Exploring fungal disease associations using genomic data and network models

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Biology, Medicine and Health

2021

Ignacio Riquelme Medina

School of Biological Sciences

List of Contents

Final word count: 44394 words.

List of Figures

List of Tables

Abbreviations: AFToL, Assembling the fungal tree of life project; AIC, Akaike information criterion; BIC, Bayesian information criterion; BC, betweenness centrality; CSF, University of Manchester computational shared facility; ERGMs, exponential random graph models; FC, fold change; FGP, 1000 Fungal Genomes Project; GCC, giant connected component; GO, gene ontology; GWAS, genome-wide association studies; HGT, horizontal gene transfer; LMMs, linear mixed models; MCL, Markov clustering algorithm; MCMC, Markov chain Monte Carlo; MRP, matrix representation with parsimony; PPD, posterior probability distribution; pI, isoelectric point; RefSeq, NCBI's reference sequence database; SSN, sequence similarity network; WoL, ways of life.

Abstract

Convergent evolution is a process by which different distantly related species can evolve the same trait, usually involving adaptation to similar environments, and it is a widespread phenomenon thorough all groups of life. One of the groups where convergent evolution could be common is the fungal kingdom since they have repeatedly and independently adapted to similar environments through their evolutionary history. One way to detect convergent evolution is by using association networks, where groups of genes that appear together more often than expected can be identified. These groups of genes are usually related to a function or process, and by considering the habitat of the species that appear in one set of associated genes we can potentially link the set to a particular phenotype. To exclude traits that are found in different species due to common ancestry, association networks need to be considered in the context of the fungal phylogeny. Fungal phylogenetics is still a very active research area, in large part due to the presence of some problematic taxa such as the Microsporidia, which are intracellular parasites that have lost most of their genome. In this thesis, in order to resolve the phylogenetic positions of problematic groups, I have used tree and data heterogeneous phylogenetic models, that are able to account for different evolutionary processes in different proteins and in different parts of the tree.

To investigate and demonstrate the utility of networks for uncovering evolutionary processes, we used bipartite networks to identify evolutionary signals in plasmids. Traditional phylogenetic methodology cannot be used to portray the overall evolutionary history of plasmids, due to the lack of common genes. Therefore, networks allow us to connect plasmids through overlapping gene sets even if there are no genes that are common to all plasmids. Through the investigation of community structure, which emerges throughout evolutionary time as a consequence of the interactions of plasmids with one another, I have been able to associate part of the networks with certain plasmid features, like host taxonomy or function. Finally, I investigated plasmid evolution further by studying how the physical properties of the nucleotide sequences that forms each plasmid can affect plasmid interactions by using Exponential Random Graph Models.

Declaration

The author of this thesis declares that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made **only** in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University's policy on Presentation of Theses

Introduction

Convergent evolution

Convergent evolution is the process by which unrelated genes in different species evolve independently, in similar ways to give raise to similar phenotypes. This process is thought to be either completely independent, where similar independent phenotypes arise by random chance, or caused by adaptation to similar environments, where similar evolutionary pressure makes independent genes evolve in similar ways (1). Convergent evolution is a widespread occurrence across all organisms and since it has been detected in all Domains of life, both at genotypical and phenotypical levels (2–7). Because of independent similar adaptation to the same environments, genes tend to show the same kinds of mutations and nucleotide sequence modifications, which can lead to high levels of sequence similarity between genes involved in convergent evolution (8). Convergent evolution has been used as one of the possible explanations of the phenomenon of homoplasy (along with parallelism and reversal), where genes with high sequence similarity appear independently in species that do not share a very distant common ancestor (9).

There are many examples in nature where similar complex traits and structures have evolved in lineages that do not have a close phylogenetic relationship. One example is the evolution of the eye, which has evolved independently in different lineages of animals resulting in similar eye structures, like spherical shape and lenses (10). Another example of convergent evolution is seen in echolocation in bats and marine mammals. Both animal groups have undergone similar sequence changes that make them more sensitive to high frequency waves required for echolocation in the *Prestin* gene, along with other changes in genes related to hearing and vision (6, 11). Finally, the rise of high intelligence in animals is also thought to be a case of convergent evolution, since it is linked to the formation of multimodal centres in the brain structure of insects, molluscs, mammals and birds (12). By using the combined information of similarities between phenotypic features and sequences that appears in cases of convergent evolution, several methods have been developed in order to detect, measure and quantify convergent evolution (13).

Fungi

One of the groups of organisms where convergent evolution is likely to be common is the fungal kingdom. Fungi diverged from the animal kingdom about one billion years ago (14). They are characterized by being highly variable, eukaryotic, with simple morphologies (typically filamentous but it can also be unicellular). Fungi have relatively small eukaryotic genomes and all of them are heterotrophic (14). They produce haploid and diploid forms in their life cycle and do not have movement except for some flagellated spores. They can live as saprobes, symbionts and parasites of plants, animals or other fungi, but they are not able to photosynthesise. Owing to the high fungal variability, these ways of life (WoL) are not tied to a particular monophyletic branch but have appeared repeatedly through all the fungal phylogeny (14). One example of adaptation to different WoL in evolutionarily close organisms is observed in the order *Hymenochaetales*, where we can find saprophytic species, mycorrhizal species, and both opportunistic plant parasites and obligate plant parasites amongst its ranks (15).

Many fungi have an impact to humans, either indirectly by infecting species of economic importance or directly as pathogens of humans. For the species that have an economic impact the most important fungi are plant parasites responsible for crop diseases. Species like *Magnaporthe oryzae*, responsible for the rice blast disease, can lead to the loss of between 10% and 30% of the harvest or even more if it is epidemic (16). Rice blast disease is not only an economical problem but it also has an impact in human populations where rice is one of the main food sources, as it is one of the most consumed cereals in the world. Another example is *Botrytis cinerea,* which can infect a wide variety of plants and fruits, thereby causing a huge economic impact due to the need to use fungicides for their control (estimated at €540 millions in 2001) apart from the damage they can cause to the crops. Fungal plant pathogens are classified in three groups: fungi that maintain the host cell alive for a long period of time (biotroph), fungi that rapidly kills the host cell (necromorph) and fungi that maintain the host cell alive during the first stages of infection but kill the host in later stages (hemibiotroph) (17). Fungi play also a big role in food spoilage as saprotrophs, resulting in many methods of preservation being developed throughout human history (18–20). As for fungi that can cause human disease, most of the fungal infections are superficial, most often affecting specially skin and nails, with a very low mortality (21). However, there are also invasive fungal infections, particularly fungi like *Candida* or *Aspergillus* species*,* that have high mortality rates and are especially dangerous for

immunosuppressed people. In animals, most fungi are obligate pathogens (mainly affecting the lungs) or part of the normal microbiota that can become infectious under certain conditions (opportunistic pathogens). In addition, there has been an emergence of new high virulent fungal diseases that are threatening biological diversity in many places of the world (22).

In addition to the negative aspects, fungi can also have a positive impact on human activity. Due to the relatively small fungal genome size, the high resilience that some fungi have to genetic manipulations and that some fungi form tissues that allow transcriptomics experiments, fungi have been regarded as good model organism for eukaryotes. Fungi have been extensively used in research, in particular, species like *Saccharomyces cerevisiae. S. cerevisiae* has been used as a model system for research into eukaryotic processes such as the cell cycle (23) or for drug discovery in cancer (24). Many other fungal species have been investigated, given the huge variability and potential for drug discovery of fungal secondary metabolites (25). Fungi have long been used by humans for as fermenting agents in both the food and the beverages industries. More recently, fungi have found use in industrial applications, such as the biodegradation of pollutants (26). Mycorrhizal fungi, which form mutualistic communities with the roots of certain plants, enhance biodiversity and variability in some ecosystems (27) and can also increase crop growth and yield (28, 29).

Given the impact fungi can have in humans, both negative and positive, the efforts to sequence the genomes of fungi have intensified in recent years, with a particular focus on the lesser known species. One of these efforts is the 1000 Fungal Genomes Project (FGP) (30) which is a project that intends to sequence 1000 fungal genomes and make them accessible to the public. Currently, at the time of writing, there are 690 genomes completed and available online. These genomes are available to download and in most cases the gene and protein sequences of the fungal species, which have been functionally annotated. The complete list of the genomes used for this project, complete with FGP identifiers and links to the FGP database, can be checked in Table 2 in the supplementary information section.

The fungal phylogeny has been controversial and there are still ongoing efforts to fully consolidate many of the smaller groups in the fungal taxonomy. Our knowledge of fungal phylogenetic relationships have been enhanced by the Assembling the Fungal Tree of Life project (AFToL) (31) due to the lack of tools to reach a consensus classification in former times (32). The main clades in fungi, from the earliest splitting branch (14), as showed in Figure 1, are:

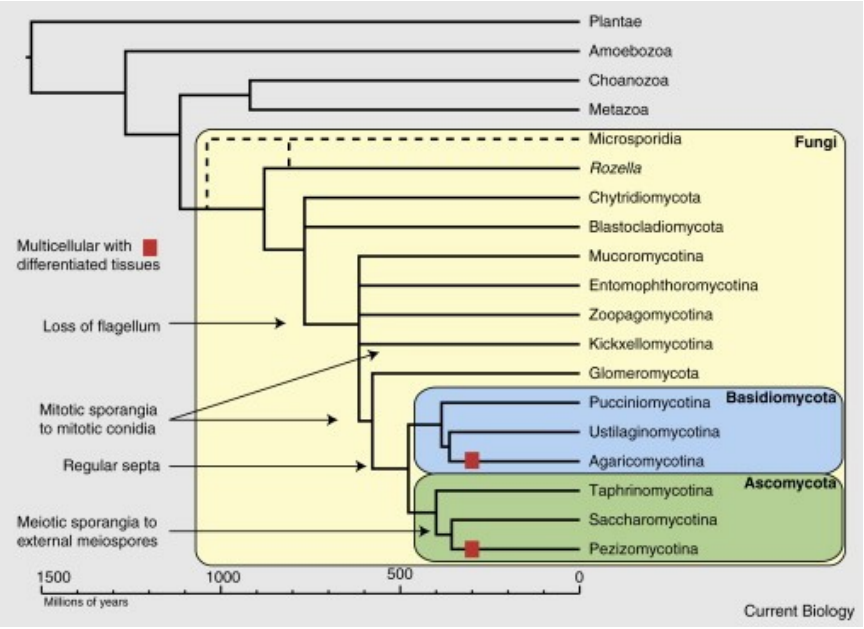

Figure 1: Current consensus phylogenetic tree of the major groups in the fungi kingdom (14).

- Chytridiomycota: They have a unicellular body surrounded by a cell wall that produces flagellated zoospores. Some of them are able to form filaments. Many of them are saprobes, but they can also be parasites or mutualists. Some important species in this group are Batrachochytrium, a parasite related to the amphibian decline, Neocallimastix, which is a mutualist found in the stomachs of ruminants and Synchytrium, responsible for the potato black wart.
- Blastocladiomycota (33): They are very similar to Chytridiomycota, but they have a more hyphal (filament) growth phase and an unusual alternation between the haploid and diploid forms. They can be saprotrophs or parasites of plants and animals. One important species is *Coelomomyces*, a parasite of mosquito larvae.
- Mucoromycotina: Saprobes. They usually grow as filaments, and form zoospores without flagella or cell walls.
- Entomophthoromycotina, Zoopagomycotina and Kickxellomycotina: With similar characteristics as the mucoromycotina. Entomophthoromycotina are insect parasites, Zoopagomycotina are fungi and animal parasites and Kickxellomycotina are also fungi and animal parasites with the addition of also being saprobes.
- Glomeromycota (34): They are one of the most ecologically important mutualists since they form the arbuscular mycorrhizae with the roots of most of the plant species.

- Dikarya: The most studied fungal phylogenetic group, its members are characterized by the lack of flagella and the formation of filaments with cells that have 2 nuclei, which constitutes a good part of their life cycle, but they can also be unicellular. This fungal subkingdom is further divided into two phyla: Ascomycota and Basidiomycota. Ascomycota are characterized by the formation of ascospores inside a sac like structure and have adapted to every lifestyle present in the fungal kingdom, from parasites to mutualists. Basidiomycota are characterized by their sexual reproduction via specialized cells called basidia and have also adapted to a variety of lifestyles except for mutualism. There are many important species in this clade, like *Penicillium* or *Saccharomyces*, which are of great value to humans, or *Fusarium*, *Coccidioides* or *Malassezia*, which are parasites and have a negative impact for humans.

As we can see in the above list, the different kinds of heterotrophic nutrition are distributed through all the phylogeny, except for a few clades. A more detailed tree of the fungi where the orders belonging to each clade are showed can be seen in Figure 2.

Owing to extensive amount of fungal variability, repeated adaptation to similar environments and WoL, and the relatively small genome size, there are good reasons to believe that fungi would be ideal candidates in which to study convergent evolutionary events.

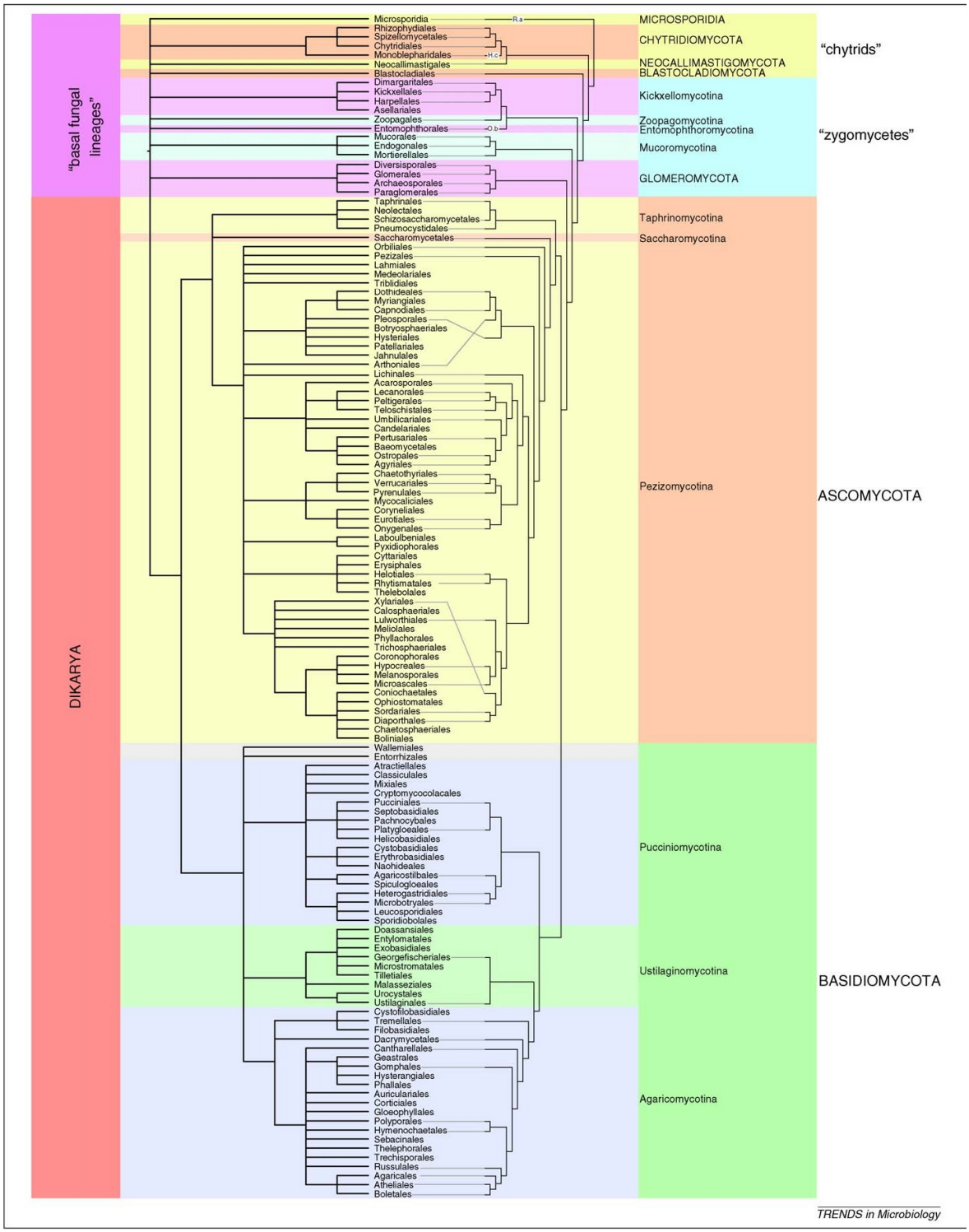

Figure 2: Current fungal classification based on the phylogeny from AFToL (left) and James et al (144) (right).(32)

Networks

Networks can provide a relatively simple representation of complex processes, so that we can better understand the interactions of the individual nodes in the network (35). Each member of the group of objects being studied is called a node of the network, and all the relations that connect a node to other nodes are called edges as it is shown in Figure 3. Edges can represent many types of interactions between nodes, from relations between people to gene homologies (36, 37). The particular networks of interest for this project, biological networks, often have two properties that separate them from other types of networks: they have few nodes with many edges and many nodes with few edges following a power law distribution (scale-free networks) and have the small world property (the average node distance is small compared to a random network of the same size). They also tend to form communities or modules, which are regions of a network of highly interconnected nodes and the nodes within these communities are likely to share some kinds of trait or characteristic (i.e. function) (38, 39).

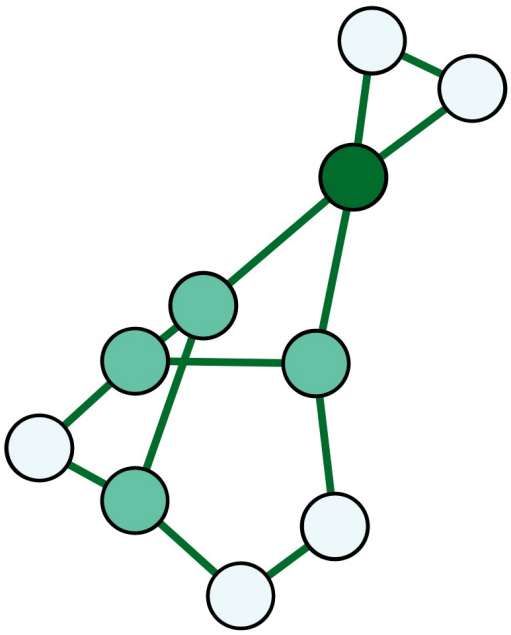

Figure 3: Example of a network. The circles represent nodes (i.e. genes, proteins) and the connections between the nodes are the edges. Darker nodes have higher degree. This network was generated using Gephi (145).

The main tools that are typically used to analyse networks are the topological measures of the network (38, 40, 41). These measures can be used for instance, to check how well a

node is integrated in the network (Degree, betweenness centrality), or to identify and measure the shortest path between nodes (distance) or how well interconnected the nodes are (clustering coefficient). For example, nodes that are highly connected in the network can be of particular interest since they may be essential pieces in a given process. The identification of communities of genes can indicate that the genes in the community might be related by function or they might possess similar genetic elements that govern their regulation. Additionally, many communities can be linked to particular pathways, functions or processes. It is also possible to identify patterns in the network (called motifs) such as loops. Thus, analysing the topological features of a the network can uncover properties of complex biological systems and help to detect key genes that play a central role in pathways, functions or diseases.

Network theory then provides us with tools that, when applied to networks, can decipher the underlying information contained in them allowing us to interpret results. The main measures we will be using are:

- Degree: the connections that a given node has with other nodes, the higher the number of connections the higher the degree.
- Betweenness centrality (BC): Measures the number of shortest paths that go through a node for all the possible shortest paths connecting every node in a network. The higher the BC the more centric a node is in the network structure. It follows the *Equation (1)* of a node *v* where σ_{st} is the total number of shortest paths and $\sigma_{st}(v)$ is the number of shortest paths between *s* and *t* that go through *v.*

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (1)$$

- Modularity: Serves to obtain communities from the network, defined as regions of highly interconnected nodes that connect more to nodes within the community than to nodes in the rest of the network. Modularity also measures how well connected are the nodes within the communities.
- Connected components: Measure to check for the number of sub-networks in a network that are independent, so they do not share connections between them.

By using these measures we can obtain a good amount of relevant biological information. Modules or communities in network theory are known to be related to some characteristic or trait, like taxonomy or function (42). The degree and BC can be used to search for the most connected and central genes in the network or communities. By combining these measurements, we can associate communities in the network with that are related to relevant traits and then we can find central genes within the community structure, as is done in other fields in biology (43–45). These genes often have a relevant function for the trait that the community is related with.

For this project I was interested in using two types of networks. First, bipartite networks, which are a type of network in which two independent groups of nodes connect to each other, but do not allow for nodes within the same group to form edges (46). An example of a bipartite network can be seen in Figure 4, which can be a representation of, as an example, a network that connects proteins to pathways, but not proteins to other proteins or pathways to other pathways. Due to this property, bipartite graphs cannot form clusters as they are defined in unipartite networks (networks with only one group of nodes, e.g. a protein similarity network). Clusters in unipartite networks are defined as regions of highly interconnected nodes measured as the coefficient of the number of triangles (three interconnected nodes) found in a region respecting all possible triangles in the region, since the minimum number of interconnected nodes in a bipartite network is four (47, 48). To get around this issue, some other definitions for clustering have been made for bipartite graphs, for example the ratio between squares (four interconnected nodes) and three paths in the network (48, 49). Bipartite graphs do tend to form communities (47). Bipartite networks can also be collapsed in what is called unipartite projections, where one group of nodes is removed and edges are drawn between nodes of the other group if they were linked to the same node in the removed group. In our project, a bipartite network can be constructed by linking groups of proteins with high sequence similarity to the plasmids where they are found. In this way, we relate traits of particular plasmids such as pathogenesis to proteins or, potentially, to other types of nodes like gene families or proteins regions. In other words, if a protein is found to be related to several plasmids with a particular trait, we can link the protein to the trait.

