS resulted more often in classification to species level than SSU and LSU in Dikarya (Fig. 7). At the same time, the often higher variability of ITS also means that for new species that are not represented in the database it can be more difficult to find comparable sequences and thus to identify them to any level. In these cases, longer sequencing reads that include more conserved regions with a stable evolutionary rate are likely to be helpful in making classifications based on sequence similarity as we did here or, by phylogenetic methods (e.g. Tedersoo et al., 2017). The phylum Chytridiomycota, which is often found in aquatic environments and was highly abundant in our environmental samples, is underrepresented in sequences databases (Frenken et al., 2017). We observed many OTUs from this phylum that could not be classified with the ITS at all, while the SSU provided at least class or family rank classifications and the LSU often even provided classifications at species rank (Fig. 7).

For the classification of the mock community, the different degrees of taxonomic resolution provided by the different markers were clear. The mock community consisted of species that are represented in the reference databases with sequences that were identical or very similar to the sequence that we found. In these cases, ITS was a superior marker region, since its greater variability allowed for higher resolution classification. While almost all classifications were correct, those obtained for ITS went down to at least family rank in all cases, and even to species rank for a third of the OTUs. LSU and SSU both provided far fewer specific classifications. Using the LSU marker, species levels classifications could be obtained for some OTUs, but others were only classified to higher taxonomic ranks (up to kingdom). Using the SSU marker, classification results were obtained between the ranks of order and family. In our environmental samples, the disadvantage of ITS becomes clear. If no closely related reference sequence was available, sequence similarity to any sequence in the database was too low to classify the sequence even to a higher taxonomic rank. In these cases, SSU and LSU markers provided at least classification at family or class level, while many OTUs stayed completely unclassified with the ITS.

The independent clustering of the different regions (SSU, ITS1, 5.8S, ITS2 and LSU) of the rRNA operon (Fig. 5) also showed the higher resolution of ITS1 and ITS2, which were the only regions that separate almost all species from each other. On the other hand, for *Metschnikowia reukaufii* they formed multiple clusters for one species. This is most likely the result of high variability of rRNA operon copies in *Metschnikowia* (Lachance et al., 2002; Sipiczki et al., 2013) in combination with the short ITS1 and ITS2 sequences (70 bp and 75 bp, respectively) which mean that very few (3) differences already constitute an identity difference of 3%.

Classification conflicts and synergies

The conflicts we observed between classifications based on different marker regions and databases can provide insights into a number of interesting problems. In some cases, they may either represent uncertainty in classification using at least one of the markers, or genuine chimeric reads. In other cases they may highlight incompatibility between the taxonomies used by the databases, or even errors in the databases (see also Nilsson et al., 2006). Many conflicts resulted from differences in naming convention and taxonomic placement in the different databases. Multiple OTUs were classified with LSU and the RDP database to the more recently defined orders Rhizophydiales (Letcher et al., 2006) and Lobulomycetales (Simmons et al., 2009), but were classified with SSU and the SILVA database as Chytridiales, the older classification for these new orders. A similar effect can be seen for the orders in the class Agaricomycetes. Three OTUs were assigned to the family Lachnocladiaceae which belongs to the order Russulales according to SILVA and to Polyporales according to RDP. Finally, one OTU was assigned to the genus *Jahnoporus* using the LSU marker. According to the RDP database this genus belongs to the order Russulales while in SILVA it belongs to the order Polyporales. Other conflicts showed that minor problems in the databases can lead to major differences in classification. In our environmental data, several high (read) abundance OTUs were classified as Chytridiomycota with SSU but as Blastocladiomycota with LSU. Closer inspection of the LSU alignments indicated that for many of these OTUs, only the second best hit was to a Blastocladiomycota, while the best match was, in fact, *Rhizophlyctis rosea*. The latter is a Chytridiomycota, but has no classification beyond kingdom in the RDP database file we used and was thus ignored for classification. In addition, the second best hit which was used for classification is to a sequence from the genus *Catenomyces* which belongs to the phylum Blastocladiomycota according to RDP, but according to SILVA belongs to the phylum of Chytridiomycota. Thus a minor error in the database file, in combination with inconsistencies in the taxonomy used by different databases, can lead to completely different classifications when using different markers.

These conflicts in classification clearly highlight problems with the databases, but classifications using three different markers from the same molecule, as obtained from the full rRNA operon, can help us to evaluate how confident we can be in our classification. A classification that is supported by three markers, with largely independent databases, can be considered more trustworthy than one that is only supported by one, or even shows conflicts when using different markers. In addition, long DNA barcodes could be used to create synergies between the databases and to support short read studies. For example, if a sequence was classified to the same family by SSU (SILVA) and LSU (RDP), the ITS region could be added to the Unite database (even if it is not classified to the species level) to help future studies that use ITS markers. The possibility to sequence SSU, ITS and LSU at the same time therefore offers the opportunity to contribute to different databases in parallel, with the future potential to generate a new reference data set with nearly full-length rRNA operon sequences.

Conclusions

We used a DNA metabarcode nearly twice the length of any used to date and created a long-read (*ca*. 4,500 bp) bioinformatics pipeline that results in rates of sequencing error and chimera detection that are comparable to typical short-read analyses. The approach enabled the use of three different rRNA gene reference databases, thereby providing significant improvements in taxonomic classification over any single marker. While ITS is likely to remain a short-metabarcode region of choice for some time, a clear limitation of ITS is that its high variability, in combination with the incompleteness of databases, often lead to classification failing. In these cases, the other rRNA markers are beneficial. In particular, classification based on SSU or LSU were superior in more basal fungal groups. The universal nature of the rRNA operon and our recovery of >100 non-fungal OTUs indicate that the method could also be suitable for more general studies of eukaryotic biodiversity.

6 Acknowledgements

We thank Lars Ganzert, Katrin Premke, and Robert Taube (IGB) for help with field sampling, Keilor Rojas and Silke Van den Wyngaert (IGB) for providing isolates, Christian Wurzbacher (Univ. Gothenberg, now TU Munich) for providing primers, and Nicole Heyer and Simone Severitt (DSMZ) for help with sequencing. Research was partially funded by the Leibniz Association Pakt/SAW project “MycoLink” (SAW-2014-IGB-1).

7 Author Contribution

F.H., E.CB., C.B., A.Y., J.O., C.J.M. and M.T.M. conceived and designed the overall study design. E.C.B. designed and performed molecular laboratory work. **F.H.** designed and implemented the analysis pipeline and carried out analysis. B.B. and C.S. advised on sequencing strategy and performed library preparation and sequencing. E.C.B, C.B. and A.Y. chose and cultivated isolates for the mock community. **F.H.**, E.C.B., C.J.M. and M.T.M. wrote the first draft, and all authors contributed to the final manuscript.

8 References

Ahn, J.-H., Kim, B.-Y., Song, J., and Weon, H.-Y. (2012). Effects of PCR cycle number and DNA polymerase type on the 16S rRNA gene pyrosequencing analysis of bacterial communities. J Microbiol. *50*, 1071–1074.

Bengtsson-Palme, J., Ryberg Martin, Hartmann Martin, Branco Sara, Wang Zheng, Godhe Anna, Wit Pierre, Sánchez García Marisol, Ebersberger Ingo, Sousa Filipe, et al. (2013). Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. Methods in Ecology and Evolution *4*, 914–919.

Blaalid, R., Kumar, S., Nilsson, R.H., Abarenkov, K., Kirk, P.M., and Kauserud, H. (2013). ITS1 versus ITS2 as DNA metabarcodes for fungi. Mol Ecol Resour *13*, 218–224.

Boers, S.A., Hays, J.P., and Jansen, R. (2015). Micelle PCR reduces chimera formation in 16S rRNA profiling of complex microbial DNA mixtures. Scientific Reports *5*.

Chaisson, M.J., and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics *13*, 238.

Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., and Tiedje, J.M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res *42*, D633–D642.

D'Amore, R., Ijaz, U.Z., Schirmer, M., Kenny, J.G., Gregory, R., Darby, A.C., Shakya, M., Podar, M., Quince, C., and Hall, N. (2016). A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. BMC Genomics *17*, 55.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460–2461.

Fonseca, V.G., Nichols, B., Lallias, D., Quince, C., Carvalho, G.R., Power, D.M., and Creer, S. (2012). Sample richness and genetic diversity as drivers of chimera formation in nSSU metagenetic analyses. Nucleic Acids Res *40*, e66–e66.

Franzén, O., Hu, J., Bao, X., Itzkowitz, S.H., Peter, I., and Bashir, A. (2015). Improved OTU-picking using long-read 16S rRNA gene amplicon sequencing and generic hierarchical clustering. Microbiome *3*.

Frenken, T., Alacid Elisabet, Berger Stella A., Bourne Elizabeth C., Gerphagnon Mélanie, Grossart Hans Peter, Gsell Alena S., Ibelings Bas W., Kagami Maiko, Küpper Frithjof C., et al. (2017). Integrating chytrid fungal parasites into plankton ecology: research gaps and needs. Environmental Microbiology *19*, 3802–3822.

Glenn, T.C. (2011). Field guide to next-generation DNA sequencers. Molecular Ecology Resources *11*, 759–769.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics *17*, 333–351.

Hauswedell, H., Singer, J., and Reinert, K. (2014). Lambda: the local aligner for massive biological data. Bioinformatics *30*, i349-355.

Hebert, P.D.N., Braukmann, T.W.A., Prosser, S.W.J., Ratnasingham, S., deWaard, J.R., Ivanova, N.V., Janzen, D.H., Hallwachs, W., Naik, S., Sones, J.E., et al. (2018). A Sequel to Sanger: amplicon sequencing that scales. BMC Genomics *19*, 219.

Judo, M.S., Wedel, A.B., and Wilson, C. (1998). Stimulation and suppression of PCR-mediated recombination. Nucleic Acids Res *26*, 1819–1825.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics *28*, 1647–1649.

Kõljalg, U., Nilsson R. Henrik, Abarenkov Kessy, Tedersoo Leho, Taylor Andy F. S., Bahram Mohammad, Bates Scott T., Bruns Thomas D., Bengtsson Palme Johan, Callaghan Tony M., et al. (2013). Towards a unified paradigm for sequence based identification of fungi. Molecular Ecology *22*, 5271–5277.

Köster, J., and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. Bioinformatics *28*, 2520–2522.

Lachance, M.-A., Bowles, J.M., and Starmer, W.T. (2002). Metschnikowia santaceciliae, Candida hawaiiana, and Candida kipukae, three new yeast species associated with insects of tropical morning glory. FEMS Yeast Research *3*, 97–103.

Lahr, D.J.G., and Katz, L.A. (2009). Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. BioTechniques *47*, 857–866.

Laver, T.W., Caswell, R.C., Moore, K.A., Poschmann, J., Johnson, M.B., Owens, M.M., Ellard, S., Paszkiewicz, K.H., and Weedon, M.N. (2016). Pitfalls of haplotype phasing from amplicon-based long-read sequencing. Sci Rep *6*.

Letcher, P.M., Powell, M.J., Churchill, P.F., and Chambers, J.G. (2006). Ultrastructural and molecular phylogenetic delineation of a new order, the Rhizophydiales (Chytridiomycota). Mycological Research *110*, 898–915.

Lindahl, B.D., Nilsson, R.H., Tedersoo, L., Abarenkov, K., Carlsen, T., Kjøller, R., Kõljalg, U., Pennanen, T., Rosendahl, S., Stenlid, J., et al. (2013). Fungal community analysis by high-throughput sequencing of amplified markers--a user’s guide. New Phytol. *199*, 288–299.

Mahé, F., Rognes, T., Quince, C., Vargas, C. de, and Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ *3*, e1420.

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

Monchy, S., Sanciu, G., Jobard, M., Rasconi, S., Gerphagnon, M., Chabé, M., Cian, A., Meloni, D., Niquil, N., Christaki, U., et al. (2011). Exploring and quantifying fungal diversity in freshwater lake ecosystems using rDNA cloning/sequencing and SSU tag pyrosequencing. Environmental Microbiology *13*, 1433–1453.

Nercessian, O., Noyes, E., Kalyuzhnaya, M.G., Lidstrom, M.E., and Chistoserdova, L. (2005). Bacterial Populations Active in Metabolism of C1 Compounds in the Sediment of Lake Washington, a Freshwater Lake. Appl Environ Microbiol *71*, 6885–6899.

Nilsson, R.H., Ryberg, M., Kristiansson, E., Abarenkov, K., Larsson, K.-H., and Kõljalg, U. (2006). Taxonomic Reliability of DNA Sequences in Public Sequence Databases: A Fungal Perspective. PLOS ONE *1*, e59.

Nilsson, R.H., Taylor, A.F.S., Adams, R.I., Baschien, C., Bengtsson-Palme, J., Cangren, P., Coleine, C., Daniel, H.-M., Glassman, S.I., Hirooka, Y., et al. (2018). Taxonomic annotation of public fungal

ITS sequences from the built environment – a report from an April 10–11, 2017 workshop (Aberdeen, UK). MycoKeys *28*, 65–82.

Ohsowski, B.M., Zaitsoff, D.P., Öpik, M., and Hart, M.M. (2014). Where the wild things are: looking for uncultured Glomeromycota. New Phytologist *204*, 171–179.

Porras-Alfaro, A., Liu, K.-L., Kuske, C.R., and 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*, 829–840.

Porter, T.M., and Golding, B.G. (2011). Are similarity or phylogeny based methods more appropriate for classifying internal transcribed spacer (ITS) metagenomic amplicons? New Phytologist *192*, 775–782.

Qiu, X., Wu, L., Huang, H., McDonel, P.E., Palumbo, A.V., Tiedje, J.M., and Zhou, J. (2001). Evaluation of PCR-Generated Chimeras, Mutations, and Heteroduplexes with 16S rRNA Gene-Based Cloning. Appl Environ Microbiol *67*, 880–887.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res *41*, D590–D596.

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

Rojas-Jimenez, K., Wurzbacher, C., Bourne, E.C., Chiuchiolo, A., Priscu, J.C., and Grossart, H.-P. (2017). Early diverging lineages within Cryptomycota and Chytridiomycota dominate the fungal communities in ice-covered lakes of the McMurdo Dry Valleys, Antarctica. Sci Rep *7*.

Roy, J., Reichel, R., Brüggemann, N., Hempel, S., and Rillig, M.C. (2017). Succession of arbuscular mycorrhizal fungi along a 52-year agricultural recultivation chronosequence. FEMS Microbiol Ecol *93*.

Schlaeppi, K., Bender, S.F., Mascher, F., Russo, G., Patrignani, A., Camenzind, T., Hempel, S., Rillig, M.C., and van der Heijden, M.G.A. (2016). High-resolution community profiling of arbuscular mycorrhizal fungi. New Phytol. *212*, 780–791.

Schloss, P.D., Jenior, M.L., Koumpouras, C.C., Westcott, S.L., and Highlander, S.K. (2016). Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. PeerJ *4*, e1869.

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., Chen, W., and Consortium, F.B. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. PNAS *109*, 6241–6246.

Schrader, C., Schielke, A., Ellerbroek, L., and Johne, R. (2012). PCR inhibitors – occurrence, properties and removal. Journal of Applied Microbiology *113*, 1014–1026.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res *13*, 2498–2504.

Simmons, D.R., James, T.Y., Meyer, A.F., and Longcore, J.E. (2009). Lobulomycetales, a new order in the Chytridiomycota. Mycological Research *113*, 450–460.

Singer, E., Bushnell, B., Coleman-Derr, D., Bowman, B., Bowers, R.M., Levy, A., Gies, E.A., Cheng, J.-F., Copeland, A., Klenk, H.-P., et al. (2016). High-resolution phylogenetic microbial community profiling. The ISME Journal *10*, 2020–2032.

Sipiczki, M., Pfliegler, W.P., and Holb, I.J. (2013). Metschnikowia Species Share a Pool of Diverse rRNA Genes Differing in Regions That Determine Hairpin-Loop Structures and Evolve by Reticulation. PLOS ONE *8*, e67384.

Sommer, S., Courtiol, A., and Mazzoni, C.J. (2013). MHC genotyping of non-model organisms using next-generation sequencing: a new methodology to deal with artefacts and allelic dropout. BMC Genomics *14*, 542.

Stielow, J.B., Lévesque, C.A., Seifert, K.A., Meyer, W., Iriny, L., Smits, D., Renfurm, R., Verkley, G.J.M., Groenewald, M., Chaduli, D., et al. (2015). One fungus, which genes? Development and assessment of universal primers for potential secondary fungal DNA barcodes. Persoonia *35*, 242–263.

Tedersoo, L., Bahram, M., Põlme, S., Kõljalg, U., Yorou, N.S., Wijesundera, R., Ruiz, L.V., Vasco-Palacios, A.M., Thu, P.Q., Suija, A., et al. (2014). Global diversity and geography of soil fungi. Science *346*, 1256688.

Tedersoo, L., Ave, T.-K., and Anslan Sten (2017). PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. New Phytologist *217*, 1370–1385.

Travers, K.J., Chin, C.-S., Rank, D.R., Eid, J.S., and Turner, S.W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res *38*, e159.