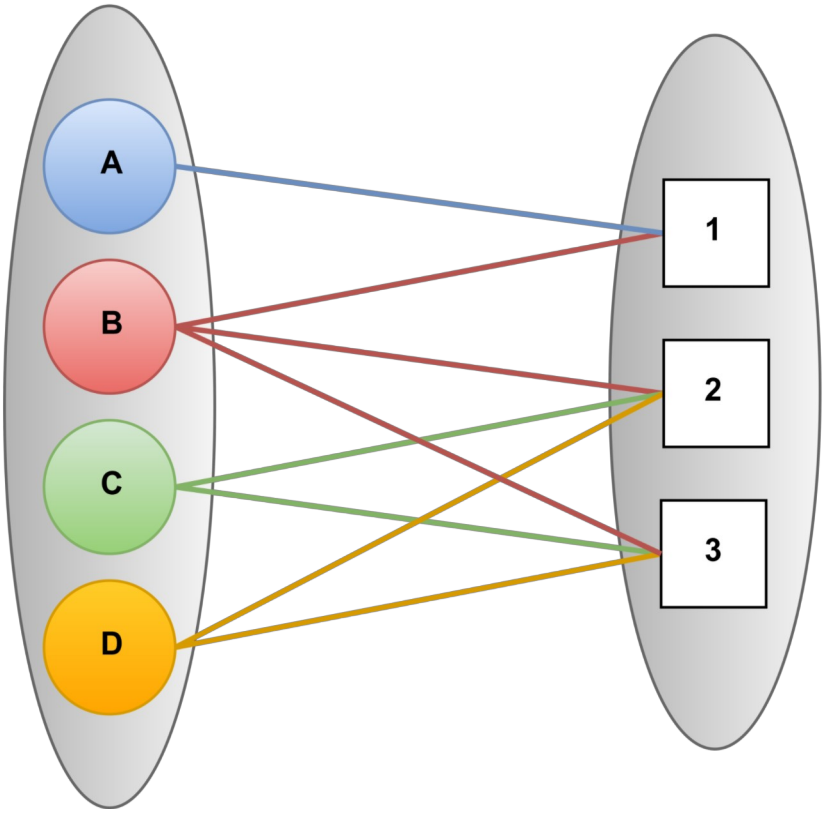

Figure 4: Example of a bipartite network. In this example there are two groups of nodes, letters (left) and numbers (right). Edges can be formed between nodes of different groups, but not within nodes of the same group.

The second type of network we used was association and disassociation networks, which can be particularly helpful to detect convergent evolution events. Unlike bipartite networks, Association networks have only one type of nodes. Instead, Association networks link genes to other genes by their tendency to appear together more often than it is expected by random chance. To avoid making gene associations with genes that only appear together because they are specific to the same phylogenetic clade, the phylogenetic relationships of species in the dataset are taken in account. Therefore, association networks can be used in a similar way to Genome-wide Association Studies (GWAS) where certain genes or proteins are tied to particular traits or phenotypes. In our case, it is assumed that proteins appear together in species because there is some benefit in having both of them that is reflected in the fitness of a species. It could be that both proteins are required for a pathway to function, or that one gene enhances the function of the other. In contrast, disassociation networks try to find nodes that avoid being in the same genome more than it is expected by random change. In this case, proteins would have negative effects on the fitness of a species if they are found together. It may be that both proteins have opposite effects in a same trait, or they may cause toxicity when found together. It is also possible

that redundant proteins are eliminated by natural selection, since they can be costly for the individual. In prokaryotes, co-incidence networks have been successfully used to detect functional associations between proteins (50–52) and to associate the presence of certain functions with habitats (53, 54). However, in order to detect events of convergent evolution, considering only protein co-occurrence in the phylogeny is not enough. In convergent evolution, genes and proteins evolve in a similar way in species that have a relatively distant common ancestor, where the similarity between the genes and proteins cannot be attributed to the common ancestor. In events where groups of proteins co-occur or co-avoid but also share a recent common ancestor, the presence or absence of a group proteins can be explained simply because of common ancestry and not because of independent similar adaptation. Therefore, we sought to find groups of co-occurrence or co-avoiding proteins that do not share a recent common ancestor, thus increasing our chances of finding convergent evolution events that could be linked to convergent evolution. The program we chose to find convergent evolution events is CoinFinder (55), a software program that can detect phylogenetically independent association and disassociation events in groups of genes or proteins from a given similarity network and a phylogeny. The main advantages that CoinFinder has over other similar programs is that it is scalable, which allows its use with large datasets. Other similar programs like Copap (56) were also considered for the project, but Copap can only detect co-occurrence, is not as scalable and much more computationally intensive than CoinFinder, which makes Copap unsuitable for our dataset. There have been previous attempts to use co-occurrence and co-avoidance networks to find genes linked to fungal pathogenicity (57), but this study did not take into account the phylogenetic independence of divergent and convergent genes and therefore it is hard to discern if some results are linked to common ancestry or are actual events of convergent evolution.

Phylogenetics

As stated in the previous section, CoinFinder needs a phylogeny to be able to detect phylogenetically independent associated and disassociated genes in a given network. Since the fungal phylogeny is still under discussion (58), we decided that the best course of action would be to make a phylogeny of the species present in our dataset.

Phylogenetics is a field of systematic biology that aims to produce phylogenetic trees, which represent relationships between different biological entities according to their shared evolutionary history (59, 60). Phylogenetic trees have been used mainly in taxonomy but in recent years there have been appearing new applications like in epidemiology (61) or linguistics (62). In the case of this project the phylogenetic analysis was be used to infer the phylogenetic relationships of the fungi.

The phylogenetic trees are represented in phylograms or cladograms in which each biological entity is represented by a node. All the nodes in the tree are joined by branches, which split representing speciation events, forming a hierarchy of the taxa in the tree related by their shared evolutionary history. Nodes in the very tip of the branches are also named leaves or taxa, and usually represent species.

In the early stages of the phylogenetic field, phylogenetic trees were obtained using the only available data: morphological characteristics from the organisms. With the first sequenced DNA fragments, it was proposed to use changes in the DNA sequences as an objective comparison of phylogenetic relationships between species, giving birth to the field of molecular phylogenetics (63). The development of sequencing and the invention of PCR made it possible to obtain the genetic sequence of whole organisms, which were incorporated to molecular phylogenetic trees. In recent years, the appearance of next generation sequencing (NGS) techniques has produced a remarkable abundance of sequenced genomes of many organisms, which has further transformed the field of phylogenetics in to phylogenomics (64).

Before the appearance of NGS only a few chosen genes or proteins could be used for phylogenetic inference, which carried with it the disadvantage that sometimes the datasets did not have enough informative sites to obtain a robust phylogenetic tree. Additionally, the scarcity of data had the potential to introduce sampling bias. Due to the current availability of fully sequenced genomes, the number of genes and proteins that can be used for phylogenetic inference has increased accordingly. This abundance ameliorates the previously mentioned problems of phylogenetics, by providing large numbers of informative sites to infer robust phylogenies whilst also reducing the potential for sampling bias. Additionally, phylogenomics can be used for comparative genomic analysis to elucidate the function of genes and proteins.

With particular regard to the fungi, they have played a central role in the development of eukaryotic genetics and phylogenetics due to the use of *S. cerevisiae* as an eukaryotic model organism (65). Both the first sequenced nucleic acid sequence and the first

sequenced eukaryotic genome were achieved using S. Cerevisiae in 1965 and 1996 respectively (66, 67). Later, many of the early sequencing efforts were centred around fungi, due to their relative simplicity when compared with other eukaryotes (68–71). These early efforts provided enough data to perform molecular phylogenetic analysis of the fungi, as well as the use of comparative genetics to predict the function of many of the sequenced loci. Nowadays fungi are still one of the most sequenced groups of organisms with projects like the 1000 FGP, which is used in this study.

In order to obtain a phylogenetic tree for a set of taxa, the tree has to be inferred by identifying and quantifying changes in the DNA or amino-acid sequence of the taxa. To check for changes in sequences of several proteins or genes, the sequences first must be aligned to identify homologous nucleotide or amino-acid positions in different sequences. This process is called alignment, and by aligning sequences we are trying to compensate for the effects of sequence modifications (insertions, deletions, etc.) that can have happened to the same protein or gene in different species. Once an alignment is complete, we can systematically examine every possible tree topology for a set of taxa if the number of taxa is small. However, as the number of taxa increases the number of possible trees grows exponentially. An exploration of all the possible trees is not feasible with bigger datasets so searches of treespace for these datasets are carried out using heuristics. By using heuristics, we are not searching for the best tree amongst all possible trees, but instead we are scouting for the best tree or set of trees using an approach that reduces the size of the search space. For this purpose, several methods have been developed since the first uses of phylogenetics. All these methods use nucleotide or amino-acid substitution models of varying complexity that simulate the evolutionary process to generate and score phylogenetic trees based on how well the tree fits the given dataset, but the scoring and the process vary between methods.

Phylogenies can be inferred through several methods. The two most popular methods for phylogenetic inference are alignment based methods: super-matrices and super-trees. In super-matrices, several protein or gene sequences are aligned and concatenated together in what is called the super-matrix, and a phylogenetic search is carried on this matrix (65). The most widely used methods used in heuristic tree searches in molecular phylogenetics applied to super-matrices nowadays are maximum likelihood and Bayesian inference. In a maximum likelihood search the different phylogenetic tree parameters are modified randomly, and the tree is then assessed for its likelihood to fit our data in search of the most likely tree. Bayesian inference uses instead a Markov Chain Monte Carlo (MCMC)

process to randomly change tree parameters in search of a set of likely trees given our super-matrix alignment (posterior probability of the phylogeny), which can then be used to obtain a consensus phylogeny for our dataset. Super-matrices use direct evidence from the sites in the alignment and are especially useful the resolution of deep phylogenies. However, the need of these methods to use of sequences present in every organism in the dataset limits the number of genes and proteins that can be considered for the analysis and may induce some compositional bias.

Another approach to calculate phylogenies is super-trees (72). This method combines previously generated single gene or protein phylogenetic trees (gene trees) whose phylogeny is individually inferred using methods like the previously described maximum-likelihood or Bayesian inference. The gene trees are posteriorly combined into a super-tree, which takes the branch splits for every gene tree and combines them by using different methods, like parsimony or Bayesian inference. Since super-trees use individual Gene trees, which do not have to include all the taxa in the dataset, they are more flexible by allowing the usage of a higher number of candidate sequences for the analysis. However, super-trees are generally less likely to resolve deep phylogenies than super-matrices and they do not retain direct evidence from the alignment to support the tree topology.

Other less popular approaches to phylogenetic inference that do not rely on alignments to infer a phylogeny are Gene Content and Composition Vectors. The Gene Content approach assumes that taxonomically close species should have a similar composition of genes, thus having more shared genes than more distantly related species. Based on this assumption some approaches like a presence-absence gene matrix or a distance matrix based on the shared genes amongst all species in the dataset can be used to derive a phylogeny. Gene Content based method can use additional information present in the genomes that other sequence-based methods overlook, but they also ignore more detailed phylogenetic information present in the sequence. Furthermore, Gene Content assumes a simple process of evolution so some evolutionary events like the loss and posterior reacquisition of a gene in a species could induce errors.

Finally, Composition Vectors, which instead of using alignments, its focused on the particular characteristics that each species has in its sequence, and that phylogenetically close species would have similar characteristics. These characteristics are usually measured as the frequency of certain short DNA or peptide sequences, or the use of particular protein folds. These characteristics of each genome are measured in a composition vector, which can be then compared with other genomes by calculating the differences using a distance

matrix. The main advantage of this method is that by searching in the whole genome, removes any bias that could have been introduced by the selection of genes for the analysis, like it happens in super-matrix approaches, in addition of being fast even in large datasets. However, Composition Vectors are still susceptible to composition bias and need to be validated with previously inferred phylogenies, so its accuracy is low.

Alignment based phylogenetic methods are constantly being improved upon by using newer and more complex nucleotide and amino-acid substitution models that can simulate the evolution process more accurately. However, in some cases a single model is not enough to accurately form a phylogenetic hypothesis for an alignment. Phylogenetic hypotheses that only use one substitution model, called homogeneous models, are making the assumption that all the proteins or genes in the alignment evolve at similar rates and that these rates are the same across the phylogeny. In recent years, with the increasing availability of computer power, new methods that consider more than one substitution model are being used. Models that can account for more than one substitution model are called heterogeneous models. There are two types of heterogeneous models: data-heterogeneous and tree-heterogeneous. Data-heterogeneous models consider that different biological entities in an alignment (like genes, proteins or different regions of a gene or a protein) can evolve at different rates. Therefore, these models can assign a different substitution model to each biological entity in the alignment. Tree-heterogeneous models consider that the rate of evolution of the same biological entity can vary in different parts of the taxonomy of a biological group, and can assign several substitution models to the same biological entity to allow for rates to change across a phylogenetic tree. Currently, there are methods that combine both data and tree heterogeneous models like P4 (73). In the case of our fungal dataset, *Microsporidia* and *Cryptomycota* are notoriously hard to place in the fungal phylogenetic tree due to their nature as intracellular parasites, which has led to genome reductions, loss of their mitochondria and high nucleotide substitution rates when compared to the rest of the fungi (74–77). As high substitution rates can be problematic when doing phylogenetic reconstruction (78, 79), heterogeneous models were useful to place these groups in the fungal phylogeny with high confidence.

By using a data-heterogeneous and tree-heterogeneous model for the phylogenetic analysis of the fungi we aim to provide a robust phylogeny of the fungi that helps to resolve some of the most contested branches over the last years. Once we obtain a robust fungal phylogeny it will be used in conjunction to a fungal sequence similarity network to find sets of co-occurring and co-avoiding genes that are phylogenetically independent with the

intention of finding sets that are related to a particular ecotype. These coinciding and co-avoiding genes are potentially related by function to one another, which will potentially allow us to discover similar functional independent adaptations to a particular environment.

Plasmids

On account of the computational and time consuming nature of many of the steps in the analysis of the fungal dataset, mainly owing to its big size, we decided to test some of the methods that would be used in the project on a smaller dataset. For this purpose, we chose to analyse a plasmid dataset using networks. Plasmids are small DNA molecules that do not belong to the chromosome of any organism and can be found freely in the environment. Plasmids can be found in cells of all domains of life (80) and can be transferred to other cells, so they play a role in horizontal gene transfer (HGT) (81). Because they can be transferred to and from cells within the same generation (lateral gene transfer), plasmids are also important in the exchange of genetic information and can affect the chromosomal DNA evolution (82). Plasmids usually carry genes that provide a beneficial impact to their host organism fitness (83), and are frequently replicated by the host. Because plasmids can carry genes that are beneficial to their host organisms, events of co-evolution between the plasmid and the host can occur, where both the plasmid and the organism adapt to the functions provided by a plasmid (84). Plasmids are not able to replicate by themselves and need a host organism in order to replicate, which makes plasmid evolution different to cellular life evolution (85). Because plasmids can't replicate by themselves, they do not encode the molecular mechanisms necessary to do so. As a consequence, there are no gene families that are common across all plasmids, unlike cellular evolution where certain conserved gene families can be found in all known organisms (86, 87). Despite plasmids needing a host to replicate, they carry their own origins of replication (88).

Both plasmids and their hosts can form communities where there exists preferential sharing of genetic material (89). This communities are not linked to the physical proximity of a replicon but rather to characteristics of both plasmids and hosts, such as genome size, similarity of GC content, type of nutrition (heterotrophic or autotrophic) and oxygen tolerance (90). These communities can be rapidly changing since many plasmids possess

promoter regions that enable rapid evolution (91, 92). Additionally, the relationship between a plasmid and it hosts can change over time, due to adaptation and recombination between the hosts and the plasmids (93), where hosts can incorporate plasmid functions into their genomes and plasmids can acquire DNA segments from the host's chromosome (94–96).

Because conserved genes do not exist amongst the plasmids, it is impossible to use phylogenetic methods to discern the evolutionary relationships between plasmids. However, by using sequence similarity networks where we can relate plasmids without needing a common pool of genes present in all plasmids, we can discern the evolutionary dynamics that happen in the genetic material exchange between plasmids and their hosts, and what factors influence these dynamics. By using bipartite networks, we can link the functional information from the proteins with the taxonomy and ecological niche of each species to try to find what characteristics define communities of genetic material exchange.

ERGMs

As a way to further investigate which factors are affecting the interaction between plasmids and their hosts, we decided to use Exponential Random Graph Models (ERGMs) to simulate how the physical properties of amino-acid and nucleotide sequences affect the plasmid network (97, 98). ERGMs can indicate if an attribute of the nodes in a network has a significant effect on the formation of connections and quantify this effect.

First, ERGMs calculate the base probability of a node forming a link with any other node in the network. Then the models check how different attributes of the nodes modify the base probability of a link. Attributes of nodes in a network are represented in the models as "terms", which can interpret either continuous (numeric values) or discrete (e.g. functional categories) attributes. Furthermore, an attribute can be represented in the model by two different kinds of terms: one that measures variance of degree and another that measures homophily. For the first type of test, the covariance of degree between nodes is measured, this term being more positive the more likely nodes with high value in an attribute are to form a link in the network. The homophily terms measure how likely two nodes are to form a link with regard to how similar or dissimilar their attributes are. When an homophily term is negative, it indicates that nodes with similar values are more likely to form links and the opposite for positive values, nodes with dissimilar values are more

likely to form links. The model is calculated using a combination of terms for different attributes, and then a score is assigned to the model to check how well it represents the degree distribution of the original dataset. For this purpose, two different scoring methods are used, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). These scores both indicate a better model fit to the data the lower in value they are.

In summary, we aim to assess the of use networks to draw evolutionary relationships where traditional methods cannot be used, helping us identify the inherent evolutionary structure of the plasmids. Furthermore, by using ERGMs we can also identify the properties of the sequence of the plasmids, and not only their function, play a role in the benefit that plasmids provide to their hosts.

The plasmid section of the project was conducted as a collaboration between the author of the thesis and another member of the McInerney lab, Martin Rusilowicz, who appears as a co-author of the paper that resulted from this chapter in the results section. The sections that Martin carried out, the plasmid co-occurrence network and functional annotation of the plasmid communities, are only discussed in the corresponding paper in the results section. The parts that the author of the thesis performed will be the parts analysed in the discussion section of this thesis. Meanwhile, the fungal sections of the thesis, namely the fungal phylogenetic tree and the fungal co-occurrence networks, were carried out by the author of the thesis in their totality.

Results

Resulting from the analyses performed in this thesis several articles were written. As these articles were being written as the analyses were being finished, and because the follow the progression of the work performed in this project (method testing, phylogenetic analysis, co-occurrence networks) as discussed in the introduction, I decided that the alternative format was better suited for the presentation of this work. The resulting articles written in the duration of this project are presented in the following section.

The sequence sharing network underpinning plasmid diversity

Ignacio Riquelme-Medina[1*], Martin Rusilowicz[1*], Thomas Husja[2], Chris Knight[3], Philippe Lopez[2], Eric Bapteste[2], James O. McInerney[1,4,#]

[1.] Division of Evolution and Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester Academic Health Science Centre, Manchester M13 9PL, United Kingdom.

[2.] UMR 7205 ISYEB, Campus Jussieu, Bâtiment A, 4eme et. pièce 429, Sorbonne Université 75005, Paris, France

[3.] School of Natural Sciences, Faculty of Science and Engineering, The University of Manchester, Manchester M13 9PL, United Kingdom

[4.] School of Life Sciences, The University of Nottingham, Nottingham NG7 2UH, United Kingdom.

[#] To whom correspondence should be addressed.

Phone: +44 (0)161 2751843

Email: james.mcinerney@manchester.ac.uk; james.mcinerney@nottingham.ac.uk

Website: http://mcinerneylab.com/

Abstract

Plasmids, unlike cellular genomes, have no set of core genes common to all. Nonetheless, many genes are shared among plasmids in a complex network reflecting the co-evolution of plasmids and their hosts. In this study we sought to create a map of the gene sharing relationships between plasmid genes across a broad range of plasmids in order to characterise the gene sharing patterns in plasmids. We combined BLAST data with gene-ontological and taxonomic information to generate plasmid-gene networks and used multi and univariate analyses to investigate the underlying structure of these networks. Structured plasmid gene networks of a non-stochastic nature were identified, formed through taxonomic and functional relations of the underlying genes. These networks uncover some of the mechanisms underpinning plasmid construction and offer new insights into the dependencies of their underlying genes and functions.

Introduction

Plasmids are found in all three domains of life (1). They possess their own origins of replication, and usually multiply with the assistance of host mechanisms (2). In addition, they can be transferred to other cells, a form of horizontal gene transfer (HGT) (3). The mobile nature of plasmids makes them important agents of genomic evolution (4). Plasmids can carry genes with a selective advantage to the host (5); encoding, for example, mechanisms of resistance to toxins or antibiotics. These new genes can induce co-evolution between the plasmid and the host to adapt to the new functions, like reducing the cost of antibiotic resistance (6). The absence of an ability to replicate independently of a cellular host makes plasmid evolution different in many respects when compared with the evolution of cellular life (7). One significant difference between plasmids and host chromosomes is the absence of a core mechanism to drive replication, transcription and translation. A consequence of the absence of genes for core organismal functions like replication, transcription and translation is that there are no gene families that are universally distributed across all plasmids (8). This contrasts with cellular organisms, where several genes are universally distributed (9).

Genetic exchange communities exist for both plasmids and bacteria. Preferential sharing of genetic material is correlated with numerous factors, including genome size, similarity in GC content between the 'host' and 'donor' molecules, carbon utilization (heterotroph/autotroph), and oxygen tolerance, rather than physical proximity within a replicon (10). However, the plasmid landscape is dynamic and plasmids are known to contain certain promoter regions that facilitate rapid evolution (11, 12) and evolutionary pressures on both plasmid and host permit relationships to develop within observable time-frames between previously unfavourable pairings. Perhaps most importantly, plasmids carry a variety of factors that facilitate recombination with the host genome itself (13). With respect to antibiotic resistance, plasmids offering some benefit to the host can be lost due to incorporation into the chromosome (14). Conversely plasmids themselves present as a mosaic of DNA acquired from their different host lineages (15, 16).

Recently there have been calls for a more rigorous nomenclature for plasmids, on account of the lack of standardised annotations and naming conventions and also the absence of analogy to the taxonomic hierarchy applied to genomes (1, 17). Our objective here is not to create a new plasmid classification, but the complexity of the co-evolutionary relationships we are addressing underlies the historic challenges of plasmid classification. Firstly, the

classification of plasmids is not straight-forward and there are two principal reasons. Firstly, we cannot simply infer phylogenetic relationships from a set of universally distributed genes and use these phylogenies as a proxy for the evolutionary history of the plasmids themselves, because such genes do not exist. Secondly, we know that plasmids frequently exchange DNA with host chromosomes, and with other plasmids and viruses (18), so a single phylogenetic tree cannot capture the complexity of plasmid evolutionary history. Nonetheless, several classification approaches are used for plasmids (8, 19–21). One of the most important systems depends on the idea of incompatibility (Inc) groups, which partition the plasmids based on their origin of replication (22). Additionally, gene sets common to specific subgroups of incompatibility types have been found. For instance, shared functionality and synteny has been demonstrated between plasmids of the IncW group (17).

Here we take a novel approach to plasmid evolution by considering the totality of genes in every plasmid. We allow for HGT by not forcing a tree structure onto the data, where clearly a treelike structure does not exist. Instead, we use two different network structures in order to understand the distribution of genes among plasmids. We use a sequence similarity network (SSN), which have been used previously to detect relationships between proteins (23, 24), and we use a multipartite network structure that represents co-occurrence relationships (25). Although we do not have a universal set of genes common to all plasmids, we find that plasmids can form large networks, with identifiable communities, which do not fully map onto any existing classification system. Using aggregate statistics from the real networks, we model the factors that constrain and shaped the evolution of these networks. To this end, we investigate both the unipartite network of sequence similarity, the bipartite network of genes and plasmids and the multimodal network of sequence association, taxonomic dependency and ontological categorisation, depicted in Figure 5.

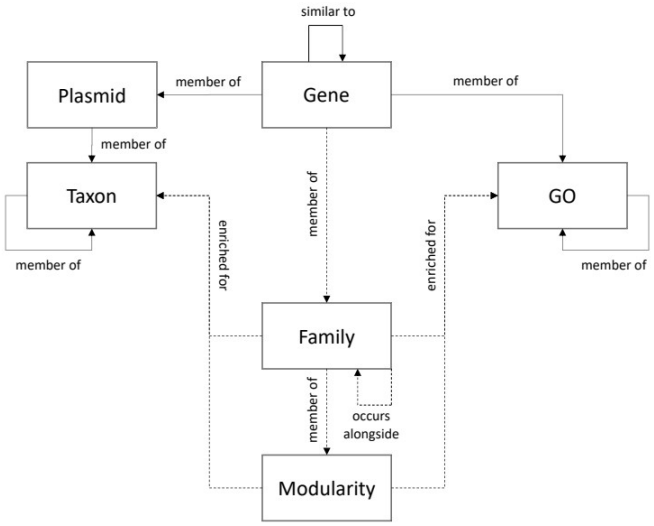

Figure 5: Metagraph showing the multimodal network of the plasmid genes. Boxes designate nodes and lines relationships. The dashed lines indicate some of the projected relationships, which can be entirely inferred from the other relationships in the graph. For example, a Family can be considered to be enriched for a particular GO category when a defined significant proportion of Genes which are members of that Family are also members of the GO category.

Network models

Plasmids consist of sets of genes and in that respect, they have some similarity with scientific papers that contain sets of authors (26), food recipes that contain sets of ingredients (27–29) and disease states that are influenced by distinct sets of genes (30–32). In each of these cases, we can use bipartite networks to explore the relationships between the groups. Bipartite networks are a network in which the edges explicitly connect elements between two groups X and Y, but not within the set X or the set Y (33). In the current study, we have used a bipartite network of gene families connected to their plasmids.

Analysing bipartite networks presents particular challenges. One approach is to decompose them into two different unipartite networks or unimodal projections. A unimodal projection of a bipartite network connects two nodes within a group if they were both linked to the same node in the other group in the bipartite network. Using the example from the last paragraph, we could obtain two different unipartite projections: one that will connect proteins that share a common pathway, and one that will connect pathways that share a

protein in common. In such unimodal projections community structure can be assessed by finding whether there are sets of highly-connected nodes within the overall network structure. Thus, clustering coefficients can be calculated as the ratio of the number of triangles (three connected nodes) in the network compared with all possible triangles in the network. Bipartite networks do not form these kinds of clusters since the minimum number of nodes in a bipartite network capable of forming an interconnected group of nodes is four for which a traditional clustering measurement cannot be used (34, 35). Although bipartite networks cannot form triangle-based clusters, they can form communities (34). To get around the issue of defining communities in bipartite networks, some other definitions of clustering have been made for bipartite networks, for example the ratio between squares (four connected nodes) and three paths in the network (35, 36). By constructing bipartite networks and their unimodal projections of plasmids and their genes, we will explore the connectivity between genes and plasmids in a number of ways.

Once a network has been constructed, we wish to understand the most important factors that have contributed to the structure we observe in our network. Of particular interest is the link between the attributes we observe in the plasmid dataset (say, the base composition of the genes in the plasmids) and the likelihood that these features were important in structuring the network, or whether the features are purely incidental. We have employed a type of logistic regression called an Exponential Random Graph Model (ERGM) (37, 38). ERGMs can quantify the significance of an attribute and its effect on the formation of edges in the network. ERGMs involve the calculation of the base probability of a node forming an edge with any other random node in the network. The presence of a feature, such as high GC content, might increase the probability of two particular nodes being connected while its absence would reduce this probability. ERGMs were used in conjunction with the metadata associated with each plasmid and its genes, in order to investigate the factors that affect plasmid evolution. The ERGM approach uses “terms” to fit the attributes into the model. The terms vary depending on whether the attribute is continuous or discrete. For each attribute, an ERGM can compute how it affects either degree (number of edges per node), or homophily (e.g. if big plasmids preferentially attach to, or do not attach to other big plasmids). The attributes used for the ERGMs in this study are plasmid size, protein length, protein isoelectric point (in order to check for acidic or basic environments), protein hydrophobicity (Kyte-Doolittle scale, in order to check for membrane proteins and environments) and the theoretical protein cost (i.e. expected energy required for synthesis of the constituent amino acids, to check whether very expensive

proteins are avoided in favour of cheaper ones, or whether expensive proteins avoid sharing the same plasmid).

The last approach we have used to explore the plasmid network was an analysis of gene co-occurrence. This refers to the case when homologs for two or more genes exist in the same genome at a frequency that is higher than we would expect by chance. Genes are found together on the same genome either because they are independently important to their host, or because they interact, and this interaction provides some measure of benefit. Recombination and HGT can, and indeed do, bring genes together in new combinations and we expect that in some cases natural selection will act to keep these genes together, if the effect on the host is beneficial. Computationally generated gene co-occurrence networks can be constructed directly from presence/absence data (39), by inferring co-incident (higher than chance) relationships (40) or by using correlation analysis to filter out indirect relationships (41).

Networks of gene co-occurrences provide a means to explore the patterns and groupings of gene interdependencies. Network level patterns relate to functional and taxonomic groupings, ecology and evolution, while individual gene-gene relationships can reveal specific mechanisms such as bacterial resistance or toxin-antitoxin systems (42, 43).

Methods

Data acquisition

The dataset used in this study was taken from NCBI and encompasses all plasmids for which the complete genome sequence is available (n = 4393, December 2014). The dataset comprises the amino acid sequences of all 338,930 proteins predicted to be encoded by these plasmids. Functional annotation metadata was obtained from Uniprot (44). Isoelectric point (pI), hydrophobicity (Kyte-Doolittle scale) and theoretical protein cost (calculated using the cost of amino acids in *Escherichia coli*) were estimated using the packages Seqinr and PeptideS (45, 46). The database of taxonomic classifications was acquired from NCBI and the GO hierarchy was downloaded from The Gene Ontology Consortium (47, 48).

Data processing

Sequence similarity was established using BLASTP v2.4.0 (49) using an e-value cut-off of 1×10-5. The BLAST output was formatted to leave only edge information and e-values for each entry. The resulting file was clustered using Markov Clustering Algorithm (MCL) version 14-137 (50) with an Inflation value (I) set to six. The clusters reflect substantial amounts of homology, though recombination means that proteins with homology (usually partial homology (51)) to proteins within one cluster can sometimes also belong to a different cluster. Recombination has made unambiguous protein cluster delineation impossible. For our purposes, these MCL clusters were used as a proxy for groups of evolutionary homologous proteins. We use the term “protein family” or simply “family” throughout the text to indicate these MCL clusters.

Bipartite Networks

Following the delineation of families using MCL, a bipartite network was built using plasmids and families as the two sets of nodes for both plasmid datasets. The edges were obtained using GenBank files in order to identify the families contained in each plasmid. In other words, an edge in the bipartite network corresponds to a protein family where a constituent protein is found on the plasmid to which it is linked. In addition, unipartite projections for both plasmids and families were derived from the networks. These

unipartite projections were only used to make ERGM models. Community structure within the bipartite network was ascertained using NetworkX (52) and a modularity measure that is based on the Louvain algorithm (53), which also detects communities within the network. Degree centrality and betweenness centrality were also calculated for every node on the network with NetworkX.

A custom script was made to check for taxonomic recall and precision measures of the communities. The script assigns the particular taxonomic rank (species, family, order…) for all plasmid hosts that are present in a module and compares them with the rest of the modules in the network. The program then reports sensitivity as the ratio between the frequency of a host species in the focal module compared with the presence of that species in the whole network and precision as the ratio between the focal host species and the total number of species present within a module.

ERGMs

ERGMs were built using the R package Statnet (38) version 2016.9 using the protein and plasmid metadata, and the modules and edges from the bipartite network. We built several models with different combinations of attributes in order to find the model that best fits the original network.

An attribute can have both a term for degree and another term for homophily. For degree terms a higher positive score indicates a higher chance of high degree nodes being connected by an edge while a negative value indicates a lower chance of high degree nodes being connected to each other by an edge. For homophily a higher positive Homophily score indicates that nodes with different values are more likely to form a link and a negative score means the same but for nodes with similar attributes. After the model is computed, the package uses the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) to measure how well the model fits the original data. The lower the values of these criteria, the better the model fits the data.

The final model was built using nine different terms, including terms such as degree effect of the molecular size, pI (Acid and Alkaline), theoretical protein cost (according to the cost of amino acids in *E. coli*), hydrophobicity and number of proteins in a family for the protein families, and plasmid size and GC content (High or low) for the plasmids. Because the bipartite network model failed to converge when homophily terms or functional categories were included the unipartite projections were used instead to explore homophily

in the network. The unipartite projection models were calculated using the terms described before for both the degree and the homophily terms. A bipartite model was used for further analysis in which homophily terms were excluded.

Coincidence network

Due to ERGMs not being able to model the functional properties of the families we explored the "coincidence" of different gene families on the same plasmids to check how the functional properties of the families shape plasmid interactions. We define coincidence as the observation that the presence of one gene family in a plasmid results in a significantly higher probability that another gene will be seen on the same plasmid. Once significance of this association is determined, we then construct a coincidence network where significantly cooccurring genes are represented as nodes linked by an edge. By creating a coincidence network, we were able to investigate the co-selection of the genes with one another.

The gene connected families themselves may be computationally inferred using the BLAST similarity network, which in our case were the isolated connected components in the network that had a size greater than one. Ideally any particular family will solely comprise homologues of a single gene, however the presence partial homologs means this is unlikely to be the case and complicates the qualitative definition of a gene or protein family. While these effects can be mitigated by raising the threshold over which genes are considered similar, this in turn results in more connected families each of which contains fewer genes, which reduces statistical power and impedes the actual detection of coincident pairs of interesting genes. For the coincidence network we therefore define a gene family simply as a connected component in the gene similarity network, where two genes are connected in this network when they share a pre-specified degree of sequence similarity according to the results of BLAST.

In the construction of the coincidence network, we seek particular pairs or groups of genes that occur together in plasmids at a rate higher than chance alone. We begin with a multi-partite network, with nodes representing the inferred families (F) connected to nodes representing the genes of which they are comprised (G):

$$F \xrightarrow{\text{contains}} G$$

Additional nodes are added to this network by including the NCBI taxonomic (T) and gene ontology (GO) classifications (O):

$$T \xrightarrow{contains} G$$

$$O \xrightarrow{contains} G$$

Given the taxonomic information now present in the network, each pair of gene families can thus be assigned a score based on their rate of coincidence within different taxa; read as the number of taxa in which genes of the two families are present in unison, against the number of taxa in which they are not. This score can be translated into a more accessible "p-value" via a binomial test, which will be outlined later.

Edges representing significant coincidences can be inferred from these scores, by drawing connections between families having a p-value of coincidence breaching a specified threshold, α. In our case α was set to a conservative limit of 0.01, modified further before being altered via the Bonferroni-correction to account for multiple testing.

$$F \xrightarrow{coincides\ with} F$$

Dropping the non-family nodes thus yields the unipartite coincidence network. This coincidence network was clustered into modules using the Louvain Method implemented in GEPHI (53)(54). The resolution of the method was set to 0.5 in order to provide a fine-grained modularity breakdown whilst still presenting classes large enough for meaningful statistical analysis. These modularities (M) can be included in the network itself:

$$M \xrightarrow{contains} F$$

Such modularities can be evaluated in terms of the annotations of the genes they contain via the network path:

$$M \xrightarrow{contains} F \xrightarrow{contains} G \xrightarrow{contained\ by} O \vee T$$

The same method as used earlier for gene family coincidence (i.e. Bonferroni correction and $P < 0.01$ threshold) can be used to identify ontologies and taxonomies that are linked to modularities at rates that are higher than chance. This gives the following edges representing significance:

$$M \xrightarrow{is} T \quad M \xrightarrow{is} O$$

The exact details of the determination of what constitutes significance are presented below.

Coincidence analysis

A binomial exact test determines if the number of times an observation to a Boolean question differs from what is to be expected. In the present case of cooccurrence we ask that, given gene families a and b, that first, how often would we expect both families to occur together in the same plasmid if co-occurrence was based on chance alone (the "rate", R). Second, we ask how many observations did we actually make (the "observations", O), and out of how many of those observations did we find both families together in the same plasmid (the "successes", S). Finally, given the null hypothesis that our successes are indeed based on chance alone, we are able to calculate a standard p-value of "rejecting" the null hypothesis.

More specifically, S is defined as the number of times gene families a and b were observed together in the same plasmid, O is defined as the total number of plasmids in the study and R is defined as the expected chance of overlap based on stochasticity alone, i.e.,

$$R=\frac{\mathrm{F_a}\,\mathrm{F_b}}{\mathrm{O}}$$

where C_i is the count of plasmids in which family i is observed.

Dissociation was also calculated, using a similar procedure, with the binomial test parameter, S being set as the number of times either a or b were observed in isolation, O as the total number of plasmids in which either were observed, and R as the expected rate of avoidance, i.e.

$$\frac{F_a\left(\mathrm{O}\text{-}F_b\right)+F_b(\mathrm{O}\text{-}F_a)}{\mathrm{O}}$$

Tests against ontological terms and taxonomy were carried out using the same procedure as for the association analysis.

Randomization analysis

In order to examine the behaviours of the ERGMs and the coincidence approaches, the analyses were re-run using random permutations of the input data. For this process, only the input data was randomised - the rest of the analyses remains the same.

100 random permutations of the plasmid network were generated by applying the Mersenne Twister randomisation method (55), which is implemented in python's random module, to the edges of the original network. This yields 100 random bipartite networks of

genes and plasmids. Note that only the edges of the input network were shuffled. This means the gene and plasmid identities themselves stay the same, the number of genes in each plasmid stays the same, and the random networks maintain the bipartite nature of the original.

For every random network derived from the families in addition to randomising the edge lists of the network the attributes of both the plasmids and the proteins families were randomised too. To accomplish this, each of the sequence of attributes was calculated for every protein (hydrophily, isoelectric point, molarity and cost) and its corresponding gene (GC content). Since the DNA sequences of plasmids are only available for the full plasmid length, we used the protein-coding DNA sequences of each gene present in the plasmid's GenBank files and then calculated the GC content for each protein coding sequence. Once these properties were generated, we proceeded to calculate the average of each attribute for each of the randomised plasmids and protein families. Finally, an ERGM was generated using the randomised edge list and its corresponding attributes, emulating in an automatic way the process that was described to obtain the metadata. Then the ERGMs were then calculated in the same way as the original ERGM.

With respect to the coincidence network, all 100 random networks were subjected to the same coincidence analysis as the original, detecting when genes from two different families tended to occur in the same plasmids at a rate higher than chance alone. The output of this process is 100 networks, with each network containing the family nodes with edges drawn between those families that tend to co-occur in the random plasmid networks.

Standard network analysis metrics (average degree, diameter, modularity) were drawn from these coincidence networks and compared with the same values calculated from the coincidence network of the true data.

Computation

The BLAST and MCL computations were performed using the Computational Shared Facility of the University of Manchester. The ERGMs and sequence analysis were done using the Statnet, SeqinR and PeptideS packages in R version 3.3. The ERGM automation scripts were written in PYTHON version 3.6. Coincidence analysis and network creation was performed using PYTHON, C++ and NEO4J. All code is open source and is available

on Bitbucket (bitbucket.org/mjr129/coinfinder). Network analysis was carried out in GEPHI v0.9.1 (56) and PYTHON. The networks were visualized using GEPHI.

Results

A total of 338,930 sequences were included in the analysis, spanning 4,393 plasmids associated with 1,646 taxa. BLAST analysis resulted in a sequence similarity network comprising 71,946,457 edges. After using MCL on the similarity network 78,012 protein families were obtained and a bipartite network was constructed connecting protein families to plasmids that share at least one sequence in common. In total, the network was composed of 82,405 nodes and 297,500 edges from which a total of 83 communities were derived represented as different colours in Figure 6. These modules indicate sets of plasmids that share more similar repertoires of protein-coding genes. We obtained a modularity score of 0.69. This score can range between -1 and 1, negative values indicate less modular networks and positive more modular networks (53, 57, 58), which implies that our network is well divided into communities.