Wright, E.S., Yilmaz, L.S., and Noguera, D.R. (2012). DECIPHER, a Search-Based Approach to Chimera Identification for 16S rRNA Sequences. Appl. Environ. Microbiol. *78*, 717–725.

Wurzbacher, C., Warthmann, N., Bourne, E., Attermeyer, K., Allgaier, M., Powell, J.R., Detering, H., Mbedi, S., Grossart, H.-P., and Monaghan, M.T. (2016). High habitat-specificity in fungal communities in oligo-mesotrophic, temperate Lake Stechlin (North-East Germany). MC *16*, 17–44.

Wurzbacher, C., Larsson, E., Bengtsson-Palme, J., Van den Wyngaert, S., Svantesson, S., Kristiansson, E., Kagami, M., and Nilsson, R.H. (2018). Introducing ribosomal tandem repeat barcoding for fungi.

V Identification of enzymes for lignocellulose degradation in *Clavariopsis aquatica*

Felix Heeger, Elizabeth C. Bourne, Christian Wurzbacher, Elisabeth Funke, Anna Lipzen, Igor Grigoriev, Vivian Ng, Guifen He, Dietmar Schlosser, Michael T. Monaghan

1 Abstract

Fungi are ecologically very important decomposers of lignocellulose. Besides “white rot” and “brown rot” Basideomycota which use different peroxidases, laccases and proteins of the cytochorme P450 super-family to degrade lignin to access cellulose and hemicellulose, limited lignin modification capabilities have also been reported for terrestrial Ascomycota. Here we investigated the presence of proteins for the modification of lignin and its constituents in the genome of the exclusively aquatic Ascomycota (hymphomycete) *Clavariopsis aquatica*. In addition we measured differential gene expression of C. aquatica when grown on lignocellulosic substrates compared to growth on a sugar rich substrate. We found differential expression of potential peroxidases, laccases and cytochrome P450s, as well as significant over representation of proteins for cellulose and hemicellulose degradation among the differential expressed genes. This observation strongly suggests that *C. aquatica* is able to modify lignin to some extent; perhaps in order to facilitate the utilization of lignocellulose as a carbon and energy source.

2 Introduction

Fungi are ecologically very important decomposers of lignocellulose, which is the main component of plant cell walls. Lignocellulose contains cellulose, hemicellulose, and lignin. Lignin, the most recalcitrant among these polymeres, is composed of phenylpropanoid monomer units and appears in the highest proportion in wood. Wood decay is prominently exclusively mediated by fungi from the phylum of Basidiomycota, and of these, only the so called white rot fungi are able to fully mineralize lignin, while brown rot fungi can only modify it to some extent. White rot fungi express multiple groups of lignin modifying enzymes like laccases and peroxidases (Lundell, Mäkelä, and Hildén 2010). Studies of full genomes of white and brown rot fungi show that the two groups are neither monophyletic nor easily separated by their protein repertoire (Riley et al. 2014; Floudas et al. 2012).

In other plant material that is not as lignin rich as wood other fungi can also degrade cell walls and play an important role in plant matter degradation. One such example is leaf litter, that is submerged in streams, where the most abundant species are aquatic hyphomycetes from the phylum Ascomycota (Duarte et al. 2015; Kubicek and Druzhinina 2007; Voříšková and Baldrian 2013). It is generally accepted, that aquatic hyphomycetes can degrade cellulose and hemicellulose from plant litter, while their ability do degrade lignin is limited at best (Gessner et al. 2007; Krauss et al. 2011). The comparative study of genomes of white and brown rot fungi has given insights into the protein families that are important for wood degradation (Riley et al. 2014; Floudas et al. 2012; Frommhagen et al. 2017) and the study of gene expression during wood degradation (e.g. Yang et al. 2012; Tang et al. 2013) offers even more fine grain insights into which proteins are produced. Studies of terrestrial Ascomycota have shown that they also possess and express genes for cellulose and hemicellulose degradation (Ries et al. 2013).

Clavariopsis auquatica is a typical aquatic Ascomycote (hyphomycete) colonizing leaf litter in streams (Iqbal and Webster 1973; Suberkropp and Klug 1976). This fungus has previously been reported to biotransform environmental pollutants such as nonylphenol and polycyclic musk fragrances in a cometabolic manner, hereby involving both extracellular laccase and intracellular oxidation reactions indicative for the action of cytochrome P450 systems (Junghanns et al. 2005; Krauss et al. 2011; M. Martin 2011; C. Martin et al. 2007).

Our aims in this study were to search for these previously described proteins and to identify peroxidases known to act on ligoncellulose components in other fungi. In addition we examined over representation of CAZy and KEGG annotations in differentially expressed genes in an effort to identify critical fungal pathways that may be involved in carbon decomposition. We assembled its genome and identified multiple laccases and peroxidases, as well as enzymes of the cytochrome P450 super-family.

Because of the various and sometimes multiple functions of these protein families and the frequent occurrence of their members in many different forms, the function of the identified genes can not easily be inferred from their sequence alone. To get further insights into which of them are involved in plant cell wall degradation, we also investigated changes in gene expression during cultivation of *C. aquatica* on two different plant materials, and also in dependence on its growth stage. Common alder (*Alnus glutinosa*) leaves were used, because they represent a possible natural substrate of the fungus in rivers. Wheat straw was applied as another natural lignocellulosic substrate typically not found in *C. aquatica* habitats. It typically possesses a clearly higher cellulose (~40%, Bjerre et al.

1996; Alemdar and Sain 2008) content than alder leaves (5-15%, Chauvet 1987; Lecerf and Chauvet 2008) and also contains a substantial proportion of lignin (9-22%, Bjerre et al. 1996; Alemdar and Sain 2008). As a control medium, we used malt extract, a mainly sugar-based substrate being essentially devoid of phenolic and further aromatic constituents. The growth phase has previously been shown to play an important role in laccase regulation (Solé et al. 2012). Liquid culturing with milled plant material allows for clear differentiation between exponential and stationary growth phase and was applied in this study. To investigate more natural conditions, additional cultures were grown on solid substrate (i.e. not milled).

The aim of our study was to provide insights into major protein families expressed by *C. aquatica* during colonization of differently composed natural lignocellulosic substrates, and their potential functions in carbohydrate and hydrocarbon metabolism. To our knowledge this is the first combined genome and gene expression study of an exclusively aquatic fungus.

3 Methods

Cultivation

Liquid cultivations of *C. aquatica* were carried out in 500-mL flasks containing 200 mL of medium. For cultivation on alder leaves and wheat straw, 10 g/L milled alder leaves and wheat straw, respectively (particle size about 2-4 mm), were autoclaved (121°C, 20 min) twice and suspended in a nitrogen-limited medium previously described for manganese peroxidase production in *Stropharia rugosoannulata* (glucose, which was used as a carbon source in the original medium composition, was omitted) (Schlosser and Höfer 2002). Control cultures were grown on liquid malt extract medium (1% malt extract, w/v; pH 5.6-5.8) (Solé et al. 2012). The flasks were inoculated with 5 mL of a mycelial suspension of the fungus prepared in sterile water (Junghanns et al. 2005). Fungal cultures were agitated at 120 rpm and incubated at 14°C in the dark. Flasks were harvested after 7 (trophophase) and 20 days of cultivation (stationary gowth phase) (Junghanns et al. 2005; Solé et al. 2012), and kept frozen at -80°C until RNA extraction.

For cultivation on solid wheat straw, 100-mL-flasks were supplemented with 2 g (dry mass) of milled wheat straw (about 2-4 mm particle size) and 8 mL of tap water, and autoclaved (121°C, 20 min) twice. The flasks were inoculated with 6 mycelia-containing agar plugs (derived from the edge of *C. aquatica* colonies on malt agar plates; (Junghanns et al. 2005)), and incubated without agitation at 14°C in the dark. Flasks were harvested after 26 days of cultivation, and kept frozen at -80°C until RNA extraction.

Genome Sequencing

Whole genome shotgun reads of the *C. aquatica* genome were available from an earlier project (Wurzbacher C., unpublished). Briefly, DNA was subjected to a NexteraXT library preparation (Illumina Inc.) and sequenced on a MiSeq instrument with the v3 chemistry (Illumina Inc.), after library verification with a Nano Kit (Illumina Inc.).

RNA Sequencing

Frozen material from each sample was ground to a fine powder using an RNase-cleaned and pre-cooled pestle and mortar and liquid Nitrogen, with a small spatula of Zirconium beads (Biospec, USA) added for additional friction. RNA extraction followed protocol 8 from (Johnson et al. 2012), using the CTAB-based extraction buffer from protocol 3. Briefly, for each sample c. 500 mg of ground, frozen tissue was added to 1.4 ml pre-heated (65 °C) CTAB buffer, vortexed until thoroughly mixed, incubated at 65 °C for 10-15 min, and centrifuged at 13,000 g for 3 min. The supernatant was transferred to a new 2 ml tube for two rounds of chloroform:isoamyl (24:1) extraction, a single phenol-chloroform extraction (5:1, pH 4.5), and a final chloroform:isoamyl (24:1) extraction. Following centrifugation, the upper phase was transferred to a new 2 ml tube. Purification was performed using the RNeasy® Mini Kit (Qiagen, Germany), with on-column DNA digestion (RNase-free DNase set, Qiagen), following the manufacturer’s guidelines. RNA was eluted using 30 µl of elution buffer added directly to the membrane and spun at 13,000 g for 1 min.

Total RNA was quantified using the QuantiFlour RNA system (Promega, USA), The presence of DNA was checked using the QuantiFlour DNA system, and samples with remaining DNA (D1, D3, D4) underwent an additional post-extraction DNAse I treatment. Integrity of the RNA was assessed with the Agilent RNA 6000 Nano Kit and Agilent 2100 Bioanalyzer (Agilent Technologies, USA) following manufacturer’s guidelines. The RNA integrity (RIN) value was determined for each sample as the ratio of the large to small ribosomal RNA subunits, and used as a proxy of the overall quality of the RNA sample. We also assessed the quality of the overall trace by eye. Samples with RIN values greater than 6 and determined to have good quality on the trace were sent on dry ice for sequencing. Multiple extractions were performed for each sample and pooled to obtain sufficient RNA for sequencing (minimum 2 µg per sample).

RNA library preparation and sequencing was performed at the DOE Joint Genome Institute in Walnut Creek, CA, USA. Stranded cDNA libraries were generated using the Illumina Truseq Stranded RNA LT kit. mRNA was purified from 1 ug of total RNA using magnetic beads containing poly-T oligos. mRNA was fragmented and reverse-transcribed using random hexamers and SSII (Invitrogen) followed by second strand synthesis. The fragmented cDNA was treated with end-pair, A-tailing, adapter ligation, and 8 cycles of PCR. qPCR was used to determine the concentration of the libraries. Libraries were sequenced on the Illumina Hiseq.

Genome Assembly

Reads were digitally normalized with khmer (0.7.1, Crusoe et al. 2015). In a first step, reads were normalized to a coverage of 20 (Brown et al. 2012). After removal of low-abundance kmers (Q. Zhang et al. 2014), another round of normalization to a coverage of 5 was applied (see Appendix 1, Supplemental Info 1) and only read pairs where none of the reads was removed were used for assembly. Assembly was performed with velvet (version 1.2.10, Zerbino and Birney 2008) and run with different kmer lengths k (see Appendix 1, Supplemental Info 1 for details) and k=27 was chosen, because it resulted in the highest N50 score. To estimate genome completeness we ran BUSCO (version 3.0.2, Simão et al. 2015) with the pezizomycotina reference set of single-copy genes. The *clean* command of the funannotate pipeline (version 1.2.0, Palmer 2018) was used to remove contigs shorter than 500 bp and redundant contigs.

Genome Annotation

Transcripts were assembled *de novo* from RNA-Seq data with Trinity (version 2.5.1, Grabherr et al. 2011). Trinity was configured to use trimmomatic for trimming and do digital normalization (see Appendix 1, Supplemental Info 1 for further details). Normalized RNA-Seq reads as produced by Trinity were mapped to the genome contigs with Star (version 2.5.3a, Dobin et al. 2013) using default parameters. The mapped reads were used to generate a genome-guided assembly with Trinity (see Appendix 1, Supplemental Info 1 further details). The pasa pipeline (version 2.2.0, Haas et al. 2003) was used to combine *de novo* and genome guided assembled transcripts into a single gff file as evidence for annotation (Apendix 3, Supplemental Info 1). The resulting gff file together with genome-guided assembled transcripts and mapped reads were used as input for the *predict* command of the funannotate pipeline. The *update* command of funannotate was then used to add UTR annotations. The predicted protein sequences were used as input for Interproscan (version 5.27, Jones et al. 2014) to generate Interpro (Finn et al. 2017) as well as Gene Ontology (GO, Ashburner et al. 2000; The Gene Ontology Consortium 2017) annotations. The *annotate* command of the funannotate pipeline was used to combine Interproscan results CAZy (Lombard et al. 2014) annotations from dbCAN (version 6.0, Yin et al. 2012).

In addition to the annotations from the funannotate pipeline (above), proteins were assigned as either secreted or not secreted using signalP (version 4.1, Petersen et al. 2011) and to KEGG (Kyoto Encyclopedia of Genes and Genoms, Kanehisa et al. 2016) pathways via assignment to KEGG Orthology groups (see snakemake workflow for details) with the BlastKOALA web service (Kanehisa, Sato, and Morishima 2016).

Besides general genome annotation, we specifically searched for gene families known to be involved in lignin degradation. We performed a blast search against the newly assembled *C. aquatica* genome described above to check for five previously described partially sequenced laccase genes (Solé et al. 2012). In addition we identified multicopper oxidases by assignment to the CAZy family AA1. They were further classified by blast search in the Laccase and Multicopper Oxidase Engineering Database (version 6.4, Sirim et al. 2011). We identified possibly relevant peroxidases by annotation with the Interpro family IPR001621 (Fungal ligninase) and verified the resulting proteins by annotation with the Peroxiscan web service (accessed May 15th 2018) of PeroxiBase (Fawal et al. 2013). Proteins possibly belonging to the cytochrome P450 family were identified by annotation with the Interpro family IPR001128 (Cytochrome P450).

Differential Expression and MGSA Analysis

Read counts per gene were generated with RSEM (version 1.3, Li and Dewey 2011) using default parameters. All of the following analyses were implemented as a snakemake (version 3.5.4, Köster and Rahmann 2012) workflow that can be found at www.github.com/f-heeger/caquatica_expression. RSEM output files were combined into a single read count matrix with the *merge_RSEM_output_to_matrix.pl* script from Trinity. Differential gene expression between different samples was modeled with the DESeq2 (version 1.10.1, Love, Huber, and Anders 2014) R package. Genes with an adjusted p-value < 0.05 and an absolute $\log_2$ fold change > 1 were considered to be differential expressed.

Multiple Gene Set Activation (MGSA) analysis uses a Bayesian network approach to predict a probability of activation for sets of genes for each comparison (e.g., straw – alder) based on differentially expressed genes (Bauer, Gagneur, and Robinson 2010; Bauer, Robinson, and Gagneur 2011). We defined gene sets in three ways for the MGSA analysis using three different annotations: (1) all genes annotated with one GO term, (2) all genes assigned to one CAZy family, and (3) all genes assigned to one KEGG pathway. The activation probability cut off , above that a gene set is considered to be “activated”, is ultimately arbitrary. The authors of the method suggest to use 0.5 (Bauer, Gagneur, and Robinson 2010), reasoning that this means the gene set is “more likely to be on than to be off”. We chose a slightly more conservative cutoff of 0.6. We note that activation is a statistical term here indicating that differential expression of genes in these sets can be best explained by some form of regulation of these sets, given the Bayesian model underlying the MGSA analysis.

4 Results

Genome Assembly and Annotation

We obtained 29.25 million read pairs and assembled them into 2,650 non-redundant contigs (longer than 500bp) with a N50 score of 30,079 bp and a total length of 34.18 Mb. These included complete single copies of 94.8% of the expected single-copy genes, indicating good completeness of our assembly. A total of 12,100 proteins were predicted by the funnanotate pipeline, of which 6,128 (50.64%) were annotated with at least one GO term, 2,322 (19.19%) were assigned to at least one KEGG pathway, and 572 (4.73%) to at least one CAZy family. 5,724 (47.31%) proteins did not receive any annotation from these databases.

All five of the previously described laccase gene sequences (Solé et al. 2012) were present in our genome, with nucleotide identity >98%. Based on annotation with CAZy auxiliary activity family AA1, we identified all five known laccases and eight additional multicopper oxidases. They all exhibited a high degree of similarity (53-100% pairwise amino acid identity) for conserved sites of laccase genes (Kumar et al. 2003) in a multiple alignment (data not shown). Of the previously described laccases one (lcc2) was classified as belonging to the “Basidomycete Laccase” super family by blast search against the Laccase and Multicopper Oxidase Engineering Database. The other four were assigned to the super family “Ascomycete MCO”. Of the newly identified potential multicopper oxidases five were assigned to “Ascomycete MCO” as well, while the other three were classified as “Fungal Ferroxidase” and will not be considered further.

Based on annotation with the Interpro family IPR001621, we identified 6 peroxidases, which were all verified as Class II peroxidases by Peroxiscan and identified as Asco Class II type A (2 cases), Asco Class II type B (1 case) and Asco Class II type C (3 cases) peroxidases.