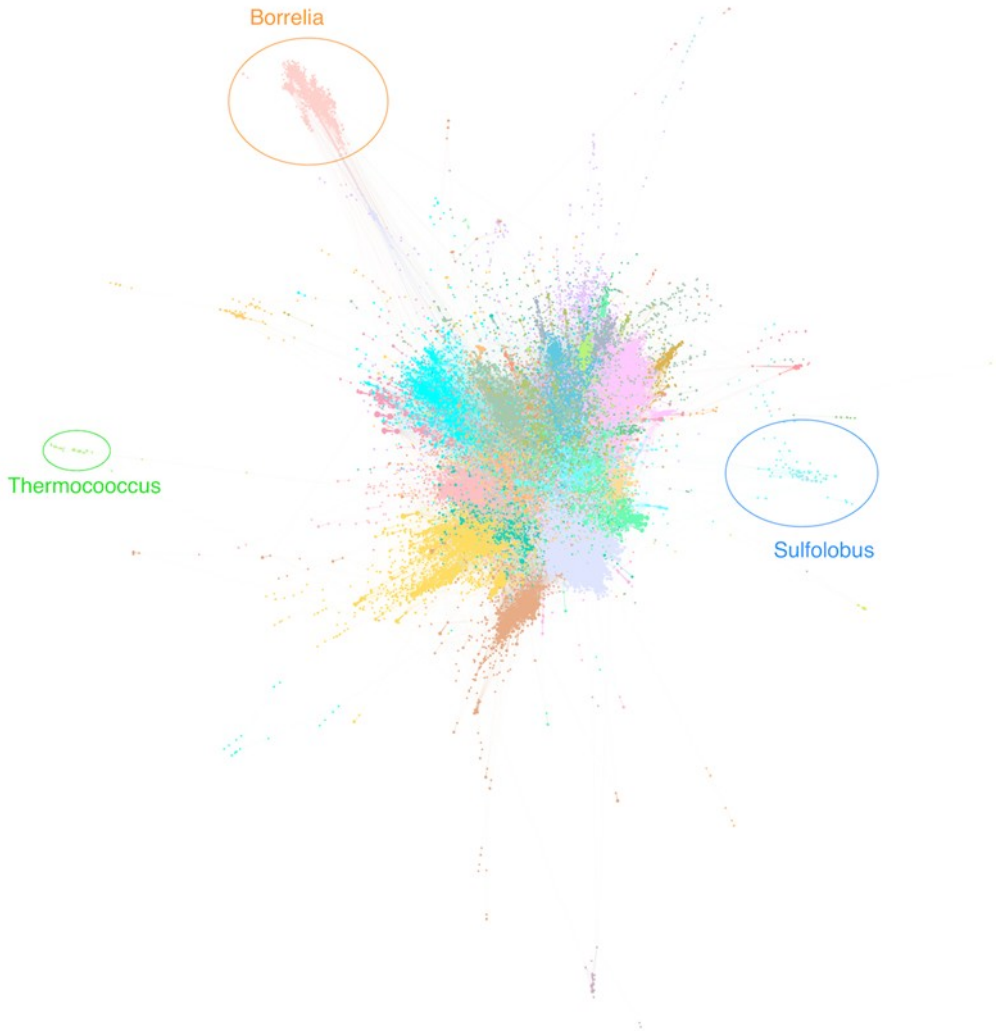

Figure 6: Visualization of the Plasmid Bipartite network. In this figure a node can be either a plasmid or a protein. The colors represent different communities. The most relevant taxonomic communities are highlighted.

ERGMs

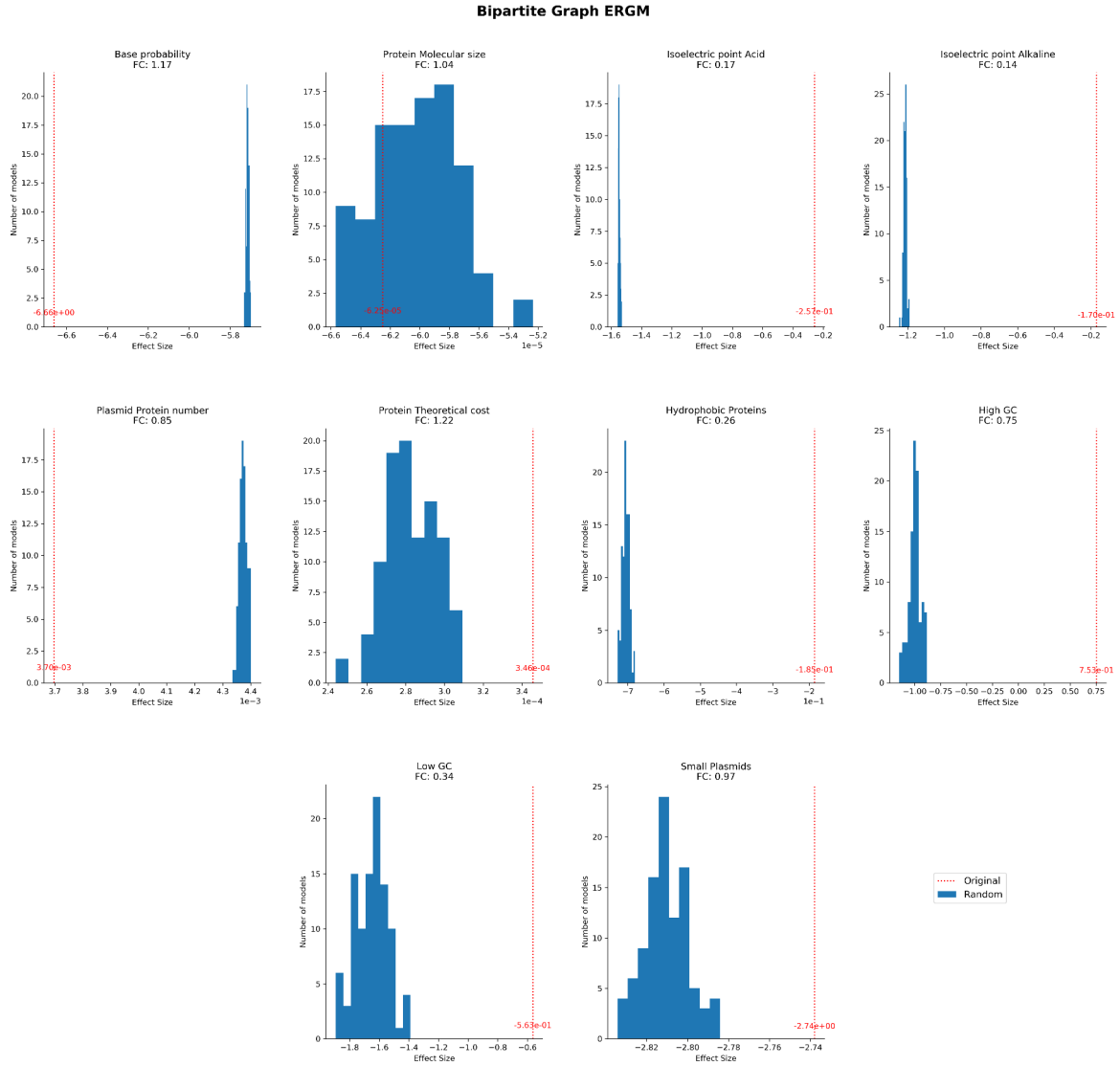

Figure 7: Random ERGM Bipartite comparison plots. Each graph represents a different term, where the 100 random ERGM term values distribution is represented as a blue histogram and the original data value for that term is represented by a red dotted line. FC is the fold change between the original data and the average of the randomised values.

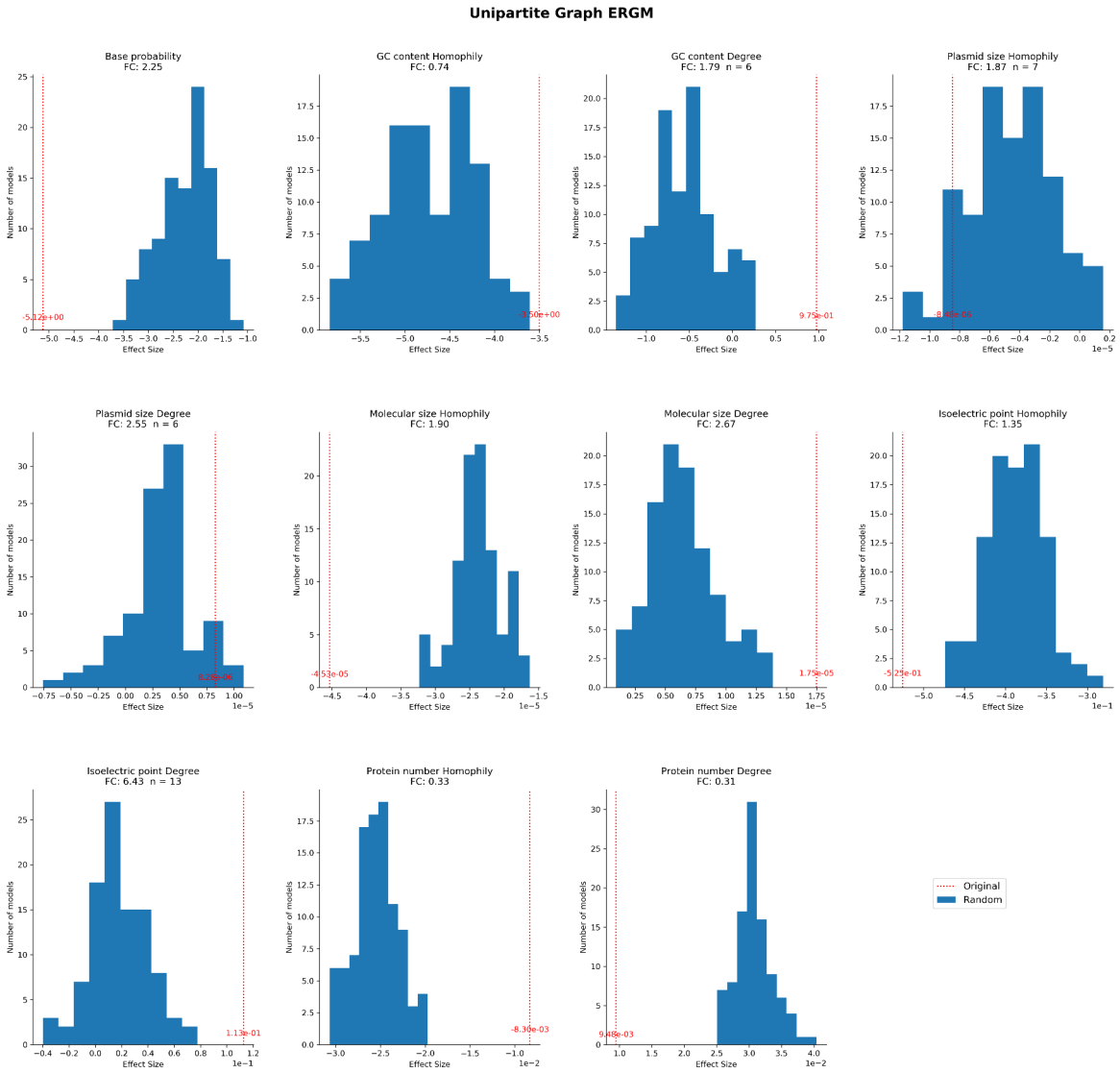

Figure 8: Random ERGM Unipartite Plasmid comparison plots. Each graph represents a different term, where 100 random ERGM term values distribution is represented as a blue histogram and the original data value for that term is represented by a red dotted line. FC is the fold change between the original data and the average of the randomised values. When a term failed to reject the null hypothesis that it affects the network in some of the runs it is marked by n = x in the title of the plot, where x is the number of runs in which the null hypothesis was not rejected.

ERGM models were calculated for this dataset for the bipartite network and for both unipartite projections (protein families and plasmids) due to the bipartite network model failing to converge when including nominal homophily measurements. The ERGM results for the bipartite network can be seen in Figure 7. and the results for the plasmid unipartite projection are in Figure 8. All attributes used in the models are significant – meaning that inclusion of these terms helps to explain the topology of the original network in the model. In general, the models returned several expected results (such as bigger plasmids being associated with higher connectivity) and other interesting features such as high GC content plasmids linking significantly more often to each other that would be predicted.

The ERGM model (which included the GC content, protein number, protein molecular size, protein cost, pI (Acid and Alkaline), hydrophobicity and plasmid size attributes for degree) of the bipartite network showed a substantial decrease in the AIC and BIC measures when compared to the base network (Base: AIC: 4,789,027; BIC: 4,789,045; Best model: AIC: 4,367,825; BIC: 4,368,001). However, the model that most improved when compared with the base network was the plasmid unipartite projection that included attributes GC content, protein number, protein molecular size, isoelectric point, hydrophobicity, plasmid size for both degree and homophily (Base: AIC: 8,771,081; BIC: 8,771,095; Best model: AIC: 6,635,845; BIC: 6,635,999). On the other hand, the model of the protein unipartite projection is less useful for our project due to the high number of edges present in the network maing the model very computationally intensive to converge and the final AIC and BIC scores being too large (the scores were essentially infinite) to allow us to assess the quality of the model. For these reasons we focused only in the bipartite network and the plasmid unipartite projection models.

The model fits to the original data degree distribution for the bipartite network and the unipartite network can be seen in Figure 9. For the bipartite network the model differs from the original data in that the original data has fewer low degree nodes and more medium degree range nodes which the model does not account for. As for the plasmid unipartite projection the model also fits the original data degree distribution well, however in this case the medium degree values are overestimated, and the low degree underestimated. In this case the model is assuming a lesser number of isolated plasmids. This may be a reflection of the lack of medium degree nodes in the bipartite model since plasmids with low connections would fall in this degree range (most plasmids in the original network are in the 3-100 degree range), having several protein families in the same plasmid, so outside of the low degree range, but poorly connected to other plasmids. This means that there are

other properties apart from the ones considered in the model that influence the distribution of medium range degree in the network. This could be related to other physical properties we didn't take in account or to protein function since the model doesn't consider it.

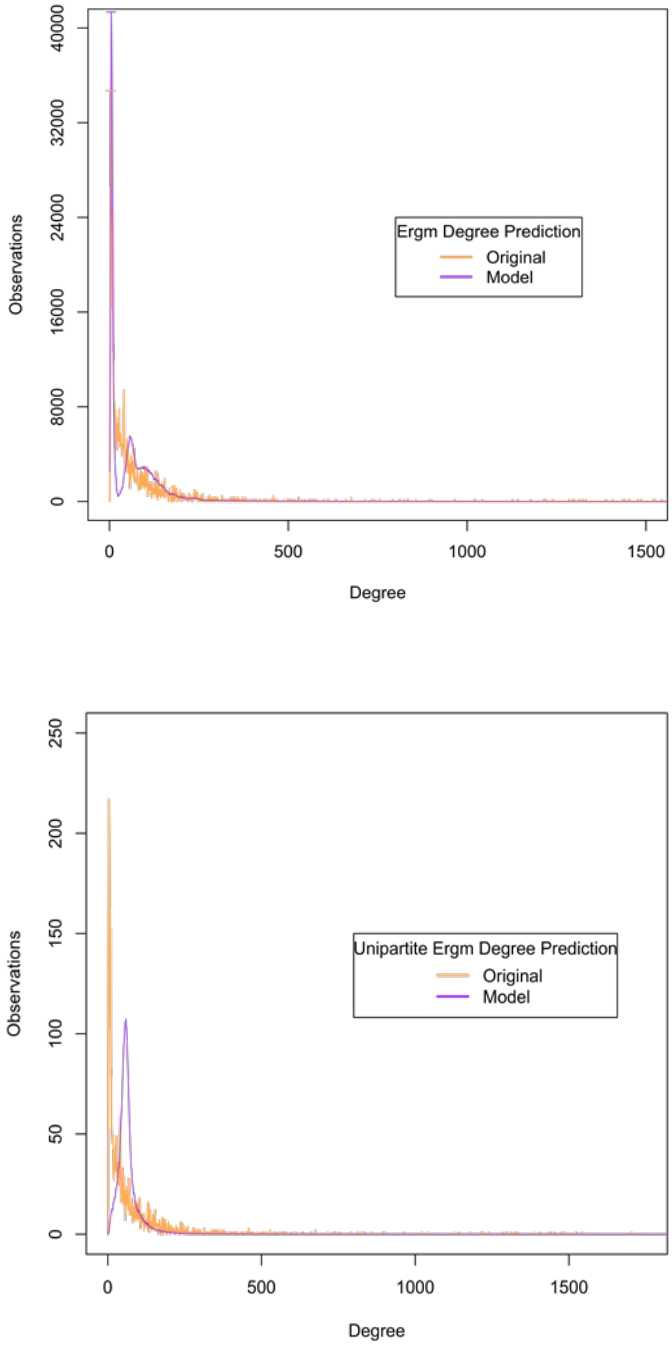

Figure 9: Degree distribution of the Original data (orange) and the ERGM (purple). The x axis represents the degree of the node and the y axis the number of nodes with that degree. Top figure represents the bipartite network model fit and the bottom figure represents the unipartite plasmid network model fit

Association Analysis

In order to carry out an analysis of the significance of association between different kinds of genes appearing on the same plasmid, we found that the protein clusters provided by MCL were not large enough to identify associations. Therefore, when determining the significance of association between different kinds of genes, we define the gene families as connected components. Our input data consisted of 59,948 families. The largest family comprised 147,425 sequences, which is 43% of the total number of sequences. A total of 41,356 sequences (12% of the total) remained isolated and not connected to any homologous sequence. These isolated sequences were removed from further analyses. This left 18,592 connected families distributed across 4,375 plasmids in 1,636 host organisms. 297,574 sequences (88% of the total) are contained within a family, with 4,375 (99.6%) plasmids and 1,636 (99.4%) taxa possessed at least one sequence in a family.

Using the same similarity network that was used for the ERGM analysis, a new bipartite network was constructed. Families were connected to plasmids that share at least one sequence in common. Coincidence analysis of this network, as outlined in the methods section, provided a unipartite "association" network, identifying families that coexist in similar groups of taxa – i.e. those families present in unison at a rate greater than expected by chance. A second unipartite “disassociation” network was also constructed, identifying families that tended to avoid each other at a rate greater than expected by chance.

The association network identified 8,910 families having significant associations with another family, representing 48% of the total number of families, whilst the disassociation network identified 2,990 families presenting significant avoidance to another family, representing 16% of the total number of families.

The Louvain algorithm decomposed the coincidence network into 229 modules, where each module typically represents a set of gene families highly dependent upon one another. These modules are explored in detail in the discussion.

Modules were annotated using the GO and taxonomic annotations of the constituent genes, where a module was assigned a specific annotation if that annotation occurred within the module at a rate significantly greater than expected by chance. The resulting network and a sample of the annotations placed upon the modules is depicted in Figure 10.

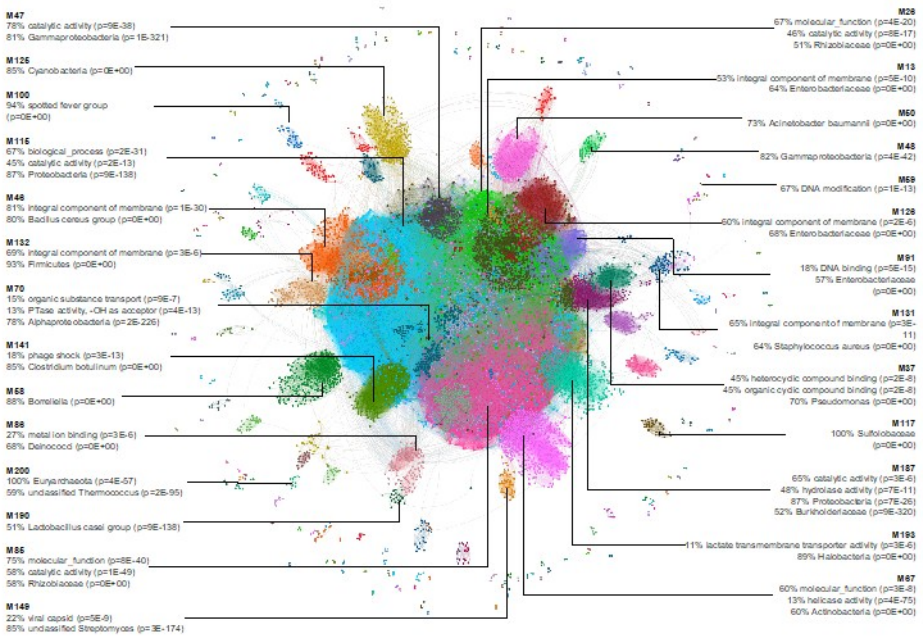

Figure 10: Unipartite association network. Nodes represent gene families and edges represent significant associations between those families. Nodes and edges are colored by module, where each module is assigned a unique color. The labels show a sample of the significant GO-class and taxonomic annotations placed upon the modules.Percentages indicate the proportion of genes with the specified annotation.

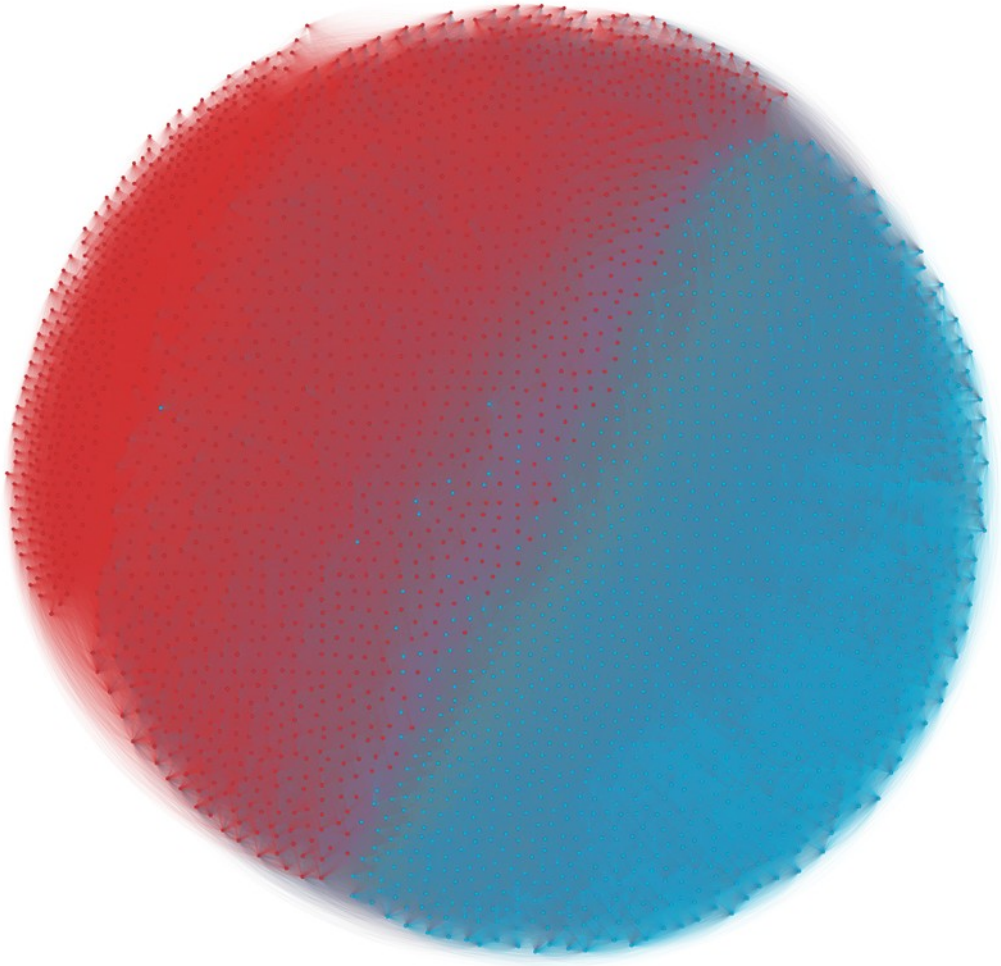

Figure 11: Unipartite disassociation network. Nodes represent gene families and edges represent significant avoidances between those families. Nodes and edges are coloured by module, where each module is assigned a unique colour. The size of each node is proportional to the number of genes within the family.