A total of 137 proteins were identified as belonging to the cytochrome P450 super-family by annotation with the Interpro family IPR001128.

RNA-Sequencing and Differential Expression

We obtained 317.18 million RNAseq reads in total with > 14 million reads for each sample (see table 5), which were deposited in the NCBI Sequence Read Archive under the IDs PRJNA440444 - PRJNA440457. 75.27% (SD 1.33%) of the reads for each sample could be mapped to the newly assembled *C. aquatica* genome with RSEM. Two samples (liquid culture, exponential phase on straw) had considerably more reads than the rest (50.50 and 57.49 million). Sub-sampling to 17 million (rounded mean number of reads in the other samples) reads and re-running RSEM mapping and differential expression analysis with DESeq2 showed only minor differences (97.55% genes with the same expression status). Because of this result and considering that read count per sample is accounted for in the DESeq2 model, we used all original reads for further analyses. We modeled differential expression between recalcitrant and rich media (wheat straw versus malt extract and alder versus malt extract), and for wheat straw between growth phases (stationary versus exponential) and method of culture (solid culture versus exponential growth in liquid culture and solid culture versus stationary growth in liquid culture).

Table 5: C. aquatica samples grown under different conditions.

sample code	culture	medium	growth phase	number of reads
A3	liquid	wheat straw	exponential	50,666,214
A4	liquid	wheat straw	exponential	57,490,827
A6	liquid	wheat straw	stationary	16,839,738
A7	liquid	wheat straw	stationary	17,367,382
A9	liquid	wheat straw	stationary	17,584,532
B1	liquid	alder leaves	exponential	17,922,153
B3	liquid	alder leaves	exponential	16,700,439
B5	liquid	alder leaves	exponential	14,857,484
D1	liquid	malt extract	exponential	17,796,408
D3	liquid	malt extract	exponential	17,863,281
D4	liquid	malt extract	exponential	19,469,408
E2	solid	wheat straw	NA*	16,859,013
E3	solid	wheat straw	NA*	16,904,004
E4	solid	wheat straw	NA*	18,861,228

* Growth phases appear simultaneously in solid culture and were not separated.

Growth on straw in solid culture compared to stationary growth on straw in liquid culture had the most differentially expressed genes, while stationary compared to exponential growth on straw in liquid culture had the least (see table 6). The differentially expressed genes when growing on the alder leaves and wheat straw compared to the control medium showed a significant (fisher exact test $p<10^{-192}$) overlap.

Table 6: Number of up- and down-regulated genes for different comparisons

comparison		differential expression	
condition 1	condition 2	up-regulated	down-regulated
exponential growth on **wheat straw** in liquid culture	exponential growth on **malt extract** in liquid culture	1,430	1,570
exponential growth on **alder leaves** in liquid culture	exponential growth on **malt extract** in liquid culture	1,033	1,462
exponential growth on wheat straw in liquid culture	**stationary growth** on wheat straw in liquid culture	1,380	883
growth on wheat straw in **solid culture**	**exponential growth** on wheat straw in **liquid culture**	2,731	2,328
growth on wheat straw in **solid culture**	**stationary growth** on wheat straw in **liquid culture**	2,683	2,478

The five known laccase genes as well as the eight newly identified laccase-like genes showed no consistent pattern of up- or down-regulation for growth on alder or straw (Fig. 9). Of the six identified putative Class II peroxidases, two were up-regulated on straw (type C and type A) and one was down-regulated on straw (type B). For growth on alder no significant differential expression could be found for the putative peroxidases (Fig. 9). Of the 137 possible cytochrome P450 proteins 33 were up- and 18 down-regulated in straw, and 20 were up- and 24 down-regulated on alder (Fig. 9).

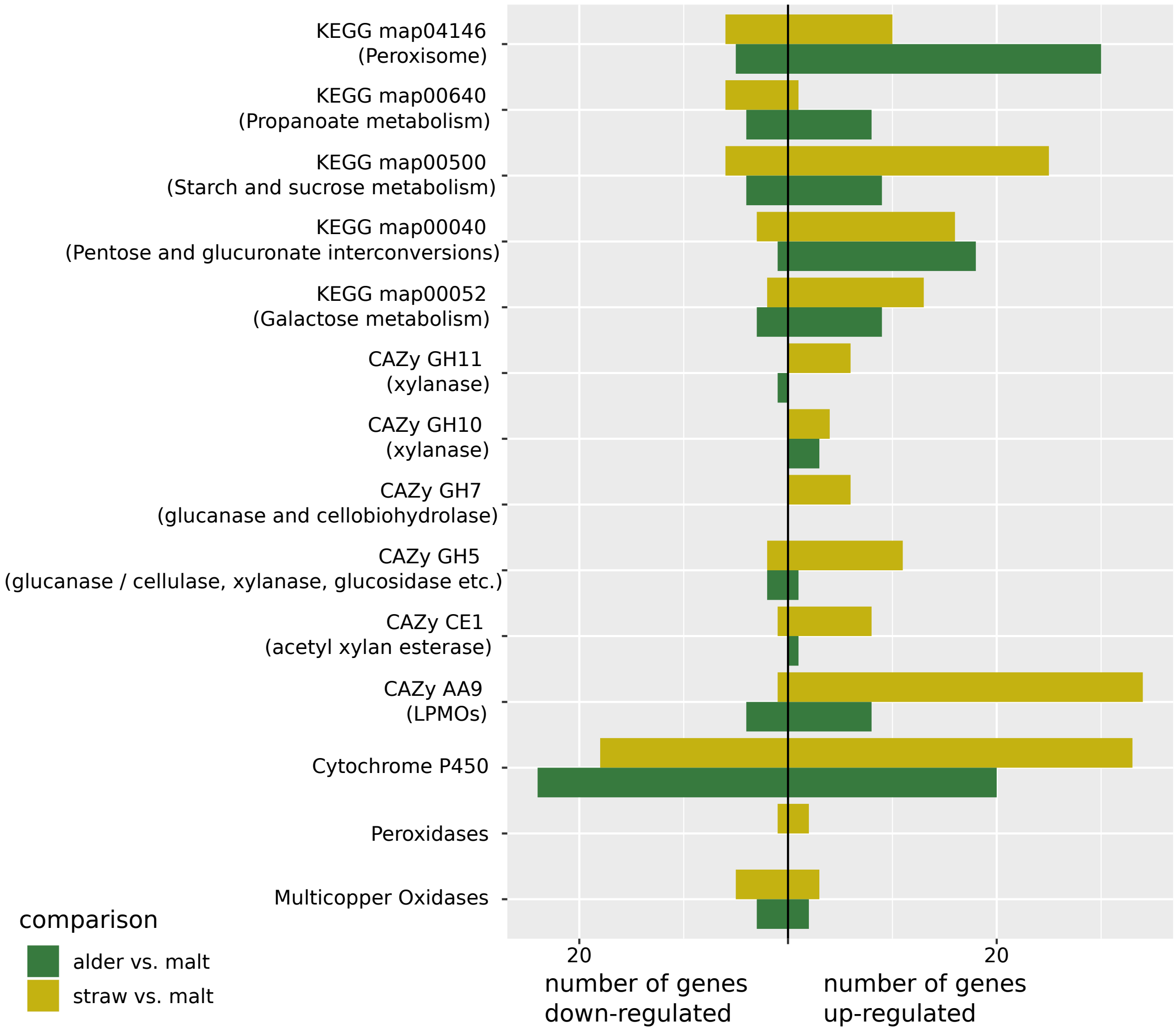


Figure 9: Number of up- and down-regulated genes in different gene groups for the comparison between wheat straw vs. malt extract (yellow) and between alder leaves vs. malt extract (green).

Gene Set Activation

We found multiple activated GO terms, KEGG pathways and CAZy families for all comparisons (Appendix 3, Supplemental Table 1-3). The only exception was the comparison between exponential growth on alder leaves and on malt extract where no active CAZy family was identified. We concentrate here on the differential expression between exponential growth on wheat straw and on malt extract, and between exponential growth on alder leaves and on malt extract, because they are the most relevant when investigating biomass degradation (Fig. 9).

From the six CAZy families that were predicted to be regulated for growth on straw (table 7), three (CE1, GH10 and GH11) were linked to xylane and thus hemicellulose degradation (J. Zhang et al. 2011) and two (GH7 and GH5_5) were linked to glucan and cellulose degradation. The CAZy family predicted to be regulated with the most genes was AA9 which contains lytic polysaccharide monooxygenases (LPMOs) acting among other on cellulose to prepare it for further enzymatic degradation and has been shown to degrade hemicellulose as well (Agger et al. 2014). Investigation of the expression of the genes assigned to these groups in the *C. aquatica* genome showed that for growth on straw they were almost all strongly up-regulated, while for growth on alder in most cases (except for GH10) there was no or only weak up regulation. For each of the families CE1 and GH7 there was one of the assigned genes, that was not up-regulated. This was also the only gene in these families predicted (by signalP) to be not secreted.

The non significant (Fisher's exact test, p=0.0512) overlap between predicted activation of KEGG pathways (table 8) for growth on alder and straw contained the two pathways map00040 (Pentose and glucuronate interconversions) and map00052 (Galactose metabolism). The up-regulation of the Pentose and glucuronate interconversions pathway was mostly caused by the up-regulation of the genes on the path from pectin to glycerol and regulation of some genes involved in conversion of Xylose to Ribulose. The up-regulated enzymes in the Galactose metabolism catalyze conversion of galactose into glucose. The pathway map00500 (Starch and sucrose metabolism) was only predicted to be regulated for growth on straw by the MGSA analysis. Most of the regulated genes are involved in cellulose degradation into glucose, but there is also down regulation of conversion of maltose into glucose. Although this pathway was not predicted to be regulated for growth on alder, many of the gene showed differential expression as well for that comparison. Two interesting pathway predicted to be activated for growth on alder, but not on straw were map04146 (Peroxisome) and map00640 (Propanoate metabolism). In map04146, besides multiple genes that are important for structure and function of the peroxisome, genes involved in the β-oxidation in the peroxisome were up-regulated. In map00640 genes for the degradation of propanoate through the β-oxidation into Acetyl-CoA were up-regulated.

The activated GO terms (table 9) were mostly connected to metabolism, but not specific enough to lead to any further conclusions. GO terms predicted as regulated for the comparison between growth on wheat straw versus growth on malt extract had a significant overlap (Fisher's exact test,

$p<10^{-15}$) with GO terms predicted to be regulated for the comparison between growth on alder leaves versus growth on malt extract.

Table 7: CAZy families predicted to be active by MGSA analysis. For comparison growth on straw compared to growth on malt extract

condition	CAZy Family	activity	genes with this annotation in the genome	differentially expressed genes with this annotation	activation probability
straw-malt	AA9	AA9 (formerly GH61) proteins are copper-dependent lytic polysaccharide monooxygenases (LPMOs); cleavage of cellulose chains with oxidation of various carbons (C-1, C-4 and C-6) has been reported several times in the literature;	49	35	1
straw-malt	CE1	acetyl xylan esterase; cinnamoyl esterase; feruloyl esterase; carboxylesterase; S-formylglutathione hydrolase; diacylglycerol O-acyltransferase; trehalose 6-O-mycolyltransferase	10	9	0.9882
straw-malt	GH11	endo-β-1,4-xylanase; endo-β-1,3-xylanase	6	6	0.96
straw-malt	GH7	endo-β-1,4-glucanase; reducing end-acting cellobiohydrolase; chitosanase; endo-β-1,3-1,4-glucanase	7	6	0.877
straw-malt	GH10	endo-1,4-β-xylanase; endo-1,3-β-xylanase; tomatinase; xylan endotransglycosylase	4	4	0.7582
straw-malt	GH5_5	endo-β-1,4-glucanase / cellulase; endo-β-1,4-xylanase; β-glucosidase; β-mannosidase; β-glucosylceramidase; glucan β-1,3-glucosidase; licheninase; exo-β-1,4-glucanase / cellodextrinase; glucan endo-1,6-β-glucosidase; mannan endo-β-1,4-mannosidase; cellulose β-1,4-cellobiosidase; steryl β-glucosidase; endoglycoceramidase; chitosanase; β-primeverosidase; xyloglucan-specific endo-β-1,4-glucanase; endo-β-1,6-galactanase; hesperidin 6-O-α-L-rhamnosyl-β-glucosidase; β-1,3-mannanase; arabinoxylan-specific endo-β-1,4-xylanase; mannan transglycosylase	5	5	0.6948

Table 8: KEGG pathways predicted to be active by MGSA analysis. For comparison growth on straw compared to growth on malt extract, and growth on alder compared to malt extract

condition	KEGG pathway ID	KEGG pathway name	genes with this annotation in the genome	differentially expressed genes with this annotation	activation probability
alder-malt	ko00040	Pentose and glucuronate interconversions	35	19	1
alder-malt	ko01120	Microbial metabolism in diverse environments	246	92	1
alder-malt	ko04146	Peroxisome	54	35	1
alder-malt	ko00280	Valine, leucine and isoleucine degradation	49	26	1
alder-malt	ko00640	Propanoate metabolism	28	12	0.9992
alder-malt	ko00460	Cyanoamino acid metabolism	26	13	0.9868
alder-malt	ko00906	Carotenoid biosynthesis	4	4	0.9804
alder-malt	ko04978	Mineral absorption	7	4	0.9438
alder-malt	ko00052	Galactose metabolism	30	12	0.7968
alder-malt	ko04920	Adipocytokine signaling pathway	9	4	0.6656
alder-malt	ko04260	Cardiac muscle contraction	12	4	0.6062
straw-malt	ko00500	Starch and sucrose metabolism	65	31	1
straw-malt	ko00040	Pentose and glucuronate interconversions	35	19	1
straw-malt	ko03008	Ribosome biogenesis in eukaryotes	64	23	1
straw-malt	ko00330	Arginine and proline metabolism	38	18	0.9978
straw-malt	ko00520	Amino sugar and nucleotide sugar metabolism	49	17	0.967
straw-malt	ko00980	Metabolism of xenobiotics by cytochrome P450	28	14	0.9512
straw-malt	ko00630	Glyoxylate and dicarboxylate metabolism	40	21	0.9136
straw-malt	ko00920	Sulfur metabolism	17	8	0.9056
straw-malt	ko00052	Galactose metabolism	30	15	0.897
straw-malt	ko00350	Tyrosine metabolism	50	20	0.838
straw-malt	ko00770	Pantothenate and CoA biosynthesis	23	11	0.7994
straw-malt	ko00910	Nitrogen metabolism	21	9	0.6518
straw-malt	ko01220	Degradation of aromatic compounds	32	17	0.6144

Table 9: GO terms predicted to be active by MGSA analysis. For comparison growth on straw compared to growth on malt extract, and growth on alder compared to malt extract

condition	GO ID	GO name	genes with this annotation in the genome	differentially expressed genes with this annotation	activation probability
alder-malt	GO:0016491	oxidoreductase activity	506	205	1
alder-malt	GO:0005975	carbohydrate metabolic process	226	72	1
alder-malt	GO:0071949	FAD binding	55	29	1
alder-malt	GO:0055085	transmembrane transport	513	165	1
alder-malt	GO:0055114	oxidation-reduction process	768	283	1
alder-malt	GO:0008152	metabolic process	387	158	1
alder-malt	GO:0008080	N-acetyltransferase activity	49	19	0.9998
alder-malt	GO:0006508	proteolysis	132	39	0.7976
straw-malt	GO:0008080	N-acetyltransferase activity	49	22	1
straw-malt	GO:0016491	oxidoreductase activity	506	192	1
straw-malt	GO:0005975	carbohydrate metabolic process	226	117	1
straw-malt	GO:0003824	catalytic activity	610	198	1
straw-malt	GO:0016787	hydrolase activity	165	54	1
straw-malt	GO:0071949	FAD binding	55	31	1
straw-malt	GO:0006508	proteolysis	132	53	1
straw-malt	GO:0055085	transmembrane transport	513	188	1
straw-malt	GO:0055114	oxidation-reduction process	768	279	1
straw-malt	GO:0000981	RNA polymerase II transcription factor activity, sequence-specific DNA binding	217	66	0.9982
straw-malt	GO:0042254	ribosome biogenesis	15	9	0.634

5 Discussion

We found that of 10 possible laccases three were up-regulated on wheat straw and two on alder leaves, with showing increased expression on both. This result is in line with findings of an earlier study (Solé et al. 2012) reporting differently regulated laccase genes in *C. aquatica* in response to metals, xenobiotics and lignocellulose breakdown products, as well as the fungal growth stage. The difference between alder and wheat straw could either indicate that different laccases act on these substrates, or that different laccases are involved at different stages of fungal growth and substrate decomposition in case that our samples were not at the same stage (although taken at the same time). We identified six potential peroxidases from the *C. aquatica* genome, that were assigned to the class II of the non animal peroxidase superfamily by Peroxiscan. This class also contains peroxidases known to be involved in lignin degradation like lignin peroxidase (LiP), manganese peroxidase (MnP), and versatile peroxidase (VP) (Hammel and Cullen 2008). Two of the putative peroxidases identified in our study, were up-regulated on straw, but not on alder. To our knowledge, the expression of active peroxidase enzymes has not yet been reported for *C. aquatica*. Their activation on the lignin rich wheat straw could indicate that they are involved in the biotransformation of certain, perhaps phenolic lignin constituents; possibly contributing to detoxification of such compounds. The third group of enzymes that we specifically investigated were cytochrome P450 family. Because the classification of these enzymes could not be further specified it is not clear which of the more than 100 enzymes from this family we found, could be acting on aromatic compounds created by lignin degradation, or on aliphatic compounds from waxes of the cuticula. The observed up-regulation of some of these putative cytochrome P450 monooxygenases on straw (33) and alder (20) could indicate such functions.