Randomization analysis

The effect size of the randomised ERGM terms and a comparison with the original ERGMs can be seen in Figure 7 for the bipartite network and in Figure 8 for the unipartite plasmid network. We expected some of the random terms to remain significant and similar to their original counterparts, especially those related to the size of the plasmids since size related properties were not affected by the randomization of the network. However, all the terms in the randomised ERGMs were significant in most of the randomisations, with similar values compared to the original ERGMs, which means that we cannot trust that the original ERGMs terms are really affecting the network. Despite this there are been changes in the signs (positive to negative or vice versa) and effect sizes of some terms which could mean that those terms are affecting the original network because it is not random.

In the bipartite random network, the term that has changed sign is the high GC content, which is positive in the original ERGM and negative in the randomised ERGM. This means that in the original network high GC content nodes tend to link more while in the randomised network the nodes with average GC are the ones that have a higher chance to link since both high and low GC have a negative impact in the chances of forming a link. In addition, some terms have the same sign between the original ERGM and the randomised ERGM, but the effect value varies between the two reflected by the Fold Change (FC) that can be seen in Figure 7. In this case all the terms that change meaningfully (FC < 0.5 or FC > 2) have < 1 FC, meaning that their effect size is smaller in the original ERGM when compared to the randomised ERGM. These terms are the isoelectric points (Acid FC: 0.17; Alkaline FC: 0.14), hydrophobicity (FC: 0.26) and low GC (FC: 0.34). All these terms have negative effect sizes, which means that the higher the attribute's value is, the lower the chances to form a link. Because all of them are less negative in the original network compared to the random one, it means that in the original network nodes with high values of these attributes have more chance to form a link than nodes with similar values in the random network, even if the effect size in the original network is still negative. Because of this it seems that nodes with extreme values in their attributes are much less prone to link by random chance, favouring nodes with average attributes in the random network, while in the original network this effect is less severe so nodes with extreme values have and increased chance to form links compared to randomness. In other words, in our original network genes with extreme values in these attributes, which are related to extremophily, are much less punished to form edges when

compared to randomness, or even have an increased chance to be shared in the case of high GC.

As for the unipartite plasmid network the terms are closer to the randomised network than in the bipartite. The random distributions for the terms GC content and isoelectric point degree are centred around 0, with GC content being more on the negative side. Because of this both terms have a mostly neutral effect in the random ERGM (more so that in both of these terms some of the models weren't affected by these terms), leaning towards a negative effect in the random ERGM GC content. In contrast, in our original network plasmids with high values in these terms are more likely to share genes than by random chance. Since GC content and Isoelectric point could be linked to extremophily or at least adaptation to certain harsh environments it would make sense that genes that have these properties that help organisms adapt to their environment are more easily shared. Another thing to note is that the base probability in the original network is much more negative than in the random distribution (FC: 2.25), and since ERGM base probability is a reflection of network density (the more positive, the denser) it means that our original network is much sparser than it would be randomly. This would mean that in our original network plasmids are much more selective with what other plasmids they share DNA, although this can be explained by plasmids that are not able to share DNA simply because they do not appear in the same environments. The rest of the terms in the unipartite networks are closer to the random distribution, but there are some properties we would like to point out. The molecular size homophily (FC: 1.9) and isoelectric point homophily (FC: 1.25) are more negative in the original network than the random distribution, meaning that in there is a stronger preference for nodes with similar isoelectric point and molecular size to share between themselves than we would expect randomly. As stated before, Isoelectric point could be linked to adaptation to some environments like acidophily so it would make sense that plasmids that appear in these environments share more between themselves. The molecular size would mean that there is a preference to share plasmids that make proteins of similar sizes, and by looking at the molecular size degree we know that there is a slight preference to share bigger proteins than in a random network. Finally, the effects of the protein number degree (FC: 0.31) and homophily (FC: 0.33) are high in the random network compared with the original one, meaning that one of the main drives to having a higher degree in the random network is simply to have more proteins. However, this effect is much smaller in the original one, so even if having more genes is beneficial to share more with other plasmids, it is not as important as we would expect, meaning that the other terms would play bigger role in shaping the connections than in a random network.

The results of the random data being significant mean that we cannot discard that some of the results from our original ERGM are significant just because of the size of our dataset. However, the differences between the normal and random ERGMs, particularly in the terms that have changed signs, lead us to believe that the original network does have an underlying structure that is explained by some of its properties.

As for the statistics calculated for the coincidence analysis of random plasmid networks, along with the results of the analysis of the true data, are presented in Figure 12.

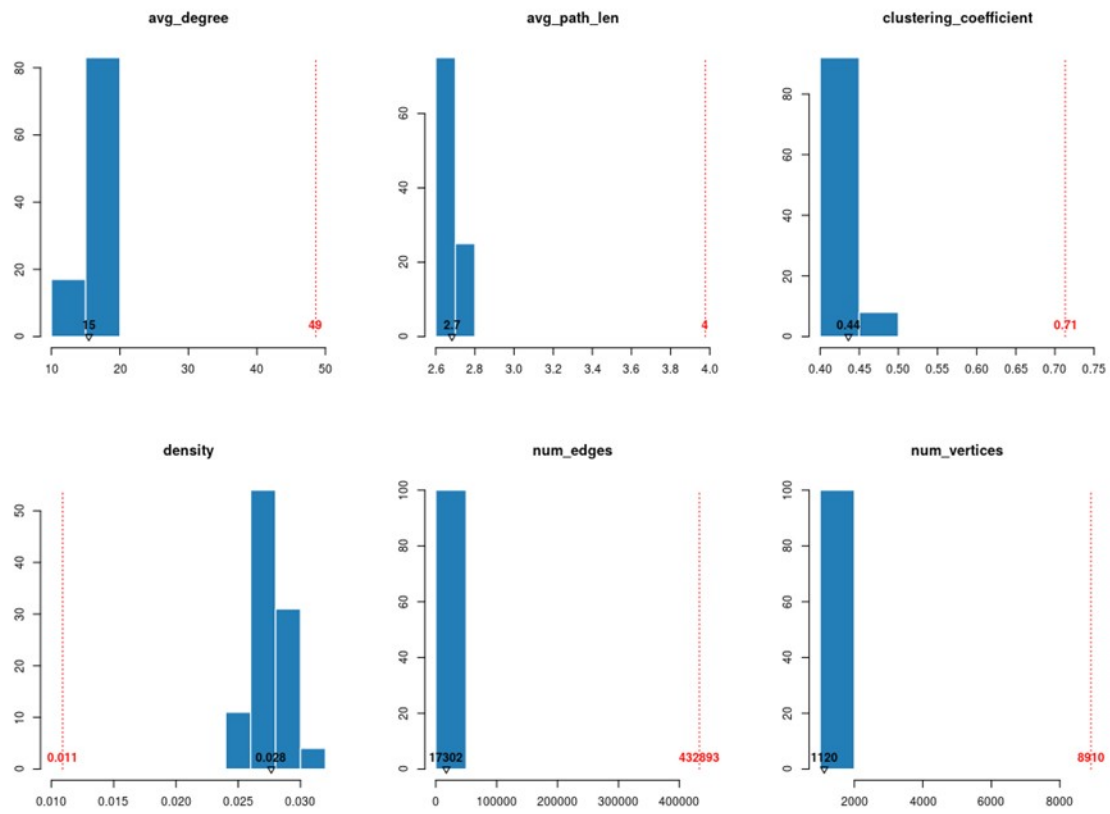

Figure 12: Histograms showing statistics calculated for the coincidence analysis of 100 random plasmid networks, versus the same value for the analysis of the true data. The blue histograms represent the analysis of the random data, with the mean value indicated by a triangle. The dotted line indicates the value for the same statistic for the analysis of the original data

Our coincidence analysis method detects families that co-occur at a rate higher than chance alone. For our random networks we would however expect some gene families to randomly co-occur. As presented in Figure 12, in all analyses of the 100 random networks far fewer coincident gene families were detected. On average, the analyses of random networks yielded 4.0% of the number of coincident families (edges) than the analysis of the true network.

Discussion

Networks

As a result of our analyses we constructed three networks. First a sequence similarity network to be used for testing how physical properties affect the plasmid network. Secondly an association network, which no longer looks at the statistics of plasmid network structure, rather it is an analysis of the identities of the genes themselves and whether particular kinds of genes tend to appear together. Lastly, we constructed an avoidance network, which analyses whether genes tend to avoid one another. Since both the ERGM network and the association network are based on similarity measures amongst the plasmids we have found them to be virtually the same, so we will discuss them together.

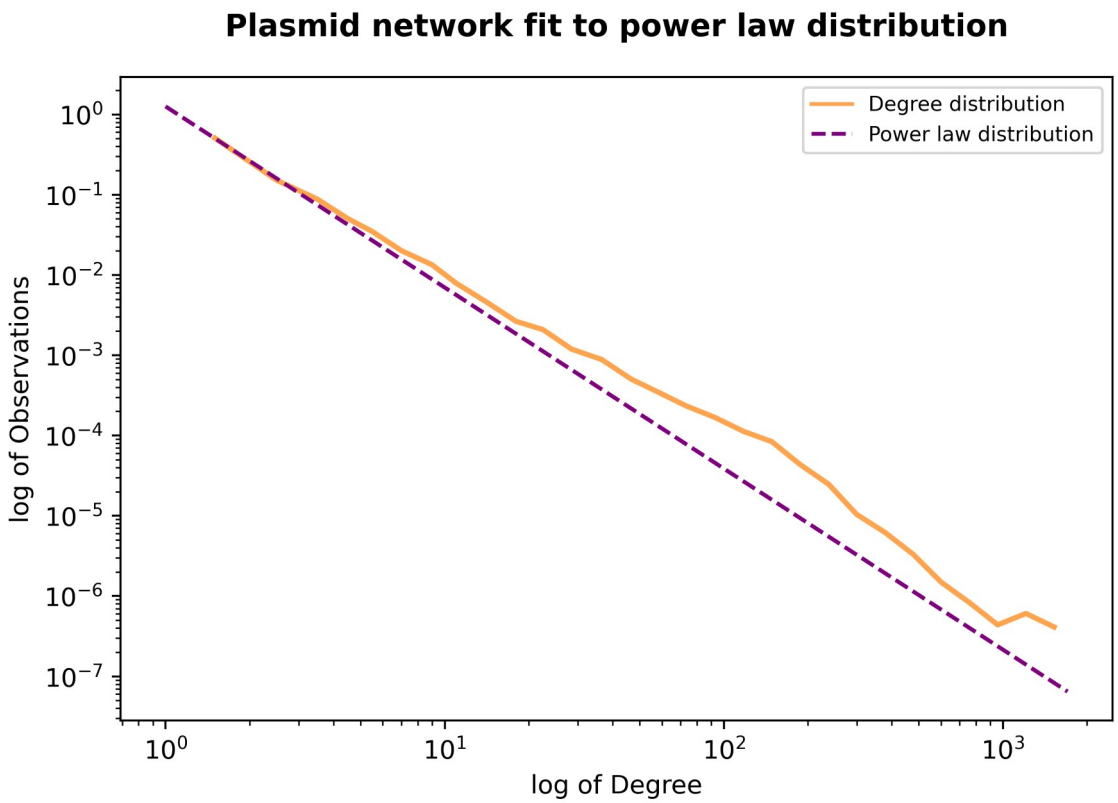

Figure 13: Log degree distribution of the plasmid bipartite network compared to a power law distribution (scale-free). The x axis represents the log degree and the y axis represents the log number of nodes. The degree distribution approximates a scale-free distribution, where there are many nodes with low degree (left of the plot) and fewer nodes with high degree (right side of the plot).

The similarity networks, shown in Figure 6 and Figure 10, display a typical scale-free biological network, with a giant connected component (GCC) surrounded by a number of small islands. We further confirmed that our networks are scale-free by comparing them to a power law distribution as can be seen in Figure 13, with a γ parameter of 2.255 (scale - free networks usually have a γ parameter within $2 < \gamma < 3$). Moreover, some of the highest degree nodes belong to the genus *Rhizobium* simply because this taxon has very large plasmids. The underlying organisation of the network is visible as an arrangement of modules with both taxonomic and functional mapping. For instance, of two of the largest modules by family count (M13, #1, 840 families, and M46, #4, 520 families), module M13 exists mostly within plasmids associated with Proteobacteria (94% of the total), while module M46 comprises the genus *Terrabacteria group* (90% of the total). Whilst both modules show enrichment of membrane part proteins (integral component of membrane: 53% and 80% respectively), module M67 (#2, 663 families) shows a significant number of genes with ion transport annotations (48%). This network can be considered as a map of the plasmid-borne genes showing communities of interdependent genes, where dependency can be characterised (given our data) by either functional or taxonomic dependency.

Within the map, known information about the plasmid world is accurately reflected, for instance module M58 depicts a set of gene families found almost exclusively (98%) in Borreliaceae, with 99% of Borreliaceae genes occurring in this module. This finding is in line with our current understanding of the atypical nature of *Borrelia* plasmids in comparison with other genera (59, 60).

Given genomic annotations are a human concept, as is the choice of sequenced genes, it is not unreasonable to assume that certain network structure may reflect human research interests. Module M86 is composed in large part (68%) of members of the genus *Deinococcus* and in terms of function, 27% of the genes are related to metal binding. The overt presence of this group may be natural, or may reflect human research into Deinococci for the purposes of industrial metal degradation or the basis of radiation resistance (61–63). Module M193 is principally *Halobacteria* (67%) yet the strongest GO annotation is *lactate transmembrane transporter activity* (11%).

When we calculated the betweenness centrality measure for every node in the network, we observed ten family nodes whose most common function annotation is related to DNA (especially DNA binding, two involved with DNA integration and one specifically involved in plasmid replication) are among those with the highest values. It is not surprising that proteins involved with plasmid replication occupy such an important

position in this network, since this is such a central function of plasmids. In cellular lifeforms, we see that the kinds of genes that are most consistently present in genomes relate to replication, conjugation and other genes related to plasmid survivability(64–66). It is perhaps not surprising that this theme is continued among plasmids, where functions that maintain and replicate plasmids have a wider distribution than other kinds of genes. Betweenness centrality is high for nodes that connect communities that might not otherwise have many connecting nodes: a high betweenness centrality score will be found for proteins that are important in many different contexts. This means that even if these proteins are not intrinsic to any community, they play a role in many different communities with differing functions and properties. We can say that high betweenness centrality protein functions therefore have more general beneficial effect under many circumstances that is well distributed in the network. In this case our results are indicating that even specialised communities need the function of DNA binding and replication in order to properly carry on with their function. Another interesting function with high betweenness centrality is related to detection of visible light so it seems this function well distributed in the network.

Network structure

Density

Where 100% represents a fully connected network, with every node connected to every other node, the density of our association network is low, at only 1.1%. Whilst there are a number of strongly connected nodes (families), even in the GCC there is no node that shares a link with all others. The average node degree (97.17) however, is substantial. There are 170 isolated connected components, and the modularity analysis further divided the network into 228 modules, 59 of which are in the GCC. The network has a relatively high total modularity of 0.67, indicating a modular network nature. Rather than a single “core-plasmid” gene-set it is therefore more reasonable to state that there are many gene-sets, all very loosely connected, but nonetheless showing strong intra-group dependencies. This is also reflected by the network diameter (14), which can be observed in the several long “tails”, such as M125, seen coming off the GCC in Figure 10. This figure is typical of biological "small world” networks. The archetypical small word network, the internet, by contrast has a diameter of around 19 (67).

Robustness

The removal of the small and large families did not substantially affect the organisation of the network, suggesting that the size of the family does not play a major role in network structure. While removing high-degree nodes does break the network up into smaller modules, repeating the modularity analysis shows that the new modules do overlap with the original set of modules, suggesting that, despite the low density, a few common nodes do not solely hold the network together. What we observe therefore, can be considered to be communities of interdependent gene families, rather than sets of families bound by few central nodes. Like many biological processes therefore, this is suggestive of high redundancy, with many gene families being replaceable when considering paths through the coincidence network.

The Borrelia peninsula

The earlier noted module M57 is primarily comprised of families and genes sourced from the Borreliaceae genome. A few other taxa (<2%) are contained within this module, namely 0.4% Helicobacteraceae, but their low incidence makes extrapolation of meaning difficult. However Borreliaceae and Helicobacteraceae can share habitats in some occasions (68), which may lead to sharing of genetic information or convergent evolution. The functionality of the module is enriched for “chromosome organisation” (9%), in comparison with the rest of the network, but no one functional class dominates the module. It is likely that the genes therefore represent a private Borrelia set of wide-ranging functionalities, rather than a specific Borrelian use-case.

Island isolates

Outside the GCC (C0) the next largest connected component is C81/M117. This encompasses just 31 families with 148 genes. The module itself primarily Sulfolobus (98%), and notable annotations include DNA integration (11%) and cytokinesis (6.8%). There seem to be a number of similar modules related to extremophily or thermophily like M200, which is related to Thermococcus and Methanocaldococcus. Extremophily may then be an important factor in structuring part of the global plasmid network, probably due to this communities’ hosts inhabiting habitats isolated from the rest of the plasmid’s hosts.

Dissociation analysis

The dissociation network, shown in Figure 11, is characterised by a single GCC comprised of 2 modules. Like the association network, the modules show some significant enrichments, however, they tend to be of mixed character. The network itself is more complete than the association network, with a high density (0.68), high average degree (1,933), short diameter (2) and very low modularity (0.06). If indeed a gene shows dissociation with another gene, it is probable that that gene shows further dissociation with others. These observations likely reflect the non-symmetrical nature of the network: if A associates with B, and B associates with C, then it stands to reason that A and C may also present an association. The same is not true in our dissociation case. It may therefore be of more use to scrutinise individual dissociative relationships, rather than infer relationships in network form.

Within these networks, the network statistics do not suggest that patterns have been identified where there are none: the clustering coefficients reveal the networks sourced from random data are generally disorganized, with no clear modules as for the analysis of actual data. The random data networks are in contrast, much more tightly packed, having higher density but lower average degree and path length, suggesting that what we do see is due to few, randomly well-connected nodes rather than an overreaching network of connections as is present within the actual data.

ERGMs

Once we finished investigating how plasmid connectivity and gene function impact the structure of the network, we proceeded to analyse what features of our dataset contributed most to the structure of the network by using ERGMs. In order to evaluate whether ERGMs had the ability identify important features in our dataset, we set plasmid size as a positive control. It is axiomatic that if a plasmid has more proteins, then it will form more connections in our networks, therefore including or excluding plasmid size in our ERGMs should have a substantial effect on the model. The resulting ERGM model showed that plasmid size has indeed a substantial effect on the network structure (for example the effect size for small plasmids in the bipartite model is -2.738, which indicates that the log odds of a small plasmids having an edge are almost 3 times lower than the baseline log odds of forming an edge. Converted to probability, this means that the baseline chance of forming

an edge is 1.30e-03 and the chance of forming an edge for a small plasmid is 8.87e-05. This is the biggest effect of an attribute in the model). The model, quite tautologically, shows that if a plasmid is large, then this substantially affects the chances of a node forming more edges.

Having established that ERGMs can identify features of importance in our network, we evaluated another eight features to see how likely they were to have influenced the network structure. A feature with high impact in our simulations is GC content. In Figure 7 we can see that plasmids with high GC content (>60% GC) have a higher chance of forming a link of any kind (0.753) and plasmids with low GC values (<40% GC) have a lower likelihood of forming a link of any kind (-0.563). Since the correlation between base composition and likelihood of forming a link is the highest amongst all the attributes, it seems to indicate that high GC plasmids interact significantly more with one another, than with other plasmids. The unipartite projection of the plasmids provides more information as seen in Figure 8. This ERGM shows that plasmids with similar GC values tend to link together considerably more than plasmids with differing CG content (-3.5). The ERGM also confirms the trend of a higher chance of forming an edge for high GC plasmids (0.975). An explanation for this phenomenon might be that having high GC content is a strategy for extremophiles (69). High GC content in DNA can result from selective pressure favoring the presence of certain amino acids, which provide more stability to proteins (70). These amino acids tend to be encoded by G or C in the 3rd position of a codon, a mechanism that some thermophilic organisms also use (71–73). High GC content can also be seen in the plasmids of some thermophiles such as the *T. thermophilus* plasmid pTT27 (74). Therefore, the plasmids of extremophiles that live in the same environments have more chance of sharing genes, and a decreased chance of sharing with low GC plasmids from non-thermophilic species. We note that not all thermophiles follow this pattern of having high GC content, and instead, have low GC content (75, 76). Other non-extremophile plasmids, whose host organisms have a selective pressure towards these stable amino acids can also show high GC content, like *Amycolatopsis* or some *Gordonia* species in our dataset.