We could only detect clear activation of CAZy families for the growth on straw. The activated families all have cellulose and hemicellulose degrading activity as expected on this substrate. The two classical glycoside hydrolase families (GH7 and GH11) that showed the highest activation probabilities have been reported to be induced by growth on straw in other fungi (Ries et al. 2013). Besides two other glycoside hydrolase families (GH10 and GH5) act on cellulose and hemicellulose main chain bonds, we also found up regulation of the CE1 family that contains acetyl xylan esterases, that cleave hemicellulose side chains and of the AA9 family that contains LPMOs that act on cellulose and hemicellulose. LPMOs have been discovered to boost the conversion of lignocellulose via oxidation. The AA9 is a large family and in our case 34 genes that were up-regulated on straw have been assigned to it. Most of the targets and the specific functions of the variety of LPMOs are not yet clarified (Vaaje-Kolstad et al. 2017), but it has been shown that they cleave cellulose as well as hemicellulose components (Frommhagen et al. 2015). Most of the up-regulated proteins in the above mentioned CAZy families are predicted to be secreted. Together this indicates that *C. aquatica* performs extracellular degradation of cellulose and hemicellulose when grown on wheat straw. On alder leaves none of the CAZy families were predicted as active and very few of the genes in them were differentially expressed compared to growth on malt.

In contrast to this difference the overall differential expressed genes on straw and alder showed a significant overlap and regulated KEGG pathways overlapped by two as well. The genes in these two KEGG pathways (map00040 and map00052) were for enzymes involved in xylose and galactose degradation and many of the gene for degradation of cellulose into glucose (map00500) were up regulated on both substrates as well (although many more were up-regulated on straw). For growth on straw this shows a clear process of extracellular cleavage of cellulose and hemicellulose followed by utilisation of the monomers as carbon sources.

It is surprising, that we could only identify the activation of genes for the down stream process, but not for the initial degradation of the polymers when C. aquatica was grown on alder. One possible explanation is the different composition of the two substrates. Wheat straw contains more cellulose (~40%) and lignin (9-22%) (Bjerre et al. 1996; Alemdar and Sain 2008), while the cellulose (5 – 15%) and lignin (6-20%) content in alder leaves are lower (Lecerf and Chauvet 2008; Chauvet 1987). Accordingly it is possible that this leads to a lower expression of cell wall degrading enzymes when C. aquatica is grown on alder leaves. In addition it is possible that alder leaves contain other carbon sources that can be utilised by C. aquatica. The propanoate metabolism (map0640) was predicted to be activated for growth on alder and the up-regulated enzymes were involved in propanoate degradation via the β-oxidation pathway found in other fungi (Otzen et al. 2014). Potentially propanoate could be produced from wax-related fatty acids in the alder leaves (for example from cutin and suberin) not present in wheat straw, or from aliphatic side chains of sterols.

Gene expression of *C. aquatica* on both lignocellulose containing materials showed indication of cellulose and hemicellulose degradation. Especially the enzymes for extracellular depolymerization were more clearly up-regulated on the more cellulose rich wheat straw. Multiple laccases, peroxidases and putative cytochrome P450 monooxygenases were identified in the genome of *C. aquatica*. The expression of several of them was increased on the lignocellulose containing substrates. This observation strongly suggests that *C. aquatica* is able to modify lignin to some extent; perhaps in order to facilitate the utilisation of lignocellulose as a carbon and energy source. It further emphasizes a role of *C. aquatica* in the breakdown of xenobiotic envionmental pollutants when dwelling in its natural riverine habitat.

6 Acknowledgments

We would like to thank Susan Mbedi and Katharina Frindte for performing genome sequencing and Madlen Schubert for carrying out fungal cultivation. We gratefully acknowledge support by the Helmholtz Association of German Research Centres under the research program "Chemicals in the Environment" (CITE) conducted at the Helmholtz Centre for Environmental Research – UFZ. RNA-Sequencing was funded by the Community Science Program (CSP) of the U.S. Department of Energy Joint Genome Institute (DOE JGI). Research was partially funded by the Leibniz Association Pakt/SAW project "MycoLink" (SAW-2014-IGB-1).

7 Author Contribution

F.H., E.C.B., C.W., D.S. and M.T.M. conceived and designed the overall study. D.S. designed and oversaw cultivation. E.C.B. and E.F. planed and carried out RNA extraction. A.L., I.G., V.N. and G.H. performed RNA library preparation, RNA sequencing and initial RNA quality assessment. **F.H.** assembled and annotated the genome, performed RNA-Seq data processing and did statistical analysis. **F.H.**, C.W. D.S. and M.T.M. wrote the manuscript.

8 References

Agger, Jane W., Trine Isaksen, Anikó Várnai, Silvia Vidal-Melgosa, William G. T. Willats, Roland Ludwig, Svein J. Horn, Vincent G. H. Eijsink, and Bjørge Westereng. 2014. "Discovery of LPMO Activity on Hemicelluloses Shows the Importance of Oxidative Processes in Plant Cell Wall Degradation." *Proceedings of the National Academy of Sciences of the United States of America* 111 (17): 6287–92. https://doi.org/10.1073/pnas.1323629111.

Alemdar, Ayse, and Mohini Sain. 2008. "Biocomposites from Wheat Straw Nanofibers: Morphology, Thermal and Mechanical Properties." *Composites Science and Technology* 68 (2): 557–65. https://doi.org/10.1016/j.compscitech.2007.05.044.

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, et al. 2000. "Gene Ontology: Tool for the Unification of Biology." *Nature Genetics* 25 (1): 25–29. https://doi.org/10.1038/75556.

Bauer, Sebastian, Julien Gagneur, and Peter N. Robinson. 2010. "GOing Bayesian: Model-Based Gene Set Analysis of Genome-Scale Data." *Nucleic Acids Research* 38 (11): 3523–32. https://doi.org/10.1093/nar/gkq045.

Bauer, Sebastian, Peter N. Robinson, and Julien Gagneur. 2011. "Model-Based Gene Set Analysis for Bioconductor." *Bioinformatics* 27 (13): 1882–83. https://doi.org/10.1093/bioinformatics/btr296.

Bjerre, Anne Belinda, Anne Bjerring Olesen, Tomas Fernqvist, Annette Plöger, and Anette Skammelsen Schmidt. 1996. "Pretreatment of Wheat Straw Using Combined Wet Oxidation and Alkaline Hydrolysis Resulting in Convertible Cellulose and Hemicellulose." *Biotechnology and Bioengineering* 49 (5): 568–77. https://doi.org/10.1002/(SICI)1097-0290(19960305)49:5<568::AID-BIT10>3.0.CO;2-6.

Brown, C. Titus, Adina Howe, Qingpeng Zhang, Alexis B. Pyrkosz, and Timothy H. Brom. 2012. "A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data." *ArXiv:1203.4802 [q-Bio]*, March. http://arxiv.org/abs/1203.4802.

Chauvet, Eric. 1987. "Changes in the Chemical Composition of Alder, Poplar and Willow Leaves during Decomposition in a River." *Hydrobiologia* 148 (1): 35–44. https://doi.org/10.1007/BF00018164.

Crusoe, Michael R., Hussien F. Alameldin, Sherine Awad, Elmar Boucher, Adam Caldwell, Reed Cartwright, Amanda Charbonneau, et al. 2015. "The Khmer Software Package: Enabling Efficient Nucleotide Sequence Analysis." *F1000Research*, September. https://doi.org/10.12688/f1000research.6924.1.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Duarte, S., F. Bärlocher, J. Trabulo, F. Cássio, and C. Pascoal. 2015. "Stream-Dwelling Fungal Decomposer Communities along a Gradient of Eutrophication Unraveled by 454 Pyrosequencing." *Fungal Diversity* 70 (1): 127–48. https://doi.org/10.1007/s13225-014-0300-y.

Fawal, Nizar, Qiang Li, Bruno Savelli, Marie Brette, Gisele Passaia, Maxime Fabre, Catherine Mathé, and Christophe Dunand. 2013. “PeroxiBase: A Database for Large-Scale Evolutionary Analysis of Peroxidases.” *Nucleic Acids Research* 41 (Database issue): D441-444. https://doi.org/10.1093/nar/gks1083.

Finn, Robert D., Teresa K. Attwood, Patricia C. Babbitt, Alex Bateman, Peer Bork, Alan J. Bridge, Hsin-Yu Chang, et al. 2017. “InterPro in 2017—beyond Protein Family and Domain Annotations.” *Nucleic Acids Research* 45 (D1): D190–99. https://doi.org/10.1093/nar/gkw1107.

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

Frommhagen, Matthias, Sumanth Kumar Mutte, Adrie H. Westphal, Martijn J. Koetsier, Sandra W. A. Hinz, Jaap Visser, Jean-Paul Vincken, et al. 2017. “Boosting LPMO-Driven Lignocellulose Degradation by Polyphenol Oxidase-Activated Lignin Building Blocks.” *Biotechnology for Biofuels* 10 (May): 121. https://doi.org/10.1186/s13068-017-0810-4.

Frommhagen, Matthias, Stefano Sforza, Adrie H. Westphal, Jaap Visser, Sandra W. A. Hinz, Martijn J. Koetsier, Willem J. H. van Berkel, Harry Gruppen, and Mirjam A. Kabel. 2015. “Discovery of the Combined Oxidative Cleavage of Plant Xylan and Cellulose by a New Fungal Polysaccharide Monooxygenase.” *Biotechnology for Biofuels* 8 (July): 101. https://doi.org/10.1186/s13068-015-0284-1.

Gessner, M. O., V. Gulis, K.A. Kuehn, E. Chauvet, and K. Suberkopp. 2007. “Fungal Decomposers of Plant Litter in Aquatic Ecosystems.” In *Environmental and Microbial Relationships*, 4:301–24. The Mycota.

Grabherr, Manfred G., Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, et al. 2011. “Trinity: Reconstructing a Full-Length Transcriptome without a Genome from RNA-Seq Data.” *Nature Biotechnology* 29 (7): 644–52. https://doi.org/10.1038/nbt.1883.

Haas, Brian J., Arthur L. Delcher, Stephen M. Mount, Jennifer R. Wortman, Roger K. Smith, Linda I. Hannick, Rama Maiti, et al. 2003. “Improving the Arabidopsis Genome Annotation Using Maximal Transcript Alignment Assemblies.” *Nucleic Acids Research* 31 (19): 5654–66. https://doi.org/10.1093/nar/gkg770.

Hammel, Kenneth E, and Dan Cullen. 2008. “Role of Fungal Peroxidases in Biological Ligninolysis.” *Current Opinion in Plant Biology*, Physiology and Metabolism - Edited by Markus Pauly and Kenneth Keegstra, 11 (3): 349–55. https://doi.org/10.1016/j.pbi.2008.02.003.

Iqbal, S. H., and J. Webster. 1973. “Aquatic Hyphomycete Spora of the River Exe and Its Tributaries.” *Transactions of the British Mycological Society* 61 (2): 331–46. https://doi.org/10.1016/S0007-1536(73)80155-X.

Johnson, Marc T. J., Eric J. Carpenter, Zhijian Tian, Richard Bruskiewich, Jason N. Burris, Charlotte T. Carrigan, Mark W. Chase, et al. 2012. “Evaluating Methods for Isolating Total RNA and Predicting the Success of Sequencing Phylogenetically Diverse Plant Transcriptomes." *PLOS ONE* 7 (11): e50226. https://doi.org/10.1371/journal.pone.0050226.

Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, et al. 2014. "InterProScan 5: Genome-Scale Protein Function Classification." *Bioinformatics* 30 (9): 1236–40. https://doi.org/10.1093/bioinformatics/btu031.

Junghanns, Charles, Monika Moeder, Gudrun Krauss, Claudia Martin, and Dietmar Schlosser. 2005. "Degradation of the Xenoestrogen Nonylphenol by Aquatic Fungi and Their Laccases." *Microbiology* 151 (1): 45–57. https://doi.org/10.1099/mic.0.27431-0.

Kanehisa, Minoru, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. 2016. "KEGG as a Reference Resource for Gene and Protein Annotation." *Nucleic Acids Research* 44 (Database issue): D457–62. https://doi.org/10.1093/nar/gkv1070.

Kanehisa, Minoru, Yoko Sato, and Kanae Morishima. 2016. "BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences." *Journal of Molecular Biology* 428 (4): 726–31. https://doi.org/10.1016/j.jmb.2015.11.006.

Köster, Johannes, and Sven Rahmann. 2012. "Snakemake—a Scalable Bioinformatics Workflow Engine." *Bioinformatics* 28 (19): 2520–22. https://doi.org/10.1093/bioinformatics/bts480.

Krauss, Gerd-Joachim, Magali Solé, Gudrun Krauss, Dietmar Schlosser, Dirk Wesenberg, and Felix Bärlocher. 2011. "Fungi in Freshwaters: Ecology, Physiology and Biochemical Potential." *FEMS Microbiology Reviews* 35 (4): 620–51. https://doi.org/10.1111/j.1574-6976.2011.00266.x.

Kubicek, Christian P., and Irina S. Druzhinina, eds. 2007. "Fungal Decomposers of Plant Litter in Aquatic Ecosystems." In *Environmental and Microbial Relationships*, 301–24. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-71840-6_17.

Kumar, S. V. Suresh, Prashant S. Phale, S. Durani, and Pramod P. Wangikar. 2003. "Combined Sequence and Structure Analysis of the Fungal Laccase Family." *Biotechnology and Bioengineering* 83 (4): 386–94. https://doi.org/10.1002/bit.10681.

Lecerf, Antoine, and Eric Chauvet. 2008. "Intraspecific Variability in Leaf Traits Strongly Affects Alder Leaf Decomposition in a Stream." *Basic and Applied Ecology* 9 (5): 598–605. https://doi.org/10.1016/j.baae.2007.11.003.

Li, Bo, and Colin N. Dewey. 2011. "RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome." *BMC Bioinformatics* 12 (August): 323. https://doi.org/10.1186/1471-2105-12-323.

Lombard, Vincent, Hemalatha Golaconda Ramulu, Elodie Drula, Pedro M. Coutinho, and Bernard Henrissat. 2014. "The Carbohydrate-Active Enzymes Database (CAZy) in 2013." *Nucleic Acids Research* 42 (Database issue): D490-495. https://doi.org/10.1093/nar/gkt1178.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (December): 550. https://doi.org/10.1186/s13059-014-0550-8.

Lundell, Taina K., Miia R. Mäkelä, and Kristiina Hildén. 2010. “Lignin-Modifying Enzymes in Filamentous Basidiomycetes - Ecological, Functional and Phylogenetic Review.” *Journal of Basic Microbiology* 50 (1): 5–20. https://doi.org/10.1002/jobm.200900338.

Martin, Claudia, Monika Moeder, Xavier Daniel, Gudrun Krauss, and Dietmar Schlosser. 2007. “Biotransformation of the Polycyclic Musks HHCB and AHTN and Metabolite Formation by Fungi Occurring in Freshwater Environments.” *Environmental Science & Technology* 41 (15): 5395–5402. https://doi.org/10.1021/es0711462.

Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.Journal* 17 (1): 10–12.

Otzen, Christian, Bettina Bardl, Ilse D. Jacobsen, Markus Nett, and Matthias Brock. 2014. “Candida Albicans Utilizes a Modified β-Oxidation Pathway for the Degradation of Toxic Propionyl-CoA.” *The Journal of Biological Chemistry* 289 (12): 8151–69. https://doi.org/10.1074/jbc.M113.517672.

Palmer, Jon. 2018. *Funannotate: Eukaryotic Genome Annotation Pipeline*. Python. https://github.com/nextgenusfs/funannotate.

Petersen, Thomas Nordahl, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2011. “SignalP 4.0: Discriminating Signal Peptides from Transmembrane Regions.” *Nature Methods* 8 (10): 785–86. https://doi.org/10.1038/nmeth.1701.

Ries, Laure, Steven T. Pullan, Stéphane Delmas, Sunir Malla, Martin J. Blythe, and David B. Archer. 2013. “Genome-Wide Transcriptional Response of Trichoderma Reesei to Lignocellulose Using RNA Sequencing and Comparison with Aspergillus Niger.” *BMC Genomics* 14 (August): 541. https://doi.org/10.1186/1471-2164-14-541.

Riley, Robert, Asaf A. Salamov, Daren W. Brown, Laszlo G. Nagy, Dimitrios Floudas, Benjamin W. Held, Anthony Levasseur, et al. 2014. “Extensive Sampling of Basidiomycete Genomes Demonstrates Inadequacy of the White-Rot/Brown-Rot Paradigm for Wood Decay Fungi.” *Proceedings of the National Academy of Sciences* 111 (27): 9923–28. https://doi.org/10.1073/pnas.1400592111.

Schlosser, Dietmar, and Christine Höfer. 2002. “Laccase-Catalyzed Oxidation of Mn2+ in the Presence of Natural Mn3+ Chelators as a Novel Source of Extracellular H2O2 Production and Its Impact on Manganese Peroxidase.” *Applied and Environmental Microbiology* 68 (7): 3514–21. https://doi.org/10.1128/AEM.68.7.3514-3521.2002.

Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. “BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs.” *Bioinformatics* 31 (19): 3210–12. https://doi.org/10.1093/bioinformatics/btv351.

Sirim, Demet, Florian Wagner, Lei Wang, Rolf D Schmid, and Jürgen Pleiss. 2011. “The Laccase Engineering Database: A Classification and Analysis System for Laccases and Related Multicopper Oxidases.” *Database: The Journal of Biological Databases and Curation* 2011 (April). https://doi.org/10.1093/database/bar006.

Solé, Magali, Ines Müller, Marek J. Pecyna, Ingo Fetzer, Hauke Harms, and Dietmar Schlosser. 2012. “Differential Regulation by Organic Compounds and Heavy Metals of Multiple Laccase Genes in the Aquatic Hyphomycete Clavariopsis Aquatica.” *Applied and Environmental Microbiology* 78 (13): 4732–39. https://doi.org/10.1128/AEM.00635-12.

Suberkropp, K., and M. J. Klug. 1976. “Fungi and Bacteria Associated with Leaves During Processing in a Woodland Stream.” *Ecology* 57 (4): 707–19. https://doi.org/10.2307/1936184.

Tang, Juliet D., Leslie A. Parker, Andy D. Perkins, Tad S. Sonstegard, Steven G. Schroeder, Darrel D. Nicholas, and Susan V. Diehl. 2013. “Gene Expression Analysis of Copper Tolerance and Wood Decay in the Brown Rot Fungus Fibroporia Radiculosa.” *Applied and Environmental Microbiology* 79 (5): 1523–33. https://doi.org/10.1128/AEM.02916-12.

The Gene Ontology Consortium. 2017. “Expansion of the Gene Ontology Knowledgebase and Resources.” *Nucleic Acids Research* 45 (Database issue): D331–38. https://doi.org/10.1093/nar/gkw1108.

Vaaje-Kolstad, Gustav, Zarah Forsberg, Jennifer SM Loose, Bastien Bissaro, and Vincent GH Eijsink. 2017. “Structural Diversity of Lytic Polysaccharide Monooxygenases.” *Current Opinion in Structural Biology*, Carbohydrates: A feast of structural glycobiology • Sequences and topology: Computational studies of protein-protein interactions, 44 (June): 67–76. https://doi.org/10.1016/j.sbi.2016.12.012.

Voříšková, Jana, and Petr Baldrian. 2013. “Fungal Community on Decomposing Leaf Litter Undergoes Rapid Successional Changes.” *The ISME Journal* 7 (3): 477–86. https://doi.org/10.1038/ismej.2012.116.

Yang, Yang, Fangfang Fan, Rui Zhuo, Fuying Ma, Yangmin Gong, Xia Wan, Mulan Jiang, and Xiaoyu Zhang. 2012. “Expression of the Laccase Gene from a White Rot Fungus in Pichia Pastoris Can Enhance the Resistance of This Yeast to H2O2-Mediated Oxidative Stress by Stimulating the Glutathione-Based Antioxidative System.” *Applied and Environmental Microbiology* 78 (16): 5845–54. https://doi.org/10.1128/AEM.00218-12.

Yin, Yanbin, Xizeng Mao, Jincai Yang, Xin Chen, Fenglou Mao, and Ying Xu. 2012. “DbCAN: A Web Resource for Automated Carbohydrate-Active Enzyme Annotation.” *Nucleic Acids Research* 40 (Web Server issue): W445. https://doi.org/10.1093/nar/gks479.

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

Zhang, Junhua, Matti Siika-aho, Maija Tenkanen, and Liisa Viikari. 2011. “The Role of Acetyl Xylan Esterase in the Solubilization of Xylan and Enzymatic Hydrolysis of Wheat Straw and Giant Reed.” *Biotechnology for Biofuels* 4 (December): 60. https://doi.org/10.1186/1754-6834-4-60.

Zhang, Qingpeng, Jason Pell, Rosangela Canino-Koning, Adina Chuang Howe, and C. Titus Brown. 2014. “These Are Not the K-Mers You Are Looking For: Efficient Online K-Mer Counting Using a Probabilistic Data Structure.” *PLOS ONE* 9 (7): e101271. https://doi.org/10.1371/journal.pone.0101271.

VI Discussion

The two aspects of this thesis are diversity and function of aquatic fungi. The aim was to apply genomics approaches to these aspects and improve upon certain points of the methods where necessary.

To study species diversity directly from the habitat metabarcoding can be used. Metabarcoding is already widely used to study terrestrial fungi and has also been applied to study aquatic fungi. In a metabarcoding project a barcode region has to be chosen a prior based on different factors. The two properties mostly considered are the ability to amplify the barcode from all species that should be studied and the difference between intra-species variability (that should be low) and inter-species variability (that should be high). These were also the criteria to choose the ITS region as the standard marker for fungi (Schoch et al., 2012). In general the rRNA markers showed the best amplification success and of them the ITS region had the best capability to separate reads from different species into OTUs. For ecologically meaningful analysis the OTUs need to be assigned to taxonomic groups (ideally species). This is difficult for fungi, because only small proportion of existing fungi is described and even less are represented in reference database. Aquatic fungi are especially poorly studied and therefore database coverage is especially low for them.

The first aspect that the thesis tackled was to quantify the effect of incomplete databases for the ITS region. In chapter III the problem of incomplete databases is demonstrated by the *in silico* analysis of the sequences of the UNITE database. Missing reference sequences from the same species, genus or family not only led to the inability to classify the sequence to that level, but also caused problems in higher level classification. In an understudied group like aquatic fungi this leads to a high proportion of sequences not even being classified to the phylum rank and fungal sequence not being identified as such as can be seen in the the test on fresh water lakes in chapter III.

This means that the choice of barcode comes with an additional crucial trade-off when many novel species are expected. A very variable marker like the ITS1 or ITS2 has the advantage of being able to identify sequences to the species level, if a sequence from the same species is in the database, but might fail to classify novel species to any meaningful level, when that is not the case. A less variable region like the 5.8S, the SSU or the LSU can give better classifications for novel species, but will not be able to classify sequences to the species level. A possible solution, that can be implemented with Illumina sequencing is to use a part of the 5.8S as complementary marker to the ITS2. This was implemented and tested in chapter III.

When the ITS region was suggested as standard barcode for fungi by Schoch et al., this was done on the basis of using the whole ITS region. Because of the read-length constraints of Illumina sequencing most studies now focus on only the ITS1 or only the ITS2. With third generation sequencing this restriction no longer applies. On the one hand full ITS region sequencing at a reasonable cost and sufficiently high throughput is possible. On the other hand we can go one step further and include the SSU, the LSU or both into an amplicon. In chapter two we used an amplicon of the full eukaryotic rRNA operon for barcoding of fungal fresh water communities. This has the advantage that information from all three regions and the according databases can be used, but also comes with the problem how to integrate this information. Because all three parts of the rRNA operon have been used as independent markers, there are different databases for them. An approach using the full operon can benefit from the independent confirmation fro three datanase, but also needs to deal with the differences. Considering the frequent changes in the taxonomy of fungi it is unsurprising, that the taxonomies underlying the three databases used in chapter IV, are different. SILVA and RDP both use trees that have been computed from the sequences in the database and in the case of SILVA manually curated (Munoz et al., 2011). UNITE until recently used the tree underlying Index Fugorum (www.indexfungorum.org), but has now also started to use classifications from a big phylogenetic study by Tedersoo et al. (2017a). Differences between the taxonomies are substantial. For example at any taxonomic rank more than 60% of taxa are unique between RDP and SILVA (Balvočiūtė and Huson, 2017). Projects like the Open tree taxonomy (Rees and Cranston, 2017) are making efforts towards unifying the taxonomies and linking corresponding taxa from different databases to each other. This would make direct comparisons of classifications from the different databases possible. Classifications that are confirmed by more than one of the rRNA markers would higher confidence.

In chapter two we used a primer pair to amplify almost the full rRNA operon. This amplicon has a length of around 4.5kb and is much longer than what is normally used for barcoding. We were concerned that a long amplicon would form more PCR chimeras. We did not find any evidence that given a reasonable number of PCR cyles the chimera rate is higher than in short read studies, but chimera formation during PCR is influenced by so many factors, that it would need a far more comprehensive study to investigate the influence that amplicon length has on chimera formation rates. Another drawback of the very long amplicon would be a possible amplification bias for shorter sequences (Shagin et al., 1999), this is another factor that would need to be studied in more detail.

Lastly, the long amplicon length also poses a challenge for sequencing. For the CCS to significantly reduce the error rate of PacBio reads, multiple “passes” are needed (Travers et al., 2010). This means the polymerase enzyme has to pass around the circular single-stranded DNA molecule multiple times. With an amplicon length of 4.5kb this means that for example to get three passes a raw reads length of 13.5 kb is necessary even when ignoring the hairpin adapters. In our study we could solve this problem with very stringent quality filtering, but at the cost of removing many reads that did not have enough passes and as a result a too high error rate.

Considering all these problems of the very long amplicon used in chapter two, it might be advisable to use a shorter amplicon for metabarcoding. One possibility would be to use only the ITS region and the SSU reducing the length of the amplicon almost by half. This would combine the most conserved with the most variable region and thus hopefully give the full advantage of both similar to the approach in chapter III. The drawback would be that fungal groups that have so far been identified mainly by the LSU could be less we classified. Another option would be to use primers that amplify less of the SSU and LSU or to use the ITS and the LSU (Schlaeppi et al., 2016; Tedersoo et al., 2017b). Third generation sequencing has opened up different possibilities and our results in chapter two show it has great potential to improve metabarcoding for fungi and other eukaryotes.

Besides PacBio the other the big vendor of third generation sequencing technology is Oxford Nanopore. Nanopore reads come with a very similar base error rate (~13%) as PacBio and do not offer the possibility to do CCS. This would make them very hard to use for metabarcoding, because the error rate is higher than the typical OTU clustering thresholds of 3%. Besides for metabarcoding we also used PacBio to barcode isolate samples in chapter IV. It has already been suggested that PacBio CSS could replace Sanger sequencing for single isolate barcoding (Hebert et al., 2018). For a shorter barcode (~800 bp), like the ones sequenced with Sanger up to now, more CCS passes than we used in chapter IV can be expected reducing error rates below the ones of Sanger sequencing. For this application Nanopore error rates are also not prohibitive, because a consensus can be generated from all reads from one sample if only one species is present. With the MinIon Oxford Nanopore also provides a cheaper sequencing option than PacBio that could be used for barcoding in the future (Srivathsan et al., 2018; Wurzbacher et al., 2018).

Because of the high number of novel species that are regularly encountered in metabarcoding studies and the fact that many of them can not easily be cultured, the number of formally described species has not kept up with the number of species known only from their barcode sequences. The official naming of a species requires a type specimen which is hard to obtain for environmental species that have not been cultured. It has been suggested, that it should be possible to name a fungal species based on a barcode sequence as type material (Hawksworth et al., 2016) to be able to attract more attention and research interest to new species (Ryberg and Nilsson, 2018). Longer barcode sequences that can be sequenced with third generation sequencing from pure cultures or the environment with high accuracy could make it easier to make the case that a barcode sequence is sufficient as type material.

Overall third generation sequencing holds a lot of promise for barcoding and metabarcoding in fungi. It opens up the possibility to sequence the whole rRNA operon in a metbarcoding study, which will make it possible to use information from the databases for SSU, ITS and LSU, that have so far been developed completely independently. It could even offer synergies where a sequence that can be assigned by one database, but not by the other can be added to the database that it is missing from with the taxonomic information from the one where it was found.
The possibility to sequence longer amplicons also could make it possible to find new markers, that were so far not feasible to sequence, because the were to long for Illumina.

The second topic of the thesis was besides the diversity of aquatic fungi was their function in the ecosystem. Specifically their role as degraders of plant biomass. Genome sequencing and RNA-Seq have been extensively used on terrestrial fungi, but to our knowledge the study presented in chapter V is the first time these methods are applied in combination to an exclusively aquatic fungus. Aquatic fungi that degrade lignocellulose could have specific adaptations to the aquatic lifestyle, in which extra cellular enzyms have to act fast before they are diluted. This possibility together with the fact that they can be easily cultured in liquid form makes them interesting for industrial applications like the creation of biofuels from plant waste materials. Out study design in chapter V was not appropriate to get any insights in enzyme kinetics, but gives a first overview of proteins present, and acting on lignocellulose in an exclusively aquatic Ascomycote. This presents a starting point to further investigate the enzymatic capability of aquatic fungi and compare them to terrestrial counter parts. Whole genome comparisons of terrestrial and aquatic fungi could give insights in adaptations to the aquatic lifestyle and answer the question if there is any evolutionary separation between groups of aquatic and terrestrial fungi.

The so called “reproducibility crisis” started in 2011 with an article that showed that many wide spread practices in psychological science can lead to significant results in the absence of a real effect (Simmons et al., 2011). It soon “spread” to other disciplines with doubt about the adherence to good scientific practices in medical science being raised (Begley and Ioannidis, 2015) and a study showing that many big studies could not be reproduced (Prinz et al., 2011). For genomics the focus of reproducibility has so far been on data availability (Drew et al., 2013; Ioannidis and Khoury, 2011). This is obviously a good point, because access to the raw data is a necessary condition to be able to reproduce any computational analysis. On the other hand the data alone are not sufficient. Especially in analysis that consist of many steps with different software tools, the way from raw data to the final result can be very complicated. Unfortunately in many articles the computational analysis is considered less relevant than sampling or laboratory procedures. Often not even the minimum requirement of giving software versions and parameters is met (Nekrutenko and Taylor, 2012). Like laboratory procedures computational analyses can consist of many often minute steps. In theory analyses on a computer have the possibility to be perfectly reproducible. Every step that is taken can be written down as a unambigous command that was given to the computer and that is expected to give the same result if repeated. In practice the conditions under which the command is given make a big difference. The software versions have to be exactly the same and every step that was taken to create the input data has to be exactly reproduced. Even if the software versions are known getting the exact version of a software to run on the system one has available is not always easy. One possible solutions is to generate complete self contained computing environments with a virtual machine (e.g. VirtualBox, www.virtualbox.org) or a container format like docker (www.docker.com). These are tools that have been developed for computer system administration, when faced with the similar problem to create a exactly defined environment for a software to run in. Another more light wait method is the utilization of work flow engines like Galaxy (Goecks et al., 2010) or snakemake (Köster and Rahmann, 2012).

For the analyses in this thesis I used snakemake work flows, that are deposited on github. Especially for chapter IV I followed an approach where the figures that are presented should be automatically created form the raw data. This is the case for figures 2, 4 and 5 from that chapter, which can be produced by downloading the workflow files from github and running one command. This makes it convenient to reproduce the figures from the data and more importantly guaranties, that every step in my analysis is documented in the from of program code. For figure 3 this was not possible since the last step of creating the figure was done manually in Cytoscape. The problem of acquiring and installing the correct software versions is still given, but since the code that does the analysis is available any step can be investigated down to the lowest level.

References

Balvočiūtė, M., and Huson, D.H. (2017). SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? BMC Genomics *18*.

Begley, C.G., and Ioannidis, J.P.A. (2015). Reproducibility in Science: Improving the Standard for Basic and Preclinical Research. Circulation Research *116*, 116–126.

Drew, B.T., Gazis, R., Cabezas, P., Swithers, K.S., Deng, J., Rodriguez, R., Katz, L.A., Crandall, K.A., Hibbett, D.S., and Soltis, D.E. (2013). Lost Branches on the Tree of Life. PLOS Biology *11*, e1001636.

Goecks, J., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology *11*, R86.

Hawksworth, D.L., Hibbett, D.S., Kirk, P.M., and Lücking, R. (2016). (308–310) Proposals to permit DNA sequence data to serve as types of names of fungi. Taxon *65*, 899–900.

Hebert, P.D.N., Braukmann, T.W.A., Prosser, S.W.J., Ratnasingham, S., deWaard, J.R., Ivanova, N.V., Janzen, D.H., Hallwachs, W., Naik, S., Sones, J.E., et al. (2018). A Sequel to Sanger: amplicon sequencing that scales. BMC Genomics *19*, 219.

Ioannidis, J.P.A., and Khoury, M.J. (2011). Improving Validation Practices in “Omics” Research. Science *334*, 1230–1232.

Köster, J., and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. Bioinformatics *28*, 2520–2522.

Munoz, R., Yarza, P., Ludwig, W., Euzéby, J., Amann, R., Schleifer, K.-H., Oliver Glöckner, F., and Rosselló-Móra, R. (2011). Release LTPs104 of the All-Species Living Tree. Systematic and Applied Microbiology *34*, 169–170.