The isoelectric point of a plasmid has a smaller effect on the network structure, compared with GC content or plasmid size, though this effect is still significant. In general, we observed in the bipartite ERGM that proteins with a neutral isoelectric point form more edges on average than alkaline or acidic proteins (since both acid and alkaline proteins have decreased chance to form edges (Acid: -2.566e-1, Alkaline: -1.698e-1)). In the unipartite projection we find that plasmids with similar average isoelectric point proteins

have a higher chance of forming links between themselves than plasmids with differing average isoelectric point (-5.253e-1). The tendency to share more with other organisms that have proteins with similar isoelectric point could also be due to extremophile bacteria, particularly acidophiles and alkaliphiles. Since the pH of the environment has a big effect on plasmid activity and stability (77) it is likely that plasmids living in extreme pH conditions would share proteins that are adapted to function in these conditions. Unlike thermophiles, acidophiles and alkaliphiles do not appear as distinctly in the network (Figure 6). However, like most extremophiles they appear towards the edge of the network. We also found that plasmids encoding proteins with an alkaline character tend to form more edges than plasmids encoding proteins with an acidic character, though the effect is small (1.126e-1).

The rest of the attributes have a smaller, though also significant, effect on the network. The number of proteins in a plasmid (protein number) and the molecular size of the proteins are similar in terms of their effect sizes. Plasmids with large numbers of proteins or those where the average protein size is large, result in more inter-plasmid links than we see between plasmids with few proteins or small proteins. The bipartite ERGM also confirms this trend since plasmid size is the biggest estimated attribute that has an effect in the network showing that small plasmids have a lowered probability of forming edges than big plasmids. However, the molecular size in the bipartite ERGM indicates that big proteins link more often with small proteins than with other big proteins. Since the unipartite projection of the plasmids works with the average molarity of all the proteins in a plasmid it seems to be more an indication of the size of the plasmid than the size of the proteins that the plasmid encodes. Considering the bipartite ERGM molarity value it seems that plasmids tend to encode differing protein size. In other words, they tend to avoid encoding exclusively big proteins or exclusively small proteins. Our analysis indicates that hydrophobic proteins have a slightly reduced chance of forming an edge compared with hydrophilic proteins. Finally, for the cost, we find that plasmids with costly proteins are more prone to link between themselves and these proteins have a higher chance of forming edges, but these effects are very small (3.456e-4).

Finally, for the fit of the model to the original data, although the model fits the original data well, the differing regions that can be seen in Figure 9 indicate that there are some properties of the network for which the model does not account. This could be that we are missing some important attributes or that they could be related to taxonomy or function, which are not modelled by the ERGMs.

Conclusions

We have constructed plasmidome gene networks from two different perspectives: sequence similarity and gene association. Our investigation indicates that the global plasmid network shows strong organisation by the properties of their gene sequences, the functions of the proteins encoded and the taxonomy of their bacterial hosts..

Coincidence analysis of plasmid borne genes reveals modules of both taxonomic and functional organisation. In particular our results are in agreement with previous studies (59, 78, 79) indicating the relative isolation of the Borrelia plasmids and their constituent genes within. In line with the plasmids themselves, there are modules of functionality shared between diverse host ranges, as well as modules specific to particular taxa. Whilst many of the associations form a large interconnected network, there are a number of free-floating island communities, presenting specific functions such as toxin-antitoxin systems and phage assembly.

In addition to the functional groupings highlighted by the network, the dependencies between the groups are also identified. This has the potential to be of use in cases where such dependencies themselves have not been studied. This contrasts with gene expression networks, which look at in-vivo “runtime” relationships, or non-systems analysis approaches, which require the purpose of all genes to be known up-front.

Network analysis has already been used to map plasmid-space. The "Plasmid Atlas" presents an extra-plasmid map, showing plasmid relatedness and allowing inferences about whole-structure evolution to be made (80). In contrast, our gene coincidence analysis details an intra-plasmid map, showing the constructs behind specific gene inclusion patterns.

The coincidence analysis employed in this study investigates association and dissociation of the relatedness of the underlying plasmids. We acknowledge that it is difficult to fully separate taxonomic, functional and environmental relationships between the genes and further research is required in order to fully determine the factors driving the formation of the individual modularities. The use of linear mixed models (LMMs) such as those implemented by BUGWAS (81) and GEMMA (82) in the analysis of phenotypic relations may be useful to pull apart interdependent factors, however these computationally intensive methods are typically used in analysing few traits within carefully preselected

species (83) and further research is required to determine their applicability to domain level, repeated analyses.

Bibliography

1. C. M. Thomas, et al., Annotation of plasmid genes. Plasmid (2017).

2. G. del Solar, R. Giraldo, M. J. Ruiz-Echevarría, M. Espinosa, R. Díaz-Orejas, Replication and Control of Circular Bacterial Plasmids. Microbiol. Mol. Biol. Rev. 62, 434–464 (1998).

3. C. M. Thomas, K. M. Nielsen, Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. Nat. Rev. Microbiol. 3, 711 (2005).

4. H. Ochman, J. G. Lawrence, E. A. Groisman, Lateral gene transfer and the nature of bacterial innovation. Nature 405, 299 (2000).

5. C. I. Kado, Origin and evolution of plasmids. Antonie Van Leeuwenhoek 73, 117–126 (1998).

6. M. J. Bottery, A. J. Wood, M. A. Brockhurst, Temporal dynamics of bacteria-plasmid coevolution under antibiotic selection. ISME J. 13, 559 (2019).

7. S. N. Cohen, Transposable genetic elements and plasmid evolution. Nature 263, 731–738 (1976).

8. J. Petersen, Phylogeny and compatibility: plasmid classification in the genomics era. Arch. Microbiol. 193, 313–321 (2011).

9. F. D. Ciccarelli, et al., Toward Automatic Reconstruction of a Highly Resolved Tree of Life. Science 311, 1283–1287 (2006).

10. R. Jain, M. C. Rivera, J. E. Moore, J. A. Lake, Horizontal Gene Transfer Accelerates Genome Innovation and Evolution. Mol. Biol. Evol. 20, 1598–1602 (2003).

11. H. Heuer, K. Smalla, Plasmids foster diversification and adaptation of bacterial populations in soil. FEMS Microbiol. Rev. 36, 1083–1104 (2012).

12. J. O. McInerney, A. McNally, M. J. O'Connell, Why prokaryotes have pangenomes. Nat. Microbiol. 2, 17040 (2017).

13. V. Burrus, G. Pavlovic, B. Decaris, G. Guédon, Conjugative transposons: the tip of the iceberg. Mol. Microbiol. 46, 601–610 (2002).

14. J. P. Hall, M. A. Brockhurst, C. Dytham, E. Harrison, The evolution of plasmid stability: Are infectious transmission and compensatory evolution competing evolutionary trajectories? Plasmid 91, 90–95 (2017).

15. E. Bosi, R. Fani, M. Fondi, The mosaicism of plasmids revealed by atypical genes detection and analysis. BMC Genomics 12, 403 (2011).

16. M. Kohiyama, S. Hiraga, I. Matic, M. Radman, Bacterial Sex: Playing Voyeurs 50 Years Later. Science 301, 802–803 (2003).

17. R. Fernandez-Lopez, S. Redondo, M. P. Garcillan-Barcia, F. de la Cruz, Towards a taxonomy of conjugative plasmids. Curr. Opin. Microbiol. 38, 106–113 (2017).

18. J. Davison, Genetic Exchange between Bacteria in the Environment. Plasmid 42, 73–91 (1999).

19. C. Sasakawa, N. Takamatsu, H. Danbara, M. Yoshikawa, A method of plasmid classification by integrative incompatibility. Plasmid 3, 116–127 (1980).

20. M. P. Garcillán-Barcia, M. V. Francia, F. de La Cruz, The diversity of conjugative relaxases and its application in plasmid classification. FEMS Microbiol. Rev. 33, 657–687 (2009).

21. A. Orlek, et al., Plasmid Classification in an Era of Whole-Genome Sequencing: Application in Studies of Antibiotic Resistance Epidemiology. Front. Microbiol. 8 (2017).

22. R. P. Novick, Plasmid incompatibility. Microbiol. Rev. 51, 381–395 (1987).

23. H. J. Atkinson, J. H. Morris, T. E. Ferrin, P. C. Babbitt, Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. PLOS ONE 4, e4345 (2009).

24. N. Song, J. M. Joseph, G. B. Davis, D. Durand, Sequence Similarity Network Reveals Common Ancestry of Multidomain Proteins. PLOS Comput. Biol. 4, e1000063 (2008).

25. I. D. Melamed, Models of Co-occurrence. ArXivcmp-Lg9805003 (1998) (May 28, 2019).

26. D. Zhou, S. A. Orshanskiy, H. Zha, C. L. Giles, Co-ranking Authors and Documents in a Heterogeneous Network in Seventh IEEE International Conference on Data Mining (ICDM 2007), (2007), pp. 739–744.

27. X. Mao, S. Yuan, W. Xu, D. Wei, Recipe recommendation considering the flavor of regional cuisines in 2016 International Conference on Progress in Informatics and Computing (PIC), (2016), pp. 32–36.

28. T. Simas, M. Ficek, A. Diaz-Guilera, P. Obrador, P. R. Rodriguez, Food-Bridging: A New Network Construction to Unveil the Principles of Cooking. Front. ICT 4 (2017).

29. Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, A.-L. Barabási, Flavor network and the principles of food pairing. Sci. Rep. 1, 196 (2011).

30. S. Uhlmann, et al., Global microRNA level regulation of EGFR-driven cell-cycle protein network in breast cancer. Mol. Syst. Biol. 8, 570 (2012).

31. A. Bashashati, et al., DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. Genome Biol. 13, R124 (2012).

32. B. Ozdemir, W. Abd-Almageed, S. Roessler, X. W. Wang, iSubgraph: Integrative Genomics for Subgroup Discovery in Hepatocellular Carcinoma Using Graph Mining and Mixture Models. PLOS ONE 8, e78624 (2013).

33. R. Diestel, Graph Theory (Springer Science & Business Media, 2006).

34. P. Zhang, et al., Clustering coefficient and community structure of bipartite networks. Phys. Stat. Mech. Its Appl. 387, 6869–6875 (2008).

35. P. G. Lind, M. C. González, H. J. Herrmann, Cycles and clustering in bipartite networks. Phys. Rev. E 72, 056127 (2005).

36. G. Robins, M. Alexander, Small Worlds Among Interlocking Directors: Network Structure and Distance in Bipartite Graphs. Comput. Math. Organ. Theory 10, 69–94 (2004).

37. M. Morris, M. S. Handcock, D. R. Hunter, Specification of Exponential-Family Random Graph Models: Terms and Computational Aspects. J. Stat. Softw. 24, 1548–7660 (2008).

38. M. S. Handcock, D. R. Hunter, C. T. Butts, S. M. Goodreau, M. Morris, statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data. J. Stat. Softw. 24, 1548–7660 (2008).

39. J. W. Shapiro, C. Putonti, Gene Co-occurrence Networks Reflect Bacteriophage Ecology and Evolution. mBio 9, e01870-17 (2018).

40. C. Beck, H. Knoop, R. Steuer, Modules of co-occurrence in the cyanobacterial pan-genome reveal functional associations between groups of ortholog genes. PLOS Genet. 14, e1007239 (2018).

41. P.-J. Kim, N. D. Price, Genetic Co-Occurrence Network across Sequenced Microbes. PLOS Comput. Biol. 7, e1002340 (2011).

42. A. Di Cesare, et al., Co-occurrence of integrase 1, antibiotic and heavy metal resistance genes in municipal wastewater treatment plants. Water Res. 94, 208–214 (2016).

43. C. Pal, J. Bengtsson-Palme, E. Kristiansson, D. G. J. Larsson, Co-occurrence of resistance genes to antibiotics, biocides and metals reveals novel insights into their co-selection potential. BMC Genomics 16, 964 (2015).

44. , UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 47, D506–D515 (2019).

45. D. Charif, J. R. Lobry, “SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis” in Structural Approaches to Sequence Evolution: Molecules, Networks, Populations, Biological and Medical Physics, Biomedical Engineering., U. Bastolla, M. Porto, H. E. Roman, M. Vendruscolo, Eds. (Springer Berlin Heidelberg, 2007), pp. 207–232.

46. D. Osorio, P. Rondón-Villarrea, R. Torres, Peptides: a package for data mining of antimicrobial peptides. R J. 7 (2015).

47. M. Ashburner, et al., Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25–29 (2000).

48. , The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 47, D330–D338 (2019).

49. S. McGinnis, T. L. Madden, BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 32, W20–W25 (2004).

50. S. M. Van Dongen, “Graph clustering by flow simulation.” (2000).

51. L. S. Haggerty, et al., A Pluralistic Account of Homology: Adapting the Models to the Data. Mol. Biol. Evol. 31, 501–516 (2014).

52. , Proceedings of the Python in Science Conference (SciPy): Exploring Network Structure, Dynamics, and Function using NetworkX (November 24, 2016).

53. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 2008, P10008 (2008).

54. R. Lambiotte, J.-C. Delvenne, M. Barahona, Laplacian Dynamics and Multiscale Modular Structure in Networks. IEEE Trans. Netw. Sci. Eng. 1, 76–90 (2014).

55. M. Matsumoto, T. Nishimura, Mersenne Twister: A 623-dimensionally Equidistributed Uniform Pseudo-random Number Generator. ACM Trans Model Comput Simul 8, 3–30 (1998).

56. M. Bastian, S. Heymann, M. Jacomy, others, Gephi: an open source software for exploring and manipulating networks. ICWSM 8, 361–362 (2009).

57. M. E. J. Newman, Modularity and community structure in networks. Proc. Natl. Acad. Sci. 103, 8577–8582 (2006).

58. W. Li, D. Schuurmans, Modular Community Detection in Networks in Twenty-Second International Joint Conference on Artificial Intelligence, (2011) (September 23, 2019).

59. A. G. Barbour, Plasmid analysis of Borrelia burgdorferi, the Lyme disease agent. J. Clin. Microbiol. 26, 475–478 (1988).

60. G. Chaconas, K. Kobryn, Structure, Function, and Evolution of Linear Replicons in Borrelia. Annu. Rev. Microbiol. 64, 185–202 (2010).

61. H. Brim, et al., Engineering Deinococcus radiodurans for metal remediation in radioactive mixed waste environments. Nat. Biotechnol. 18, 85 (2000).

62. H. Brim, et al., Deinococcus radiodurans engineered for complete toluene degradation facilitates Cr(VI) reduction. Microbiology 152, 2469–2477 (2006).

63. R. Dixit, et al., Bioremediation of Heavy Metals from Soil and Aquatic Environment: An Overview of Principles and Criteria of Fundamental Processes. Sustainability 7, 2189–2212 (2015).

64. W. G. Eberhard, Why do bacterial plasmids carry some genes and not others? Plasmid 21, 167–174 (1989).

65. A. Schlüter, L. Krause, R. Szczepanowski, A. Goesmann, A. Pühler, Genetic diversity and composition of a plasmid metagenome from a wastewater treatment plant. J. Biotechnol. 136, 65–76 (2008).

66. D. R. Call, M.-S. Kang, J. Daniels, T. E. Besser, Assessing genetic diversity in plasmids from Escherichia coli and Salmonella enterica using a mixed-plasmid microarray. J. Appl. Microbiol. 100, 15–28 (2006).

67. R. Albert, H. Jeong, A.-L. Barabási, Error and attack tolerance of complex networks. Nature 406, 378 (2000).

68. M. J. Middelveen, et al., Mixed Borrelia burgdorferi and Helicobacter pylori Biofilms in Morgellons Disease Dermatological Specimens. Healthcare 7 (2019).

69. K. O. Stetter, Extremophiles and their adaptation to hot environments. FEBS Lett. 452, 22–25 (1999).

70. K. Suhre, J.-M. Claverie, Genomic Correlates of Hyperthermostability, an Update. J. Biol. Chem. 278, 17198–17202 (2003).

71. D. P. Kreil, C. A. Ouzounis, Identification of thermophilic species by the amino acid compositions deduced from their genomes. Nucleic Acids Res. 29, 1608–1615 (2001).

72. Y. Nishio, et al., Comparative Complete Genome Sequence Analysis of the Amino Acid Replacements Responsible for the Thermostability of Corynebacterium efficiens. Genome Res. 13, 1572–1579 (2003).

73. H. Takami, et al., Thermoadaptation trait revealed by the genome sequence of thermophilic Geobacillus kaustophilus. Nucleic Acids Res. 32, 6292–6303 (2004).

74. A. Henne, et al., The genome sequence of the extreme thermophile Thermus thermophilus. Nat. Biotechnol. 22, 547–553 (2004).

75. L. D. Hurst, A. R. Merchant, High guanine–cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proc. R. Soc. Lond. B Biol. Sci. 268, 493–497 (2001).

76. H. Wu, Z. Zhang, S. Hu, J. Yu, On the molecular mechanism of GC content variation among eubacterial genomes. Biol. Direct 7, 2 (2012).

77. K. Talley, E. Alexov, On the pH-optimum of activity and stability of proteins. Proteins 78, 2699–2706 (2010).

78. S. Halary, J. O. McInerney, P. Lopez, E. Bapteste, EGN: a wizard for construction of gene and genome similarity networks. BMC Evol. Biol. 13, 146 (2013).

79. C. M. Fraser, et al., Genomic sequence of a Lyme disease spirochaete, Borrelia burgdorferi. Nature 390, 580 (1997).

80. T. F. Jesus, et al., Plasmid ATLAS: plasmid visual analytics and identification in high-throughput sequencing data. Nucleic Acids Res. 47, D188–D194 (2019).

81. S. G. Earle, et al., Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nat. Microbiol. 1, 16041 (2016).

82. X. Zhou, M. Stephens, Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44, 821–824 (2012).

83. D. Falush, Bacterial genomics: Microbial GWAS coming of age. Nat. Microbiol. 1, 16059 (2016).

Fungal phylogeny reconstructed using heterogeneous models reveals new placement of *Microsporidia*

Ignacio Riquelme-Medina[1], James O. McInerney[1,2,#]

[1.] Division of Evolution and Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester Academic Health Science Centre, Manchester M13 9PL, United Kingdom.

[2.] . School of Life Sciences, The University of Nottingham, Nottingham NG7 2UH, United Kingdom.

[#] To whom correspondence should be addressed.

Phone: +44 (0)161 2751843

Email: james.mcinerney@manchester.ac.uk; james.mcinerney@nottingham.ac.uk

Website: http://mcinerneylab.com/

Abstract

The fungal kingdom is a group of organisms that have a large impact both to ecosystems, mainly playing a role in organic matter decomposition, and to humans, both as disease agents with a negative impact in humans and economy and as useful organisms for research and industry. Still, the fungal phylogeny remains partly unresolved with the presence on many *incertae sedis* clades, without a robust placement on the phylogenetic tree. Some of the more problematic groups whose placement remains unclear is the intracellular parasites *Microsporidia* and Cryptomycota, who have accelerated rates of nucleotide substitution and reduced genomes, which makes the resolution of these groups hard. In this project we used tree and data heterogeneous models, which can account for different substitution rates between the proteins used to reconstruct the phylogeny and for rate changes of the same protein in different parts of the phylogeny. By using heterogeneous models we could obtain a general structure of the fungal phylogeny as well as a robust placement for both *Microsporidia* and Cryptomycota, the former being rooted deeply within the fungi and the later being a sister branch to the fungi.

Introduction

The Fungal kingdom is one of the most variable that exists on the planet. It makes a huge impact on planetary ecosystems, playing a major role in wood decay for instance (1). Fungi also have a major impact on human activity, by performing fermentations needed in the food and beverages industries, acting as pathogens in crop and human diseases, and in research, serving as model organisms, and in assisting drug discovery. Fungi and animals have a common ancestor dating to approximately one billion years ago (2). Fungi are characterized by being highly variable, eukaryotic, with simple morphologies (typically filamentous but they can also be unicellular). Fungi have relatively small genomes and are heterotrophic (3). They can have haploid and diploid stages in their life cycle and do not have movement except for some flagellated spores. They can be saprobes, symbionts and parasites of plants, animals or other fungi, but they are not able to photosynthesise. Owing to their variability they have adapted repeatedly to the same kind of environments at different points in their phylogeny and phenotypes such as parasitism and mutualism are not exclusively found in a single clade (3). As an example of how quickly fungi can adapt, we can look at the Hymenochaetales order, which has saprobes, mycorrhiza and strong and weak plant parasites amongst its members (4). A fungal phylogenetic tree has been recently corroborated by the Assembling the Fungal Tree of Life project (AFToL) (5) due to the lack of tools to reach a consensus classification in former times (6). However, many branches of the phylogeny are not well resolved, and these are classified as *Incertae sedis* as they appear in the NCBI database taxonomy (7). In particular some of the earlier branches, for example *Microsporidia* and *Cryptomycota* (*Rozella*), are still being analysed in order to address whether or not they belong to the fungal kingdom or are sister groups to the fungi (8–14).