Nekrutenko, A., and Taylor, J. (2012). Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. Nature Reviews Genetics *13*, 667–672.

Prinz, F., Schlange, T., and Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery *10*, 712.

Rees, J., and Cranston, K. (2017). Automated assembly of a reference taxonomy for phylogenetic data synthesis. Biodiversity Data Journal *5*, e12581.

Ryberg, M., and Nilsson, R.H. (2018). New light on names and naming of dark taxa. MycoKeys 31–39.

Schlaeppi, K., Bender, S.F., Mascher, F., Russo, G., Patrignani, A., Camenzind, T., Hempel, S., Rillig, M.C., and van der Heijden, M.G.A. (2016). High-resolution community profiling of arbuscular mycorrhizal fungi. New Phytol. *212*, 780–791.

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., Chen, W., and Consortium, F.B. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. PNAS *109*, 6241–6246.

Shagin, D.A., Lukyanov, K.A., Vagner, L.L., and Matz, M.V. (1999). Regulation of average length of complex PCR product. Nucleic Acids Res *27*, e23-i-e23-iii.

Simmons, J.P., Nelson, L.D., and Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychol Sci *22*, 1359–1366.

Srivathsan, A., Baloğlu, B., Wang, W., Tan, W.X., Bertrand, D., Ng, A.H.Q., Boey, E.J.H., Koh, J.J.Y., Nagarajan, N., and Meier, R. (2018). A MinION$^{\mathrm{TM}}$-based pipeline for fast and cost-effective DNA barcoding. Molecular Ecology Resources.

Tedersoo, L., Bahram, M., Puusepp, R., Nilsson, R.H., and James, T.Y. (2017a). Novel soil-inhabiting clades fill gaps in the fungal tree of life. Microbiome *5*, 42.

Tedersoo, L., Ave, T.-K., and Anslan Sten (2017b). PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. New Phytologist *217*, 1370–1385.

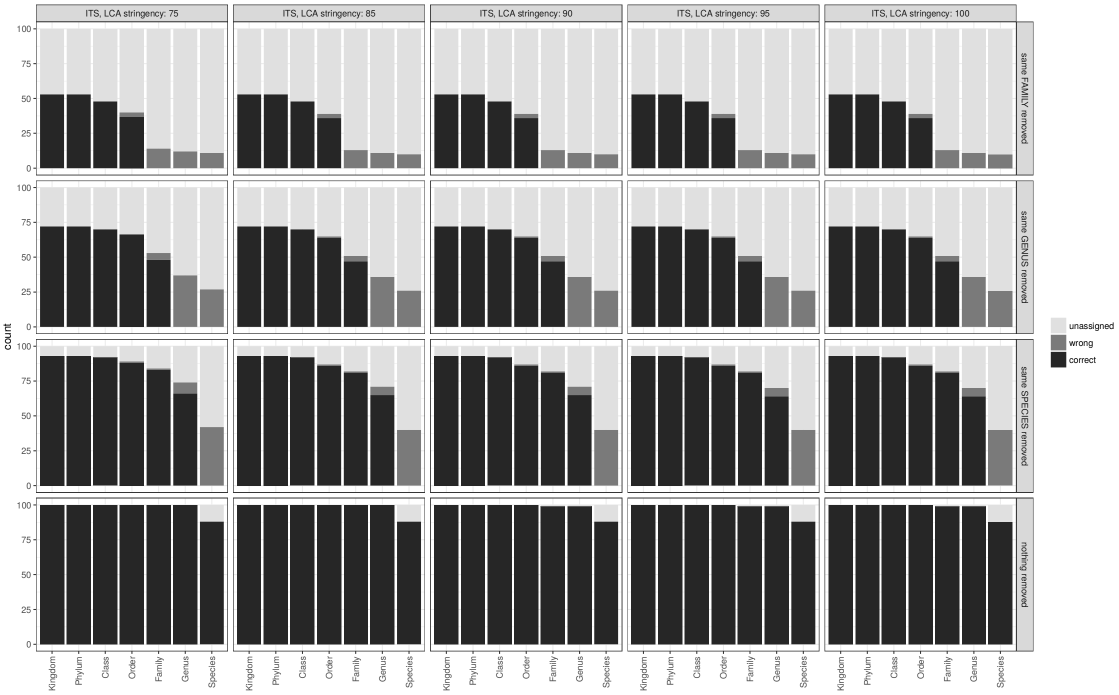
Travers, K.J., Chin, C.-S., Rank, D.R., Eid, J.S., and Turner, S.W. (2010). A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucleic Acids Res *38*, e159.

Wurzbacher, C., Larsson, E., Bengtsson-Palme, J., Van den Wyngaert, S., Svantesson, S., Kristiansson, E., Kagami, M., and Nilsson, R.H. (2018). Introducing ribosomal tandem repeat barcoding for fungi.

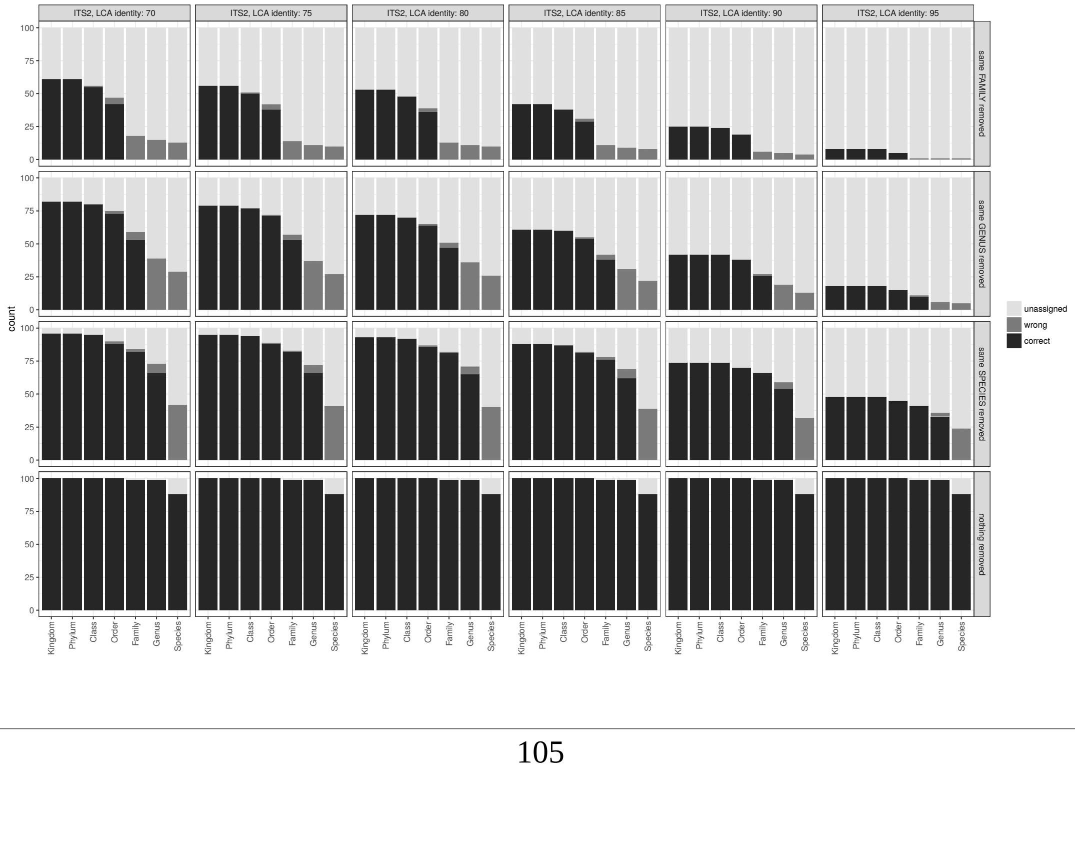
Oracle VM VirtualBox.

VII Appendix 1

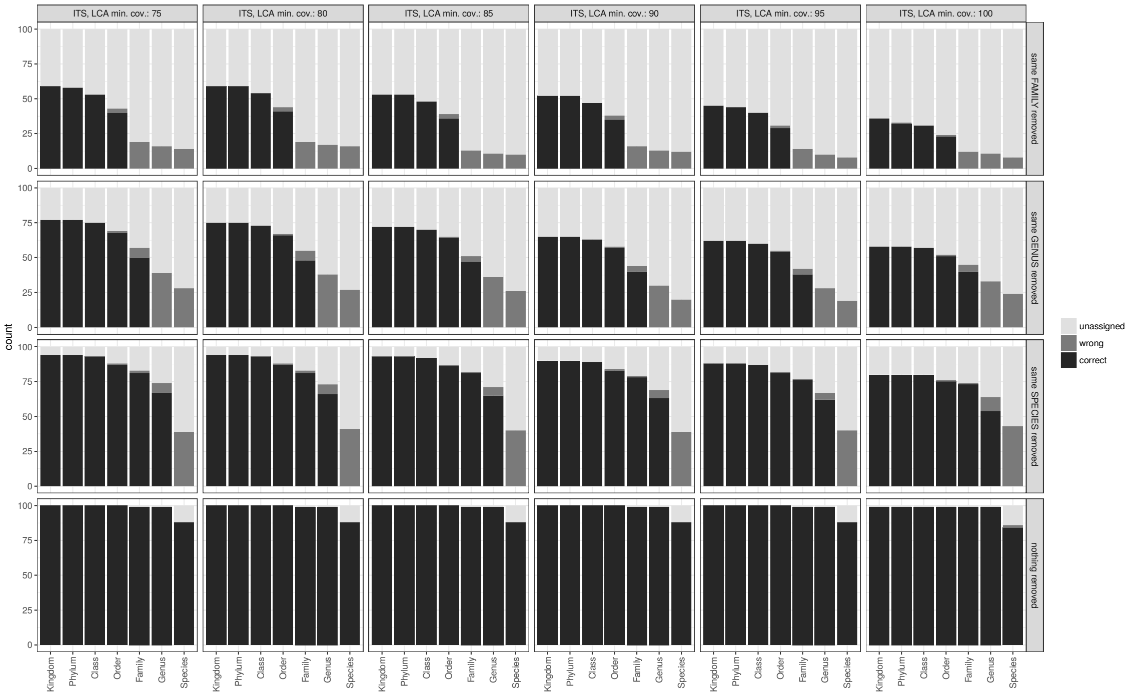
1 Supplemental Figure 1



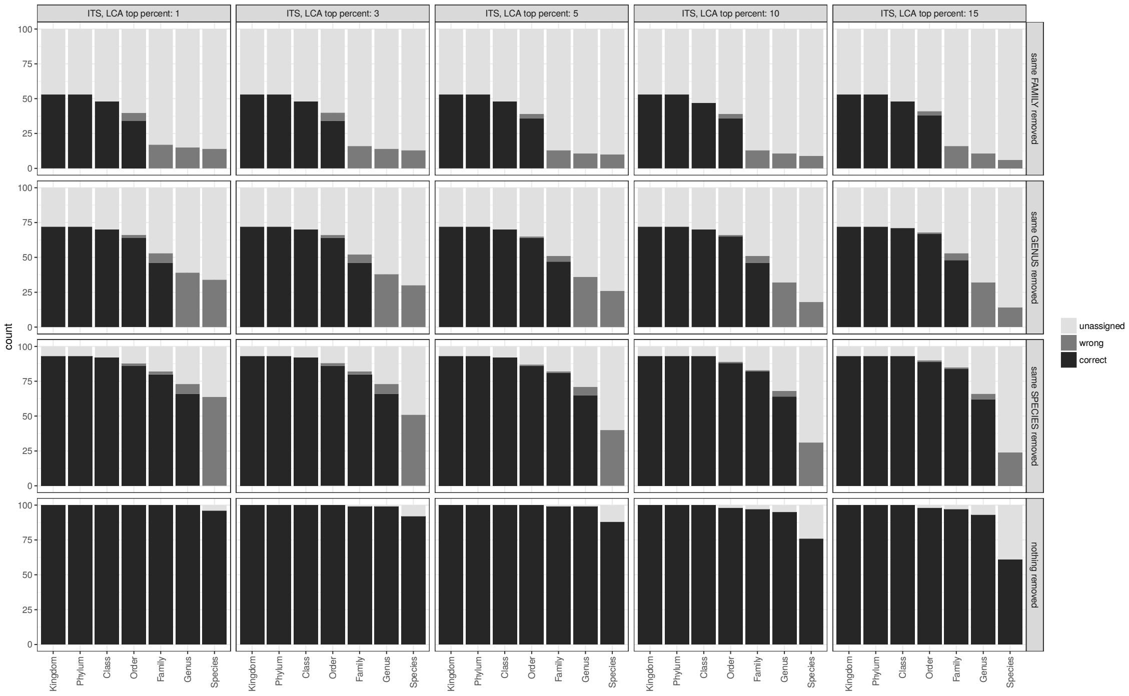
2 Supplemental Figure 2



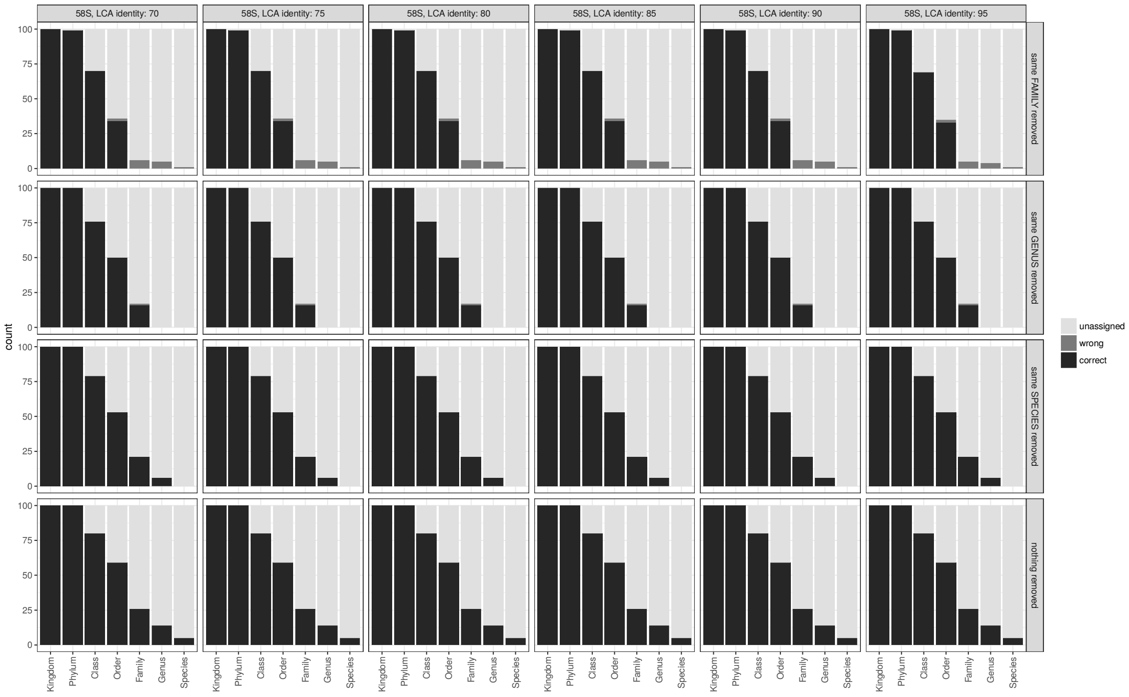
3 Supplemental Figure 3



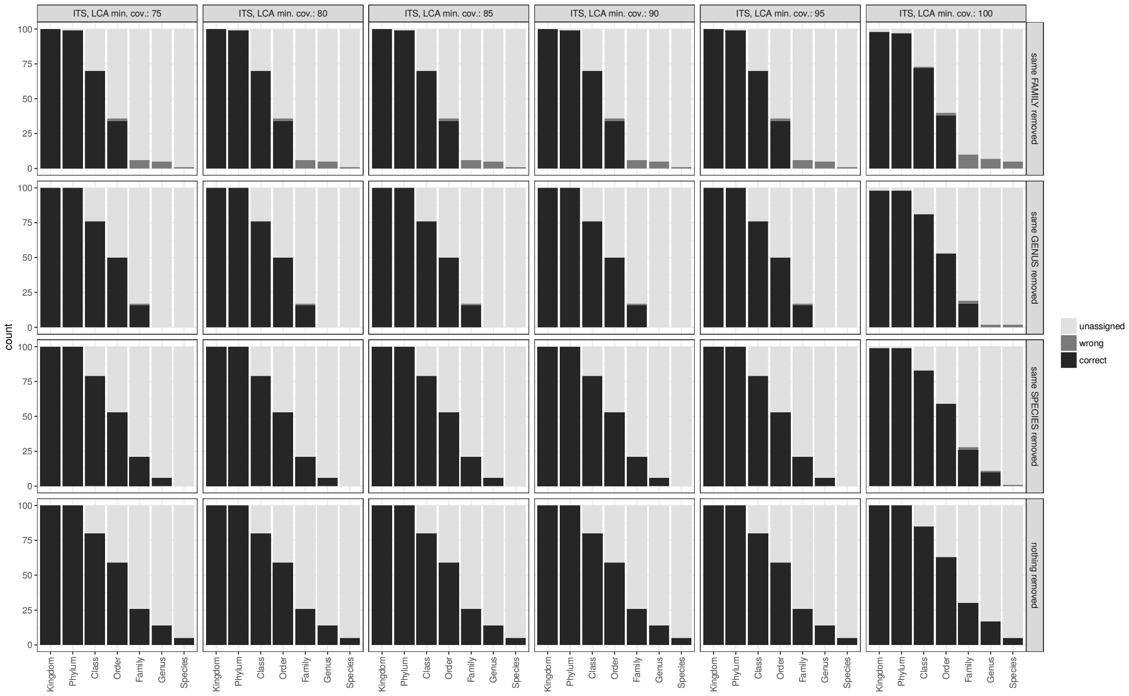
4 Supplemental Figure 4



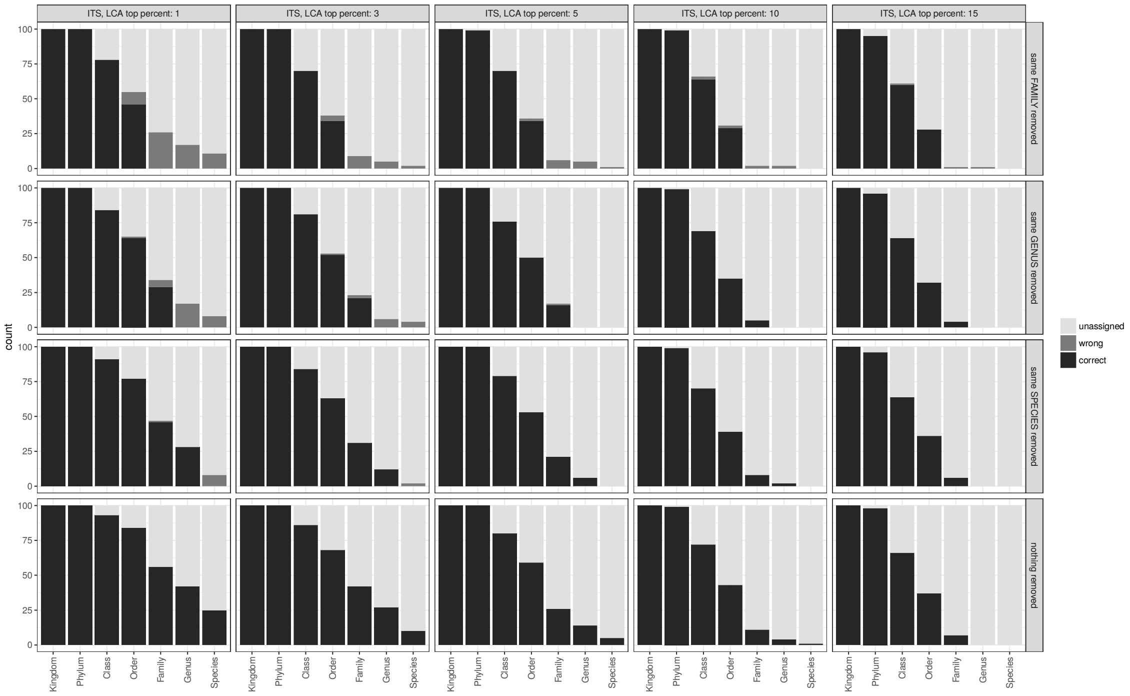
5 Supplemental Figure 5



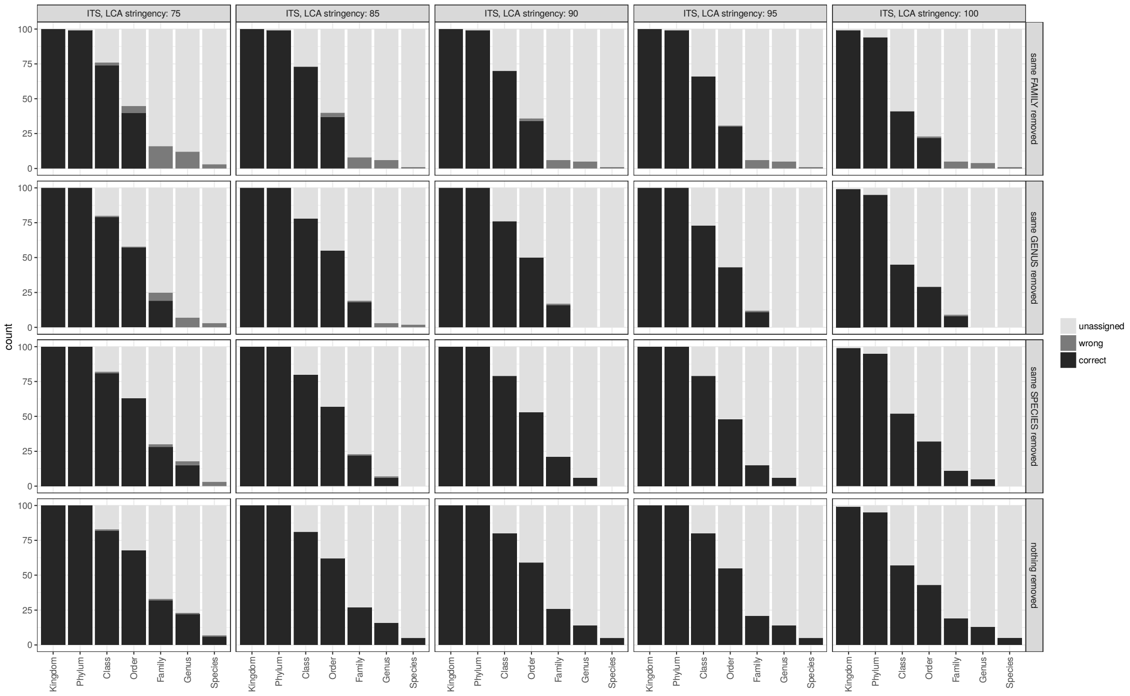
6 Supplemental Figure 6



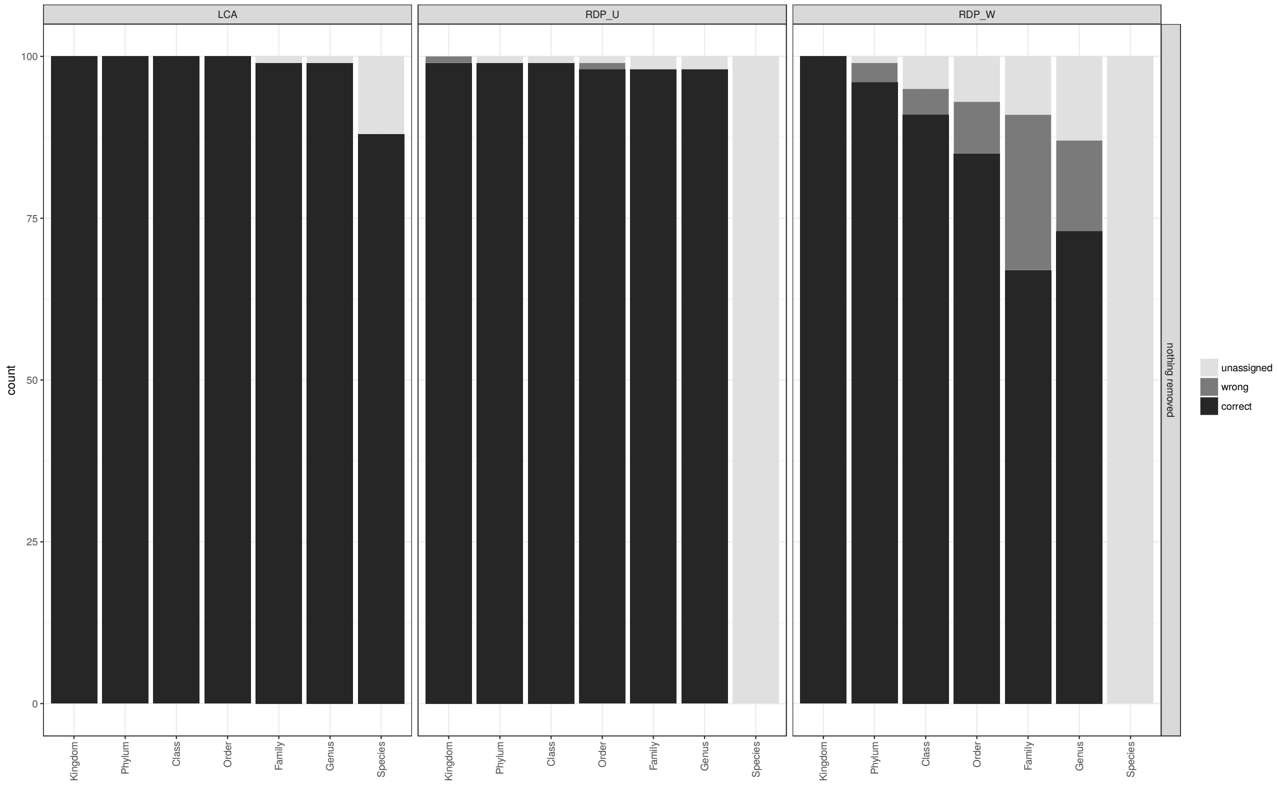
7 Supplemental Figure 7



8 Supplemental Figure 8



9 Supplemental Figure 9



VIII Appendix 2

1 Supplemental Info 1: Pipeline steps

The following is a list of the main rules in the PacBio metabarcoding workflow with a short description of what is done in each step.

getFullCls

For each pre-cluster, representative sequences get a classification. This can be either i) CHIMERA if the reference base chimera detection called this as chimeric (Y) or possibly chimeric (?), ii) UNKNOWN if the sequence was not called as chimeric, but not match to an isolate consensus sequence was found or iii) the name of species, this is given by the highest scoring match to a isolate consensus sequence for non-chimeric sequences.

fullMapping

Run blasr to map representatives of non-chimeric pre-clusters against the isolate consensus sequences. The following parameters are used: `-m 5` to get tabular output, `--bestn 50` to get a maximum of 50 hits for each query and `--minPctSimilarity 90` to get only hits with at least 90% identity.

removeChimeraRef

Run vsearch to remove chimeras with reference based approach. The `--uchime_ref` parameter is used to run the reference based chimera detection algorithm and the `--db` parameter to give the isolate consensus sequences selected by the getFullRef rule as reference sequenes.

getFullRef

For each isolate sample get the consensus sequence of the biggest pre-cluster. Will give an error if there are more than one pre-cluster with 10 or more reads for one sample. Compares sequences for replicates of each species (if available) and writes a warning to the log file if a difference is encountered.

getCorrectCls

Get “correct” classifications for reads in each OTU according to mappings to isolate consensus sequences. Each OTU might have multiple species with read numbers listed.

classifyLSU

Classify OTUs by matches of LSU sequences to the RDP LSU database. See main methods section for details and parameters.

alignToRdp

Run lambda to get local alignments of each OTU LSU representative to the RDP LSU database. Lambda is run with the following parameters: `--output-columns "std qlen slen"` to get query length and subject length along with the default columns in the output table (these are used for coverage computation later), `-p blastn` to run in blastn (nucleotide vs nucleotide) mode, `-nm 5000` to get more matches per query sequence (this is important due to the high number of similar sequences in the database), `-b -2` to the square root of the query length as width for the banded alignment optimization (because higher indel rate and, even more important, uneven insertion/deletion ratio cause the alignment to leave the band), `-x 40` to reduce the x-drop value which can cause alignments to be terminated prematurely and `-as F` to disable adaptive seeding which normally is used to reduce number of hits. Parameters were optimized to allow for alignments for all mock community species.

classifySSU

Classify OTUs by matches of SSU sequences to the SILVA database. See main methods section for details and parameters.

alignToSilva

Run lambda to get local alignments of each OTU SSU representative to the SILVA database. Lambda is run with the following parameters: `--output-columns "std qlen slen"` to get query length and subject length along with the default columns in the output table (these are used for coverage computation later), `-p blastn` to run in blastn (nucleotide vs nucleotide) mode, `-nm 20000` to get more matches per query sequence (this is important due to the high number of similar sequences in the database), `-b -2` to th square root of the query length as width for the banded alignment optimization (because higher indel rate and, even more important, uneven insertion/deletion ratio cause the alignment to leave the band), `-x 30` to reduce the x-drop value which can cause alignments to be terminated prematurely and `-as F` to disable adaptive seeding which normally is used to reduce number of hits. Parameters were optimized to allow for alignments of all mock community species.

classifyITS

Classify OTUs by matches to the UNITE database. See main methods section for details and parameters.

alignToUnite

Run lambda to get local alignments of each OTU representative to the UNITE database. Lambda is run with default parameters except for: `--output-columns "std qlen slen"` to get query length and subject length along with the default columns in the output table (these are used for coverage computation later) and `-p blastn` to run in blastn (nucleotide vs nucleotide) mode.

otuCluster

Run vsearch to cluster OTUs at 97% identity threshold. The following parameters are used for vsearch: `--cluster_size` to choose cluster seeds by descending pre-cluster size (according to size annotation), `--relabel otu` to name OTUs with out and running number instead of using the first sequence as a name, `--sizein --sizeout` to read and write size annotation, `--iddef 0` to use the identity definition of CD-Hit (see vsearch manual), `--id 0.97` to use

97% identity threshold and `--minsl 0.9` to only accept alignments of at least 90% coverage for similarity computation. In addition `--centroids` is used to output a representative centroid sequence for each OTU.

itsx

Run ITSx to separate the different regions of the rRNA operon. The following parameters were set for ITSx: `-t .` to use HMM models from all available taxonomic groups, `--save_regions SSU,ITS1,5.8S,ITS2,LSU` to save separate files for all different regions of the rRNA operon, `--complement F` to not allow reverse-complement detection (all sequences were orientated in forward direction in the primerFilter rule), `--partial 500` to allow for partial matches (we do not have complete SSU and LSU sequences in the amplicon) and `-E 1e-4` to allow for HMM hits with slightly lower e-values (this was optimized to make sure that all rRNA operons in the mock community species were recognized).

removeChimera

Run vsearch to remove chimeras in *de novo* mode. Vsearch is run with the `–uchime_denovo` parameter. All parameters for the chimera detection algorithm are left at default values. Vsearch automatically uses size annotations form the pre-clustering step for its greedy algorithm.

preCluster

Run vsearch to create pre-clusters at 99% identity threshold. The following parameters are used for vsearch: `--usersort --cluster_smallmem` to choose cluster seeds in the order the sequences are sorted in the input file, `--relabel` to name pre-clusters according to the given scheme instead of using the first sequence as a name, `--sizeout` to add size annotation to the output, `--iddef 0` to use the identity definition of CD-Hit (see vsearch manual), `--id 0.99` to use 99% identity threshold and `--minsl 0.9` to only accept alignments of at least 90% coverage for similarity computation. In addition `--consout` is used to generate consensus sequences for each pre-cluster.

prepPrecluster

Reads are sorted by descending mean quality (see qualityFilter rule for computation). This helps to use high quality reads as cluster seeds for pre-clusters in the next step.

filterPrimer

Primers are found and cut with cutadapt. Cutadapt is run with default parameters except for `--trimmed-only` to only retain reads were the primer was found and `-O 10.` Cutadapt is configured to search for both the forward and the reverse primer at the start of the sequence. For sequences where the forward primer was found, the reverse-complemented reverse primer is search at the end of the sequence with an additional run of cutadapt. Accordingly for sequences where the reverse primer was found, the reverse-complemented forward primer is searched at the end of the sequence with another run of cutadapt. In the end sequences with forward-reverse primer combination and reverse-complemented sequences with reverse-forward primer combination are concatenated into one file.

windowQualFilter

For overlapping windows of size 8 the mean error rate is computed from the Phred scores with the same formula as in the qualityFilter rule (except that S is the substring in the windiw instead of the whole sequence). If any window in a sequence has a mean error rate of 0.9 or higher the sequence is removed.

qualityFilter

Mean error rate per sequence is computed from the Phred score given in the fastq file with the formula: $\frac{\sum 10^{-q/10}}{length(S)}$ with q being the quality values of sequence S. Sequences with an error rate of 0.4% or more are written to a separate file (not further used).

lengthFilter

Sequences with length above 6,500 or below 3,000 are printed to separate files (not used further).

filterSilva

Filter SILVA sequences by the quality and pintail (chimera probability) values given in the database. Only sequences with a quality value of at least 85 and pintail value of at least 50 are retained.