Fungi have long been considered to be good model organisms for eukaryotes, owing to their small genome size, the fact that they are easy and inexpensive to maintain, they are relatively amenable to genetic manipulation, and that some form tissues that allow transcriptomic experiments. In recent years efforts to sequence complete fungal genomes have intensified, particularly with the 1,000 Fungal Genome Project (FGP) (15) which is a project that intends to sequence 1,000 fungal genomes and make them accessible to the public. With the FGP hundreds of fungal genomes have already been sequenced but remain to be analysed. Here we report the use of these data in order to construct an updated phylogeny of the fungi, paying special attention to unresolved groups.

There is ongoing discussion about the inclusion of *Microsporidia* and *Cryptomycota* in the fungal kingdom (16). The difficulty associated with placing these groups on the eukaryote phylogeny is due to their specialised obligate intracellular parasitic lifestyle. This lifestyle has led to considerable genome reduction, and unusually high nucleotide and amino acid substitution rates, as well as the loss of their mitochondria (17–19). High substitution rates are known to be problematic for accurate phylogeny reconstruction (20, 21), and a number of solutions have been proposed to mitigate these effects (22–24). In this paper we tested whether using a model that specified a homogenous evolutionary process would accurately capture the signals in the data. We found that such a model was a poor fit to the data and in order to properly account for the signals in the data we used both tree-heterogeneous and sequence-heterogeneous phylogenetic reconstruction approaches (25, 26). A compositionally heterogeneous model specifies more than one evolutionary process for the different genes or proteins used for a super-matrix reconstruction as well as being able to use several models of evolution for different branches of a phylogeny. By allowing the phylogenetic reconstruction to account for different rates of evolution for different genes and branches we should be able to account for the genome reductions observed in *Microsporidia* and *Cryptomycota.*

Methods

Dataset

A total of 690 genomes were available in the FGP (15) were downloaded using the Globus data sharing tool (27). For this project the DNA sequence, the protein sequence, the annotation and the functional annotations were transferred, using the filtered models for each genome. These files are included for every genome in the FGP database, and also include GO (28), KEGG pathways (29), InterPro (30), KOG (31),and SignalP (32) annotations.

In order to evaluate the quality of the genomes we used the Fungal Genome Mapping Project (FGMP) (33) a framework designed to check if several fungi conserved and ultraconserved genes are present in the genome. These genes should be present in every fungal genome and are a good indicator of the ratio of completeness of the genome. Poor quality genomes were identified as having fewer than 75% of the conserved genes present and were discarded leaving a total of 671 genomes for further analysis. Owing to their known reduce genome size, parasitic fungi were not discarded in this step despite showing quality values lower than 30% of present conserved genes.

To deal with the size of our dataset and its associated metadata a SQL database was built using PostgreSQL versions 11 to 12.2. In this database we built a protein table where the protein identification, protein sequence and protein family of each protein was stored. Other tables where built to store the functional annotations and were connected to the proteins by the identification numbers provided by FGP. Finally, a table with all the fungal species and their categories (WoL, habitat, infectious, extremophile and notes) was connected to the protein families in the protein table by the edge information obtained in the previous step with MCL. The structure of the database can be seen in Figure 14. The python package psycopg2 was used to interact and retrieve information from the database.

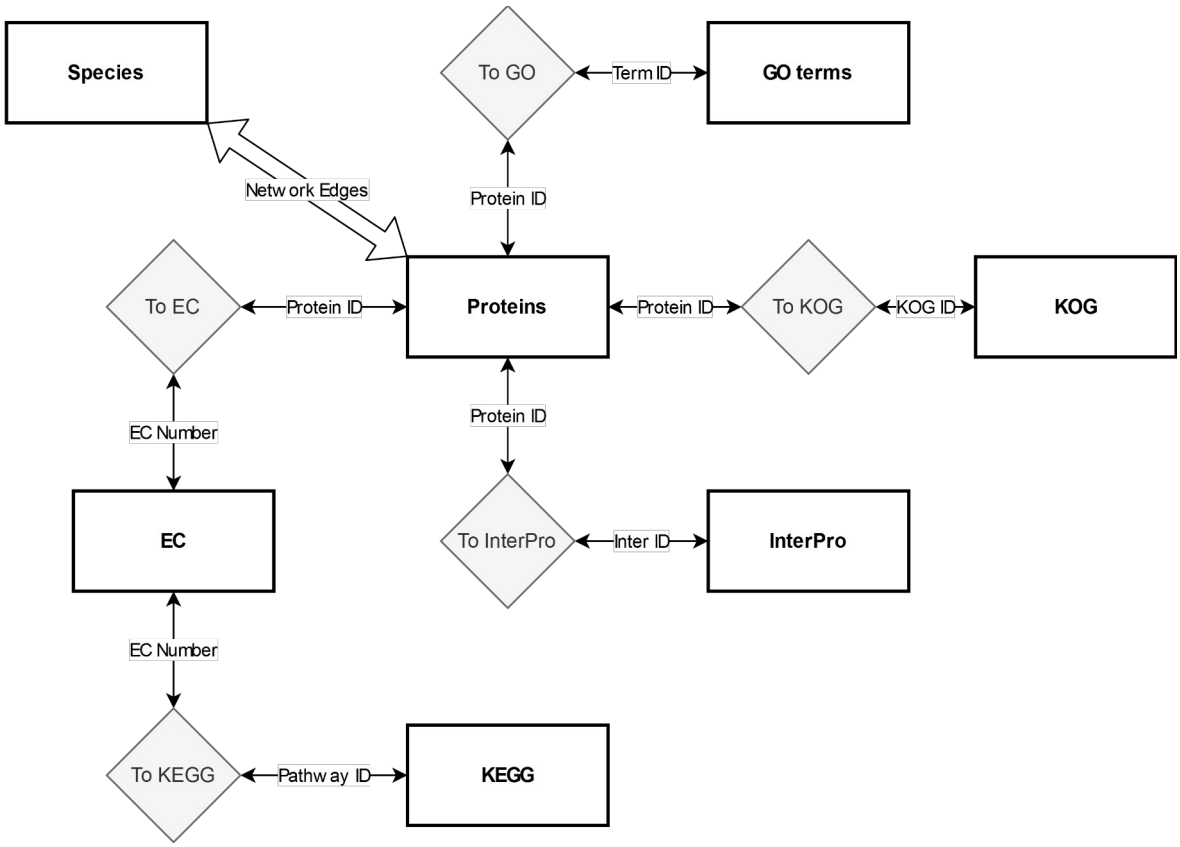

Figure 14: Diagram of the database tables and relationships. Rectangles represents tables that store information and diamonds represent relational tables that store the connections between tables. Each link indicates the identifier used to connect the tables.

Data Processing

Protein sequences were then used in order to perform an all-versus-all BLAST (34, 35) search (blastp v2.4.0, e value = 1*e-6). The BLAST output was processed using the Markov Cluster Algorithm (MCL) (36) version 14-137 in order to detect clusters. To select an appropriate inflation value for our dataset we looked for the highest inflation value where most conserved proteins (e.g. ribosomal proteins) where properly recovered under the same cluster. The final inflation value was set to 1.4. This resulted in the recovery of 965,828 clusters from the dataset.

We chose to construct phylogenetic hypotheses using a concatenation of several alignments. Therefore, we sought to identify clusters that include almost all 671 taxa, where the clusters contained little, to no, duplicated genes. We identified an initial set of 507 clusters that had 671 ± 20 taxa, allowing for some gene duplications, and that included at least 95% of all the taxa to avoid clusters with high amount of duplications.

Phylogenetic Analyses

Multiple sequence alignment was carried out using MAFFT v6.611b (37) (auto1 option) as an alignment tool. We then used Trimal v1.2 (38) (automated1 option) in order to remove poorly aligned positions in the alignments. Prottest 3 (39) was used to assess the best substitution model for each of the alignments. After assigning the best model to each alignment, phylogenetic hypotheses were constructed using RAxML v8 (40). Since gene duplications have been shown to be a source of discrepancies between a gene tree and a species tree, often because duplicated genes evolve at different rates (41), we tested further the resulting 507 trees to check whether gene duplications were few and evolutionarily close in each tree. If the duplications are evolutionarily close, it would mean that there were few nucleotide substitutions between duplicated genes and that the different rate of evolution would have less impact on the gene tree. This was checked using ETE3 (42) node distance comparison function in python v2.7. To do this a custom script was made that checks the distance between all the duplicated gene pairs in one tree and reports the gene tree as valid to construct a species tree if: a) the duplicates are in very close proximity (same branch) and b) if no more than one pair of duplicates is further away than this set distance. Using this relaxed criterion, a total of 58 gene trees were selected. In addition, out-groups were added to each gene unaligned file, using blastp as described before to search for similar proteins in 4 different organisms: a mammal (*Homo sapiens*), a cnidarian (*Nematostella Vectensis*), a choanoflagellate (*Monosiga Brevicollis*) and a plant (*Arabidopsis thaliana*) making a total of 675 taxa in the dataset. The gene sequences were aligned again with the outgroups included as previously described and then the aligned gene sequences were concatenated in order to construct the species tree. If one gene tree was missing a taxon from the complete species tree, the gap was filled with missing characters of the same length as the gene.

The resulting concatenated alignment was of 11,559 amino acids in length. The programs Prottest 3 and RAxML were used as described previously in order to construct several phylogenetic hypotheses using a single model for all parts of the alignment. Bootstrap resampling (43) using 100 replicates was used to assess support for internal branches.

The same alignment file was used with a heterogeneous model tree using the P4 software program (25). P4 implements both data-heterogeneous and tree-heterogeneous models using a modified version of MCMC. Since heterogeneous models are much more computationally expensive than homogeneous models, the original alignment file could not

be used to construct a heterogeneous tree in a reasonable amount of time. Therefore, we had to make a reduced dataset using a selection of taxa from the original alignment. We chose a single taxon to represent each uncontroversial group in the maximum-likelihood tree apart from the out-group. However, the *Microsporidia* branch and *Rozella* were left intact as they were the most problematic branches in the original tree. The number of taxa in the reduced alignment was 59.

First in order to account for the data heterogeneity, the alignment file was used as input for the PartitionFinder 2.11 program (44). PartitionFinder searches for differing rates of substitution across regions described by the user in an alignment with the purpose of dividing them in partitions of proteins or genes with similar evolutionary rates and assigning a phylogenetic model to each partition. In our case we split the alignment in to each one of the partitions indicated by PartitionFinder (Partitions = 27). Afterwards the partition data and the alignment were analysed using P4, assigning a model to each partition in accordance with PartitionFinder results. Furthermore, each partition was also assigned the second-best model found by PartitionFinder in order to account for the heterogeneity across the tree. Finally, MCMC was run for 300,000 generations, assessing whether the MCMC chains had converged at the end of the process. Additionally, the tail area probability of the MCMC run was calculated using a function included in P4 to asses whether the phylogeny recovered by MCMC could plausibly explain our original data.

The final step in the analysis was to combine the results obtained from the heterogeneous model tree to the full dataset. To do this we used PartitionFinder again with the full dataset since RAxML allows for data-heterogeneous models, obtaining a total of 51 partitions. Each partition was assigned the substitution model indicated by PartitionFinder for RAxML. To avoid problems with long branch attraction in this tree hypothesis we forced RAxML to root the tree on the out-group, as it was supported by the reduced dataset phylogenetic tree. After the phylogenetic analysis was finished, bootstrap resampling with 100 replicates was used to asses support for internal branches.

Phylogeny validation

In order to validate our phylogenetic tree, we compared the complete dataset phylogenetic tree with other phylogenies using the program TOPD/FMTS (45). The phylogenies were obtained and downloaded from the TreeBASE database (46, 47), selecting phylogenies that

included taxa from our dataset. This software allows to calculate the nodal distance and the split distance, also known as the Robinson–Foulds metric (48), between any given pair of phylogenetic trees.

As a further way to test the validity of our methodology, we decided to apply the same analysis done to our phylogeny to a well supported fungal phylogeny proposed by *McCarthy* and *Fitzpatrick (49),* that uses a similar but smaller dataset from the FGP. This validation dataset does not contain the more problematic groups in the fungi, like *Microsporidia*, which would remove any artefact introduced by these groups in the phylogeny.

We first checked whether the species present in *McCarthy* and *Fitzpatrick* dataset were present in our dataset, having passed our genome quality evaluation step. For species in the paper dataset that were not present in our dataset but had other close species in the same family, we used this species instead. Specifically, *Zymoseptoria tritici* was changed for *Zymoseptoria ardabiliae, Candida albicans* for *Candida tanzawaensis, Microbotryum lychnidis-dioicae* for *Microbotryum violaceum* and *Rhizopus oryzae* for *Rhizopus microsporus.* Species that did not appear in our dataset and did not have a close relative where excluded from the analysis, specifically *Endocarpon pusillum, Orpinomyces sp. C1A,* and *Batrachochytrium dendrobatidis.* Instead of using an external source genome for *Allomyces macrogynus* that was used in the paper dataset, we used a sequence genome from the FGP instead. This made a total of 81 species that were used in this validation phylogenetic analysis. This dataset will be referred as validation dataset from this point onwards.

We then extracted a subset from our blast output that included only edges between the species present in the validation dataset and proceeded to apply the same steps done in our phylogenetic analysis. We used MCL with the same inflation value, 1.4, to obtain a total of 155,969 protein clusters. Afterwards, we filtered the clusters to select candidate genes for the phylogenetic analysis with a more strict filter due to the high amount of genes returned if we applied the same criteria as with our dataset. We filtered only for clusters that had every species in the dataset present and at most two duplications, which left a total of 35 candidate clusters. Using the same procedure as with our dataset, we selected a total of 21 clusters that had either no duplications or duplications in close proximity to each other. For clusters that had duplications, one of them was removed randomly to leave only one protein per species. As done with our analysis, the selected clusters were aligned, trimmed and concatenated with the same settings as with the original dataset to form a concatenated

alignment with a final length of 7,611 amino acids. We used PartitionFinder again to determine the heterogeneity of substitution rates in the data (Validation Partitions = 13). Finally, we used two different P4 analyses, one considering only data-heterogeneity, which included the model indicated by PartitionFinder for each partition, and one considering data and tree-heterogeneity, which used two models indicated by PartitionFinder for each partition. Both analyses MCMC processes were executed until convergence was achieved, using 100,000 generations for the data-heterogeneous analysis and 240,000 generations for the data and tree-heterogeneous analysis. The tail-area probability of each analysis was also calculated using P4.

Super-tree

As a complementary analysis to the heterogeneous models super-matrix we decided to use an alternative approach and make a phylogeny using super-trees. We decided to use Clann (50), which implements several methods to construct super-trees but focuses on matrix the representation with parsimony (MRP) method. MRP uses many individual gene trees to collect evidence for branch split support, i.e. checks how many gene trees support species A being in the same branch with species B or the opposite for every branch split in the super-tree. This split support is stored in the form of a matrix and then a maximum parsimony search is used in the matrix to construct the final super-tree.

Thus, we used the clusters obtained by applying MCL to the complete dataset, as each cluster is equivalent to a gene family, to construct the individual gene trees. We filtered the clusters to only include clusters with no duplications, since gene duplications introduce problems that are hard to deal with by current super-tree methods (51), and that have four taxa or more, leaving a total of 49,261 gene families. Each gene family was then aligned and trimmed as the super-matrix dataset (MAFFT auto1 option and Trimal automated1 option). Afterwards, Prottest 3 was used to assign the appropriate substitution model to each gene alignment and then gene trees were constructed using RAxML with the indicated substitution model for each tree. Finally, we used Clann to construct the super-tree by using the MRP method with default parameters for heuristic tree search (Parsimony analysis and tree bisection and reconnection as the type of heuristic search) with no repetitions.

Results

Phylogenetic Analysis

As an initial step in our phylogenetic analysis we estimated several phylogenetic hypotheses under the LG+I+G model in RAxML with bootstrap resampling support for the total amount of taxa in the dataset. These trees had 675 leaves (671 fungi and 4 out-group taxa). The hypotheses recovered the phylum and class groups in the fungi with high confidence, particularly in the *Dikarya*. However, there was no identifiable split between the in-groups and the out-groups. Moreover, in some hypotheses *Microsporidia* appear nesting within the out-group, an example of which can be seen in Figure 15. Deep branches in hypotheses where the out-group appears split showed low support values. So while the maximum-likelihood tree has provided good results for resolving short fungal branches, it failed to resolve the deep branches of the tree, due to how high the substitution rates are in the *Microsporidia* branch compared to the rest of the tree. This difference in branch length suggest a very different rate of evolution, as it has been proved that *Microsporidia* have many fast evolving genes (52), when compared to the rest of the fungi for which the model did not account in this tree, so this branch couldn't be resolved using this approach.

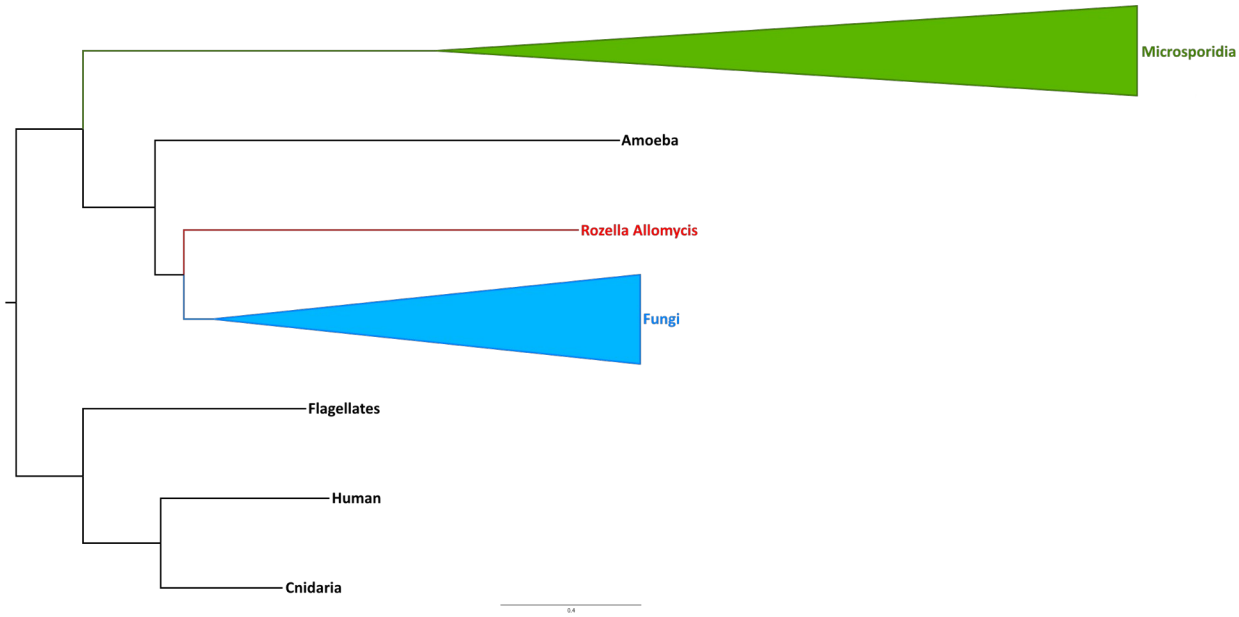

Figure 15: Example of a problematic phylogeny where the out-groups appear as a non-monophyletic clade, with Microsporidia appearing within the out-groups

In order to overcome problems associated with the low support of the long branches on our tree, we used heterogeneous models, which can account for variation in rates of evolution across the different proteins and in across the tree. Because heterogeneous models are much more computationally intensive than the homogeneous models, we used a reduced dataset that consisted of a representative taxon for each clade in the fungal in-group, though we retained all the out-group sequences and Microsporidia.

Figure 16: Reduced dataset data and tree heterogeneous Phylogenetic tree. Each branch is marked with the Posterior Probability Distribution (PPD) percentage of support. The out-groups are in black, early fungi are coloured red, Microsporidia are coloured green and Dikarya are coloured blue.

The resulting reduced tree was reconstructed in P4 using two models per partition to account for tree-heterogeneity of substitution rates using the substitution models indicated by PartitionFinder, resulting in two LG+I+G substitution model per partition for twenty four partitions and two LG+G substitution models per partition for the remaining three partitions (LogLikelihood = -557,634.03). From the 300,000 generations of the MCMC run, the first 200,000 were discarded as the MCMC had not converged, and a consensus tree of the remaining 100,000 generations was calculated. The branch split support values were obtained from the posterior probability distribution after MCMC had converged as indicated in the P4 1.2 documentation (The documentation is available online in the following link: P4 documentation). This phylogenetic hypothesis has reliably placed Microsporidia as a sister group to Dikarya and placing Rozella outside of the Fungi as can be seen in Figure 16. There is a complete split between the in-groups and the out-groups of the tree unlike the previous tree, which also indicates that when accounting for different rates of evolution with the heterogeneous models helps resolve deep branches.