2 Supplemental Table 1

Table 10: PCR conditions for chimera tests and resulting chimera formation rates

	Barcode used	Conditions	Template input (ng)	No. PCR cycles	chimera formation rate
first runs	0009	emulsion PCR	first PCR: 25ng, scale up PCR: 2ng	25, 25	4.44%
	0018	standard PCR	8ng	13	0.21%
	0027	standard PCR	8ng	15	0.00%
	0056	standard PCR	8ng	18	1.36%
	0075	standard PCR	8ng	25	14.14%
	0095	standard PCR	2ng	18	0.60%
	0034	standard PCR	20ng	18	0.29%
additional runs	0018	standard PCR	8ng	30	16.27%

IX Appendix 3

1 Supplemental Info 1

Digital normalization was done with the khmer package with the following steps and parameters:

1. `interleave-reads.py`
2. `normalize-by-median.py -C 20 -k 20 -N 4 -x 2.5e8`
3. `filter-abund.py clavariopsis.kh`
4. `normalize-by-median.py -C 5 -k 20 -N 4 -x 1e8`
5. `extract-paired-reads.py`

Genome assembly was done was done with velvet with the digitally normalized read with the following commands:

1. `velveth –fastq.gz –shortPaired`
2. `velvetg –ins_length 900 –exp_cov 140 –cov_cutoff 50`

***De Novo* transcriptome assembly** was done with Trinity with the following command:

```
Trinity --seqType fq --max_memory 100G --single $INDATA --CPU 12
--trimmomatic --normalize_reads --SS_lib_type R
```

Genome guided transcriptome assembly was done with Trinity with the following command:

```
Trinity --genome_guided_bam $INDATA --genome_guided_max_intron
1000 --max_memory 30G --CPU 12
```

The **PASA pipeline** was run with the following commands:

1. `cat Trinity.fasta Trinity-GG.fasta > ${TRANSCRIPTS}`
2. `accession_extractor.pl < Trinity.fasta > tdn.accs`

3. `Launch_PASA_pipeline.pl -c alignAssembly.conf -C -R -g ${GENOME} -t ${TRANSCRIPTS} --ALIGNERS blat,gmap --TDN tdn.accs --transcribed_is_aligned_orient --MAX_INTRON_LENGTH 3000`
4. `build_comprehensive_transcriptome.dbi -c alignAssembly.conf -t ${TRANSCRIPTS} --min_per_ID 95 --min_per_aligned 30`

Supplemental Table 1

condition	CAZy Family	activity	genes with this annotation in the genome	differentially expressed genes with this annotation	activation probability
alder-straw	AA9	AA9 (formerly GH61) proteins are copper-dependent lytic polysaccharide monooxygenases (LPMOs); cleavage of cellulose chains with oxidation of various carbons (C-1, C-4 and C-6) has been reported several times in the literature;	49	35	1
alder-straw	CE1	acetyl xylan esterase; cinnamoyl esterase; feruloyl esterase; carboxylesterase; S-formylglutathione hydrolase; diacylglycerol O-acyltransferase; trehalose 6-O-mycolyltransferase	10	9	0.9826
alder-straw	GH11	endo-β-1,4-xylanase; endo-β-1,3-xylanase	6	6	0.935
alder-straw	GH7	endo-β-1,4-glucanase; reducing end-acting cellobiohydrolase; chitosanase; endo-β-1,3-1,4-glucanase	7	6	0.8518
alder-straw	GH131	broad specificity exo-β-1,3/1,6-glucanase with endo-β-1,4-glucanase activity;	4	4	0.7446
alder-straw	GH10	endo-1,4-β-xylanase; endo-1,3-β-xylanase; tomatinase; xylan endotransglycosylase	4	4	0.7428
alder-straw	GH5_5	endo-β-1,4-glucanase / cellulase; endo-β-1,4-xylanase; β-glucosidase; β-mannosidase; β-glucosylceramidase; glucan β-1,3-glucosidase; licheninase; exo-β-1,4-glucanase / cellodextrinase; glucan endo-1,6-β-glucosidase; mannan endo-β-1,4-mannosidase; cellulose β-1,4-cellobiosidase; steryl β-glucosidase; endoglycoceramidase; chitosanase; β-primeverosidase; xyloglucan-specific endo-β-1,4-glucanase; endo-β-1,6-galactanase; hesperidin 6-O-α-L-rhamnosyl-β-	5	5	0.6984

		glucosidase; β-1,3-mannanase; arabinoxylan-specific endo-β-1,4-xylanase; mannan transglycosylase			
solid-liquidExp	GH7	endo-β-1,4-glucanase; reducing end-acting cellobiohydrolase; chitosanase; endo-β-1,3-1,4-glucanase	7	7	0.937
solid-liquidExp	AA9	AA9 (formerly GH61) proteins are copper-dependent lytic polysaccharide monooxygenases (LPMOs); cleavage of cellulose chains with oxidation of various carbons (C-1, C-4 and C-6) has been reported several times in the literature;	49	26	0.813
solid-liquidExp	GH72	β-1,3-glucanosyltransglycosylase	7	6	0.7628
solid-liquidExp	AA7	glucooligosaccharide oxidase; chitooligosaccharide oxidase	25	14	0.6684
solid-liquidExp	GH3	β-glucosidase; xylan 1,4-β-xylosidase; β-glucosylceramidase; β-N-acetylhexosaminidase; α-L-arabinofuranosidase; glucan 1,3-β-glucosidase; glucan 1,4-β-glucosidase; isoprimeverose-producing oligoxyloglucan hydrolase; coniferin β-glucosidase; exo-1,3-1,4-glucanase; β-N-acetylglucosaminide phosphorylases	19	11	0.6038
solid-liquidSta	AA9	AA9 (formerly GH61) proteins are copper-dependent lytic polysaccharide monooxygenases (LPMOs); cleavage of cellulose chains with oxidation of various carbons (C-1, C-4 and C-6) has been reported several times in the literature;	49	33	1
solid-liquidSta	GH7	endo-β-1,4-glucanase; reducing end-acting cellobiohydrolase; chitosanase; endo-β-1,3-1,4-glucanase	7	7	0.9634
solid-liquidSta	CE16	acetylesterase active on various carbohydrate acetyl esters	5	5	0.847

solid-liquidSta	AA12	The pyrroloquinoline quinone-dependent oxidoreductase activity was demonstrated for the CC1G_09525 protein of Coprinopsis cinerea.	6	5	0.745
solid-liquidSta	GH55	exo-β-1,3-glucanase; endo-β-1,3-glucanase	4	4	0.7364
solid-liquidSta	GH11	endo-β-1,4-xylanase; endo-β-1,3-xylanase	6	5	0.7024
stat-exp	CE8	pectin methylesterase	6	5	0.6276
straw-malt	AA9	AA9 (formerly GH61) proteins are copper-dependent lytic polysaccharide monooxygenases (LPMOs); cleavage of cellulose chains with oxidation of various carbons (C-1, C-4 and C-6) has been reported several times in the literature;	49	35	1
straw-malt	CE1	acetyl xylan esterase; cinnamoyl esterase; feruloyl esterase; carboxylesterase; S-formylglutathione hydrolase; diacylglycerol O-acyltransferase; trehalose 6-O-mycolyltransferase	10	9	0.9882
straw-malt	GH11	endo-β-1,4-xylanase; endo-β-1,3-xylanase	6	6	0.96
straw-malt	GH7	endo-β-1,4-glucanase; reducing end-acting cellobiohydrolase; chitosanase; endo-β-1,3-1,4-glucanase	7	6	0.877
straw-malt	GH10	endo-1,4-β-xylanase; endo-1,3-β-xylanase; tomatinase; xylan endotransglycosylase	4	4	0.7582
straw-malt	GH5_5	endo-β-1,4-glucanase / cellulase; endo-β-1,4-xylanase; β-glucosidase; β-mannosidase; β-glucosylceramidase; glucan β-1,3-glucosidase; licheninase; exo-β-1,4-glucanase / cellodextrinase; glucan endo-1,6-β-glucosidase; mannan endo-β-1,4-mannosidase; cellulose β-1,4-cellobiosidase; steryl β-glucosidase; endoglycoceramidase; chitosanase; β-primeverosidase; xyloglucan-specific endo-β-1,4-glucanase; endo-β-1,6-galactanase; hesperidin 6-O-α-L-rhamnosyl-β-	5	5	0.6948

		glucosidase; β-1,3-mannanase; arabinoxylan-specific endo-β-1,4-xylanase; mannan transglycosylase			

2 Supplemental Table 2

condition	KEGG pathway ID	KEGG pathway name	genes with this annotation in the genome	differentially expressed genes with this annotation	activation probability
alder-malt	ko00040	Pentose and glucuronate interconversions	35	19	1
alder-malt	ko01120	Microbial metabolism in diverse environments	246	92	1
alder-malt	ko04146	Peroxisome	54	35	1
alder-malt	ko00280	Valine, leucine and isoleucine degradation	49	26	1
alder-malt	ko00640	Propanoate metabolism	28	12	0.9992
alder-malt	ko00460	Cyanoamino acid metabolism	26	13	0.9868
alder-malt	ko00906	Carotenoid biosynthesis	4	4	0.9804
alder-malt	ko04978	Mineral absorption	7	4	0.9438
alder-malt	ko00052	Galactose metabolism	30	12	0.7968
alder-malt	ko04920	Adipocytokine signaling pathway	9	4	0.6656
alder-malt	ko04260	Cardiac muscle contraction	12	4	0.6062
alder-straw	ko00500	Starch and sucrose metabolism	65	32	1
alder-straw	ko01120	Microbial metabolism in diverse environments	246	96	1
alder-straw	ko04146	Peroxisome	54	33	1
alder-straw	ko00040	Pentose and glucuronate interconversions	35	16	0.9972

alder-straw	ko00190	Oxidative phosphorylation	73	25	0.9796
alder-straw	ko00520	Amino sugar and nucleotide sugar metabolism	49	18	0.974
alder-straw	ko00770	Pantothenate and CoA biosynthesis	23	10	0.9574
alder-straw	ko04142	Lysosome	40	13	0.955
alder-straw	ko00330	Arginine and proline metabolism	38	17	0.9466
alder-straw	ko00906	Carotenoid biosynthesis	4	4	0.9202
alder-straw	ko00600	Sphingolipid metabolism	21	11	0.8762
alder-straw	ko00910	Nitrogen metabolism	21	6	0.8466
alder-straw	ko00780	Biotin metabolism	9	5	0.7978
alder-straw	ko00740	Riboflavin metabolism	11	5	0.7602
alder-straw	ko04978	Mineral absorption	7	4	0.727
alder-straw	ko00380	Tryptophan metabolism	40	21	0.702
alder-straw	ko05130	Pathogenic Escherichia coli infection	13	5	0.6742
alder-straw	ko00640	Propanoate metabolism	28	13	0.6634
solid-liquidExp	ko00970	Aminoacyl-tRNA biosynthesis	40	28	0.9998
solid-liquidExp	ko01230	Biosynthesis of amino acids	116	58	0.9152
solid-liquidExp	ko00520	Amino sugar and nucleotide sugar metabolism	49	28	0.8352
solid-liquidExp	ko03030	DNA replication	32	27	0.7924

solid-liquidExp	ko03008	Ribosome biogenesis in eukaryotes	64	31	0.7612
solid-liquidExp	ko00280	Valine, leucine and isoleucine degradation	49	26	0.7048
solid-liquidExp	ko04111	Cell cycle - yeast	75	43	0.6586
solid-liquidExp	ko01524	Platinum drug resistance	29	18	0.6358
solid-liquidExp	ko03050	Proteasome	35	18	0.6236
solid-liquidSta	ko03008	Ribosome biogenesis in eukaryotes	64	38	1
solid-liquidSta	ko00500	Starch and sucrose metabolism	65	38	0.997
solid-liquidSta	ko00040	Pentose and glucuronate interconversions	35	25	0.9964
solid-liquidSta	ko03030	DNA replication	32	22	0.983
solid-liquidSta	ko00970	Aminoacyl-tRNA biosynthesis	40	22	0.9042
solid-liquidSta	ko00052	Galactose metabolism	30	20	0.882
solid-liquidSta	ko00564	Glycerophospholipid metabolism	39	22	0.8494
solid-liquidSta	ko00770	Pantothenate and CoA biosynthesis	23	15	0.7834
solid-	ko00965	Betalain biosynthesis	17	12	0.6102

liquidSta					
stat-exp	ko00040	Pentose and glucuronate interconversions	35	21	1
stat-exp	ko03010	Ribosome	99	29	1
stat-exp	ko01120	Microbial metabolism in diverse environments	246	60	0.9772
stat-exp	ko00280	Valine, leucine and isoleucine degradation	49	17	0.9676
stat-exp	ko00520	Amino sugar and nucleotide sugar metabolism	49	12	0.8906
stat-exp	ko00620	Pyruvate metabolism	41	11	0.8554
stat-exp	ko00260	Glycine, serine and threonine metabolism	47	11	0.845
stat-exp	ko00250	Alanine, aspartate and glutamate metabolism	31	9	0.7604
stat-exp	ko01110	Biosynthesis of secondary metabolites	355	94	0.7382
straw-malt	ko00500	Starch and sucrose metabolism	65	31	1
straw-malt	ko00040	Pentose and glucuronate interconversions	35	19	1
straw-malt	ko03008	Ribosome biogenesis in eukaryotes	64	23	1
straw-malt	ko00330	Arginine and proline metabolism	38	18	0.9978
straw-malt	ko00520	Amino sugar and nucleotide sugar metabolism	49	17	0.967
straw-malt	ko00980	Metabolism of xenobiotics by cytochrome P450	28	14	0.9512
straw-malt	ko00630	Glyoxylate and dicarboxylate metabolism	40	21	0.9136
straw-malt	ko00920	Sulfur metabolism	17	8	0.9056
straw-malt	ko00052	Galactose metabolism	30	15	0.897
straw-malt	ko00350	Tyrosine metabolism	50	20	0.838

straw-malt	ko00770	Pantothenate and CoA biosynthesis	23	11	0.7994
straw-malt	ko00910	Nitrogen metabolism	21	9	0.6518
straw-malt	ko01220	Degradation of aromatic compounds	32	17	0.6144

3 Supplemental Table 3

condition	GO ID	GO name	genes with this annotation in the genome	differentially expressed genes with this annotation	activation probability
alder-malt	GO:0016491	oxidoreductase activity	506	205	1
alder-malt	GO:0005975	carbohydrate metabolic process	226	72	1
alder-malt	GO:0071949	FAD binding	55	29	1
alder-malt	GO:0055085	transmembrane transport	513	165	1
alder-malt	GO:0055114	oxidation-reduction process	768	283	1
alder-malt	GO:0008152	metabolic process	387	158	1
alder-malt	GO:0008080	N-acetyltransferase activity	49	19	0.9998
alder-malt	GO:0006508	proteolysis	132	39	0.7976
alder-straw	GO:0016491	oxidoreductase activity	506	211	1
alder-straw	GO:0003824	catalytic activity	610	212	1
alder-straw	GO:0016787	hydrolase activity	165	57	1
alder-straw	GO:0071949	FAD binding	55	28	1
alder-straw	GO:0006508	proteolysis	132	53	1
alder-straw	GO:0055114	oxidation-reduction process	768	301	1

alder-straw	GO:0055085	transmembrane transport	513	176	0.9952
alder-straw	GO:0005975	carbohydrate metabolic process	226	110	0.988
solid-liquidExp	GO:0005524	ATP binding	497	231	1
solid-liquidExp	GO:0005515	protein binding	749	314	1
solid-liquidExp	GO:0003824	catalytic activity	610	256	1
solid-liquidExp	GO:0055085	transmembrane transport	513	230	1
solid-liquidExp	GO:0055114	oxidation-reduction process	768	341	0.9992
solid-liquidExp	GO:0006508	proteolysis	132	68	0.9848
solid-liquidExp	GO:0005975	carbohydrate metabolic process	226	115	0.9762
solid-liquidExp	GO:0005634	nucleus	389	180	0.9716
solid-liquidExp	GO:0016787	hydrolase activity	165	77	0.913
solid-liquidExp	GO:0016491	oxidoreductase activity	506	231	0.8836
solid-liquidExp	GO:0003676	nucleic acid binding	294	130	0.6472
solid-	GO:0003824	catalytic activity	610	283	1

liquidSta					
solid-liquidSta	GO:0005975	carbohydrate metabolic process	226	125	0.9898
solid-liquidSta	GO:0055114	oxidation-reduction process	768	335	0.9868
solid-liquidSta	GO:0008152	metabolic process	387	180	0.9798
solid-liquidSta	GO:0055085	transmembrane transport	513	264	0.975
solid-liquidSta	GO:0006508	proteolysis	132	66	0.8306
stat-exp	GO:0003824	catalytic activity	610	164	1
stat-exp	GO:0055114	oxidation-reduction process	768	232	1
stat-exp	GO:0016491	oxidoreductase activity	506	153	0.9988
stat-exp	GO:0071949	FAD binding	55	21	0.9874
stat-exp	GO:0005975	carbohydrate metabolic process	226	67	0.9518
stat-exp	GO:0055085	transmembrane transport	513	113	0.8948
straw-malt	GO:0008080	N-acetyltransferase activity	49	22	1
straw-malt	GO:0016491	oxidoreductase activity	506	192	1
straw-malt	GO:0005975	carbohydrate metabolic process	226	117	1
straw-malt	GO:0003824	catalytic activity	610	198	1
straw-malt	GO:0016787	hydrolase activity	165	54	1

straw-malt	GO:0071949	FAD binding	55	31	1
straw-malt	GO:0006508	proteolysis	132	53	1
straw-malt	GO:0055085	transmembrane transport	513	188	1
straw-malt	GO:0055114	oxidation-reduction process	768	279	1
straw-malt	GO:0000981	RNA polymerase II transcription factor activity, sequence-specific DNA binding	217	66	0.9982
straw-malt	GO:0042254	ribosome biogenesis	15	9	0.634