To further test this phylogenetic hypothesis we set other two MCMC processes to check that they recovered phylogenies with similar topology, which proved to be true (Split support standard deviation between MCMC processes = 0.0719). We also checked the tail area probability of the phylogenetic hypothesis (0.1984) which means that our dataset could have plausibly be generated by the model. Finally we wanted to check how the likelihood of the phylogenetic hypothesis was improved by using a data and tree heterogeneous model, so we ran another two MCMC processes, one with an only data-heterogeneous model and another with a homogeneous model, in order to compare. For the data heterogeneous model we used only one substitution model per partition with the substitution models indicated before and run for 300,000 generations (LogLikelihood = -559,060.90, tail area probability = 0.1236). The data heterogeneous model topology is more similar to other topologies in the literature, placing Microsporidia as a sister group to the Fungi together with Rozella. For the homogeneous model we used a single LG+I+G model and run the MCMC for 300,000 generations (LogLikelihood = -562,923.74, tail area probability = 0.409). The homogeneous tree topology is similar to the data heterogeneous topology, with Microsporidia outside of the Fungi together with Rozella. However, even if the homogeneous model hypothesis can plausibly explain the data, some *Microsporidian* taxa appear spread across the tree within the fungi with low Posterior Probability support, and the likelihood of the model is much worse when compared with the more complex models.

Next we compared the different heterogeneous likelihood models by carrying out a likelihood ratio test. According to the test of the significance of the difference in log-likelihood scores between the trees, the data and tree heterogeneous model is a significantly better fit to the data than the data heterogeneous model (P value < 0.001). In summary both the tree data heterogeneous hypothesis and the data heterogeneous hypothesis with their respective topologies are plausible hypothesis for a fungal phylogeny and well supported by the data, but the data tree heterogeneous model is a significantly better hypothesis according to the likelihood of the model.

Finally, we defined partitions again with the complete alignment using PartitionFinder to account for varying rates of amino-acid substitution between proteins in the maximum-likelihood hypothesis with the complete dataset. Then we used the root that we obtained with the reduced dataset heterogeneous model we built a final ML tree defining the out-groups as indicated by the heterogeneous models hypothesis. This last phylogenetic hypothesis was constructed using RAxML under the LG4X+I+G substitution model with the data split into thirty partitions, the LG4M+I+G substitution model for sixteen partitions, LGF+I+G substitution model for two partitions and LG+I+G for the remaining two partitions as indicated by PartitionFinder. Then we assessed support for the tree internal branches of this tree by using bootstrapping resampling (100 pseudoreplicates) as shown in Figure 17. The branches of this tree were collapsed in clades matching the NCBI's fungal taxonomy database to allow for a better visualization.

The resulting trees obtained from the concatenated analysis are very similar to the current taxonomy of the fungi according to NCBI. Since the leaves in the tree in Figure 17 were collapsed according to NCBI's taxonomy it is clear that many of the groups found by our analysis are in agreement with the current consensus. This tree will be referred as Final tree and will be the main basis of our discussion.

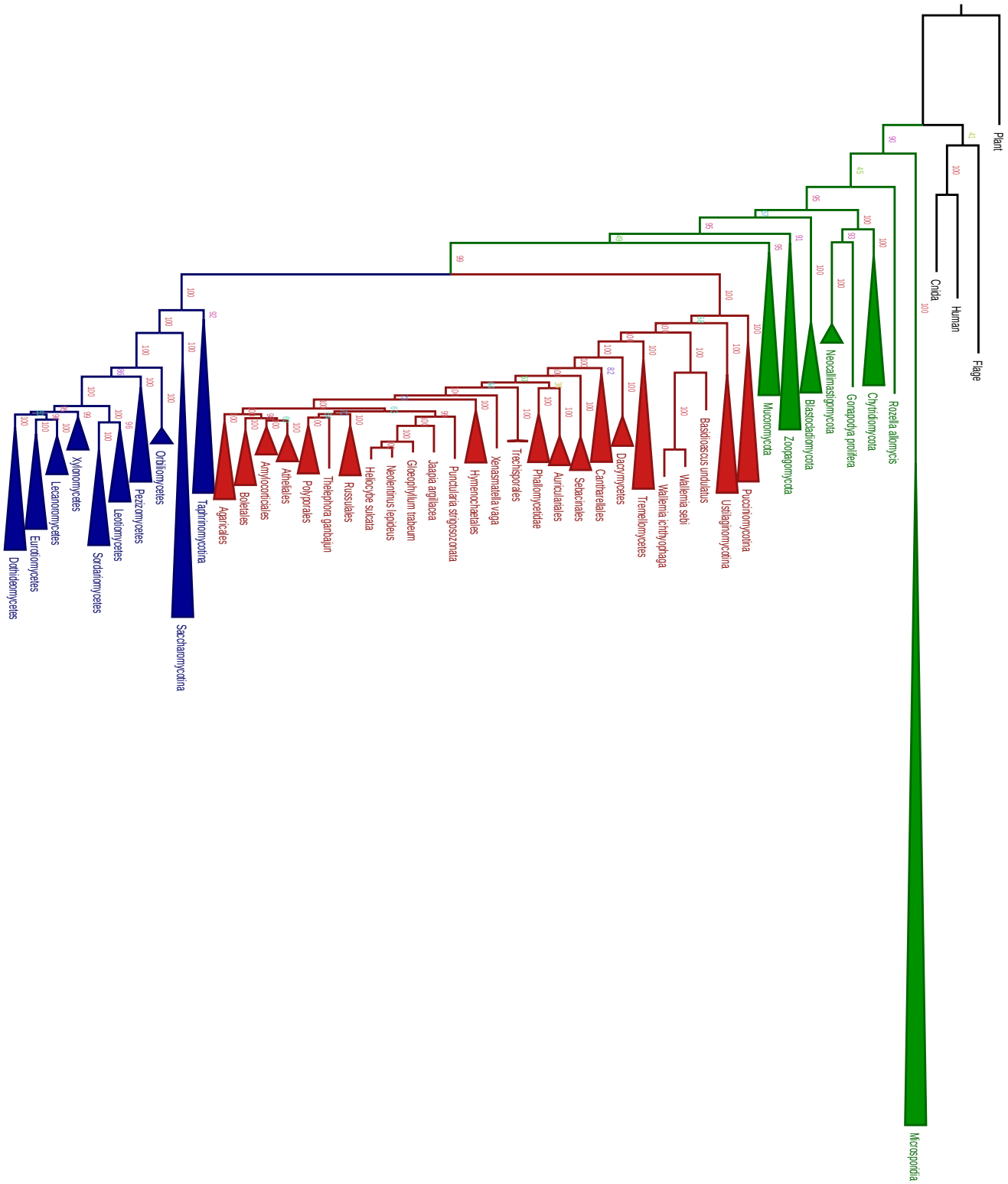

Figure 17: Fungal phylogenetic tree with bootstrap values. In Green are the groups outside of Dikarya, Red is Basidiomycota, Blue is Ascomycota and Black is the outgroup.

Phylogeny validation

Finally, we have compared our complete phylogeny with a number of previously recovered phylogenies from other studies as a means to validate the structure in our tree. The results obtained after analysing each pair of trees with TOPD/FMTS, i.e. our complete phylogeny compared to each of the literature phylogenies by using TOPD/FMTS nodal distance and split distance modes, can be observed in Table 1.

Table 1: TOPD/FMTS results of the comparison between the recovered phylogeny of the total taxa with the phylogeny proposed by each one of the following studies.

Study	Taxa in Common	Nodal Distance (Pruned / Unpruned)	Split Distance (Differences / Possible)	Disagreeing Taxa (taxa disagree / all taxa)
Zhao et al. A six-gene phylogenetic overview of Basidiomycota and allied phyla with estimated divergence times of higher taxa and a phyloproteomics perspective. (53)	4.10%	1.45 / 2.85	0.25 (22 / 88)	10 / 47
Davis et al. Genome-scale phylogenetics reveals a monophyletic Zoopagales (Zoopagomycota, Fungi). (54)	3.10%	0.41 / 0.81	0.05 (2 / 38)	3 / 22
Chen et al. Phylogenetic placement of Paratrichaptum and reconsideration of Gloeophyllales. (55)	2.10%	1.27 / 2.52	0.31 (8 / 26)	3 / 16
De Crop et al. A multi-gene phylogeny of Lactifluus (Basidiomycota, Russulales) translated into a new infrageneric classification of the genus. (56)	4.30%	0.9 / 1.77	0.15 (8/54)	5 / 30

As for the methods validation, we constructed two phylogenies, one tree and data heterogeneous phylogeny (using two LG+I+G substitution models for every partition as indicated by PartitionFinder, LogLikelihood = -571,114, tail area probability = 0.764), which can be seen in Figure 18, and one data heterogeneous phylogeny (using one LG+I+G substitution model for every partition as indicated by PartitionFinder, LogLikelihood = -572,933, tail area probability = 0.326). Both trees are identical in structure and can plausibly have explained the validation dataset, but the tree and data

heterogeneous phylogeny has a significantly higher log-likelihood (*P* value < 0.001) so this model is a better fit for the validation dataset in a similar way as previously happened with our original dataset. Branch split support from the posterior probability distribution was also lower for the data heterogeneous phylogeny, but both phylogenies had high support for every branch split.

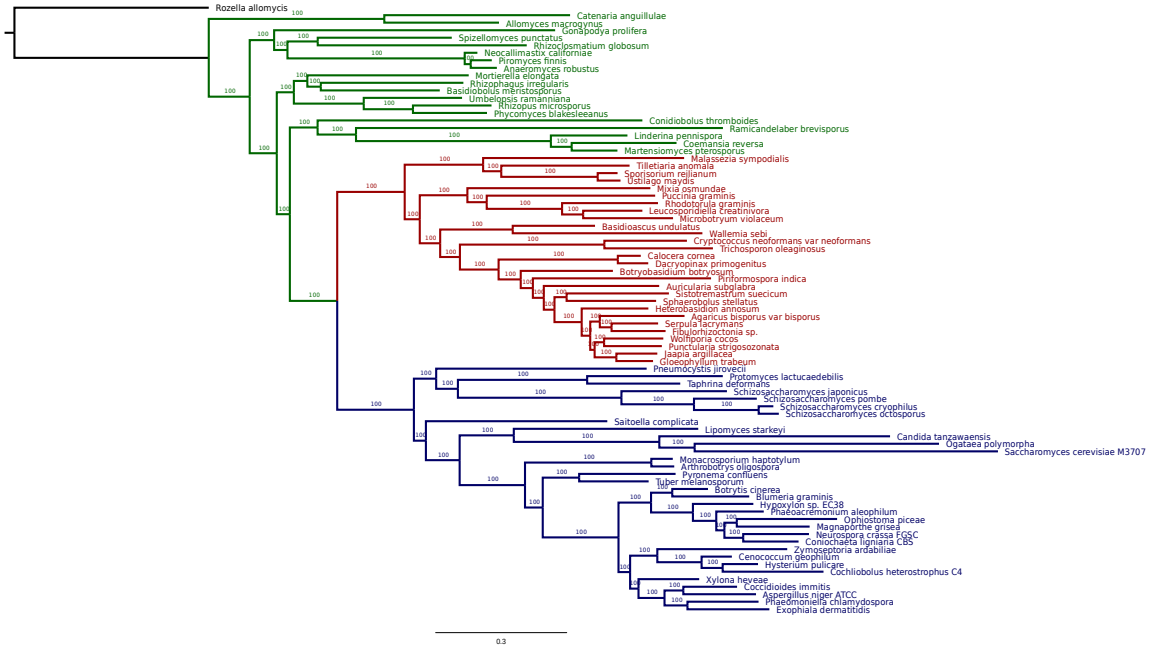

Figure 18: Validation dataset data and tree heterogeneous Phylogenetic tree. Each branch is marked with the PPD percentage of support. In Green are the early fungi, Red is Basidiomycota, Blue is Ascomycota and Black is the outgroup.

Super-tree

As for the alternative super-tree method the phylogeny that can be seen in Figure 19 was obtained by using the MRP method with a total of 49,261 single gene trees. Bootstrap analysis of the data could not be performed due to the long time it would take to complete for 100 iterations, as each iteration took several weeks to complete.

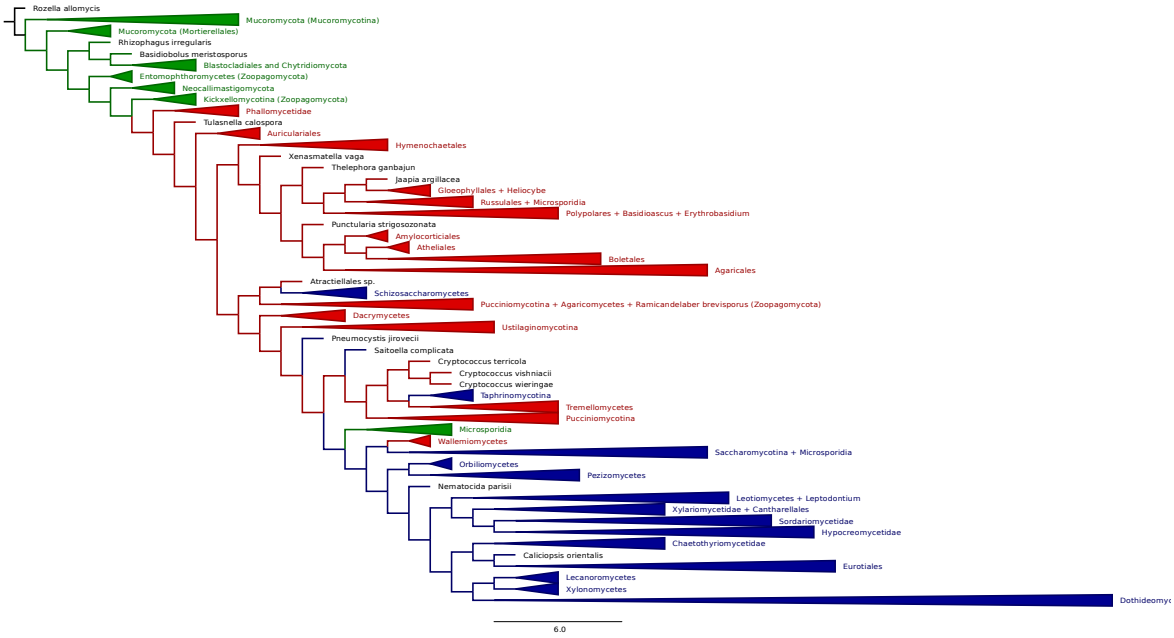

Figure 19: Super-tree reconstruction using the MRP method. Branches are collapsed when they are monophiletic according to NCBI's taxonomy. In Green are the groups of early fungi outside of Dikarya, Red represents Basidiomycota, Blue represents Ascomycota and Black is the out-group (Rozella Allomycis).

Discussion

Super-tree

We will begin our discussion with the alternative super-tree method, since it has obvious discrepancies with our super-matrix approach. First of all, the super-tree approach has been incapable of properly separating the two big clades present in Dikarya, Ascomycota and Basidiomycota. While in our super-matrix approach and in the literature these groups are clearly distinguished and separated, in the super-matrix approach Ascomycota and Basidiomycota appear mixed in some branches, for example in Figure 19 Taphrinomycotina (Ascomycota) appears as a sister branch of the Tremellomycetes (Basidiomycota) or Saccharomycotina (Ascomycota) as a sister branch of Wallemiomycetes (Basidiomycota). There are also issues with the early fungi outside of Dikarya, where clades like Zoopagomycota are not monophyletic and are instead separated and mixed with other groups like Mucoromycota, or even appear in branches within the Dikarya phylum like it is the case with *Ramicandelaber brevisporus*, which is present within Pucciniomycotina. Other problematic groups like Microsporidia appear mostly within Dikarya, but are also not a monophyletic group and are spread across several branches. On the contrary, the super-tree approach is able to recover clades with a more recent common ancestry, as many of the smaller groups present in the external branches of the phylogeny are in accordance with NCBI's taxonomy and could therefore be collapsed in Figure 19.

One possible explanation of why the super-tree phylogeny is struggling to reconstruct the deeper branches of the phylogeny can be the distribution of the gene tree size of our dataset. Almost 90% of the single gene trees used for the super-tree analysis have 7 taxa or less, which is 1% of the total species in our dataset, which may provide enough resolution to resolve smaller clades but not enough to resolve the deep branch splits in the phylogeny. The size of our dataset might be the cause of this issue as it might make the presence of single gene trees without duplication sparse. We faced the same issue previously when we were selecting the protein families to use with our super-matrix approach. For our large super-matrix approach, we needed to set lax filters and allow for some duplication events in order to find enough candidate genes. On the other hand, with the smaller validation super-matrix the filters to select candidate proteins needed to be more strict due to the large

number of candidate protein families retrieved by the previous filters. Therefore, the large size of our dataset can explain the lack of suitable protein families needed for super-tree reconstruction. The issues with problematic groups in the super-tree, namely that some of them are not monophyletic, might be explained by the presence of intracellular parasites like in Microsporidia, which have greatly reduced genomes and have even lost or repurposed ribosomal proteins (57, 58), which implies that single gene trees where these species are present are much more sparse. Other possible cause for the low accuracy of the super-tree is that single gene trees may have a similar problem as the super-matrix approach, where homogeneous substitution models are enough to model the evolutionary process in groups with recent common ancestry, but the more complex heterogeneous models are needed to model the evolutionary process of deep branches and problematic groups. These issues indicate that super-tree phylogeny reconstruction may not be a good fit for our particular dataset.

On the contrary, the super-matrix approach was capable of recovering the branch splits between the large clades in fungi, robustly splitting Dikarya from early fungi and Ascomycota from Basidiomycota, even when using only homogeneous models. Additionally, due to the size of our dataset, the super-tree methodology involved several weeks of computation time to construct a single super-tree, which also impeded us of obtaining branch split support from bootstrap analysis. Given the low success of our attempts to construct super-tree phylogenies, the long computational time involved and the impossibility of obtaining bootstrap support for branch splits, we decided to focus our efforts in using tree and data heterogeneous models with the super-matrix approach.

Taxonomy

After we got our final super-matrix, we proceeded to compare it with other recent fungal taxonomy research in order to check whether its similar to other recent fungi topologies.

Initially, we sought to compare our phylogeny with other phylogenies that intended to provide a general structure of the fungal taxonomy. However, access to the phylogenies in an appropriate format for comparison was not possible. Still, we can compare the general structure of our phylogenetic tree with such phylogenies.

First, we compared our tree with the topology that *Ebersberger et al* (59) did with the objective of providing with a general taxonomy for the fungi to be used as a backbone for

future works to expand on. Even though our tree has many more taxa than this one the both share a similar structure when it comes to the big clade splits and *Ascomycota,* showing the same groupings of the branches. There are some minor differences in *Basidiomycota* like *Gloeophyllum* appearing in a deeper branch than *Heterobasidion* in our tree and that *Schizophyllum* and *Pleurotus* are in different clades in our tree. However, these differences seem to be due to varying number of taxa in these positions of the tree between the two phylogenies. In the first case the branch of interest in our tree is very populated in contrast to *Ebersberger's* phylogeny. On the contrary, in the second case there are considerably more taxa in *Ebersberger's* phylogeny branch than in our phylogeny's branch. Hence, the difference between number of taxa in these branches between phylogenies may explain the distinct structure. Despite these small differences both trees are similar on a general level and agree in most of the major groups.

Another study that focuses in the distribution of thermophilic fungi through the tree has also helped us to assert the general accuracy of our tree (60). This paper also supports our tree topology in the species they both share, being a perfect match in most groups except for some small discrepancies in Basidiomycota. Particularly *Heterobasidion annosum* is placed differently in both trees, next to A*garicales* in this paper and in a sister branch in our tree. This could be also due to the big difference in number of taxa in this part of the tree, since they just use 12 taxa for the Agaricomycetes while our tree has approximately 100 taxa which leads to the two separate clades within the group. Apart from this part of the tree the rest of the topology is similar in both trees, from Ascomycota to the early branches and the rest of the Basidiomycota.

After using this taxonomy to perform a rough comparison of our tree general structure, we proceeded to compare our tree with several other studies by using nodal distance and split distance measurements provided by TOPD/FMTS, which can be seen in Table 1.

The first study we used for comparison with our phylogeny and the largest one, focusing on the phylogeny of the Basidiomycota, is a six-gene phylogeny by *Zhao et al* (61). A tanglegram representing the comparison of both phylogenies can be seen in Figure 20, with taxa that are shared between both phylogenies indicated by lines. Both phylogenies are in agreement for most part of the tree, specially in the Agaricomycotina subphylum, found in the top part of the phylogenies in Figure 20, which is in complete agreement between both trees. Nevertheless, disagreements between the structure of the tree can be found in the bottom part of the phylogenies, which represents the subphylums of Basidiomycota,

Pucciniomycotina and Ustilaginomycotina, deeper divergent clades in the Basidiomycota. The disagreement between both trees comes from the placement of these subphylums, where in our phylogeny Ustilaginomycotina is found as a sister clade to Agaricomycotina, in *Zhao et al* phylogeny Pucciniomycotina is found as the sister clade to Agaricomycotina. The branch split between the Ustilaginomycotina and Agaricomycotina clades in our phylogeny is not robustly supported (58%) and the groupings proposed by *Zhao et al* phylogenetic hyphotesis is has additional support from recent literature (62). Moreover, the reduced tree and data heterogeneous phylogeny recovered in our study is also in agreement with *Zhao et al* hypothesis, as can be observed in Figure 16. Therefore, we can assume that the branch split between Pucciniomycotina and Ustilaginomycotina proposed by the complete phylogeny hypothesis is incorrect. Other than this discrepancy, both phylogenies are in agreement with the remaining structure of the tree.

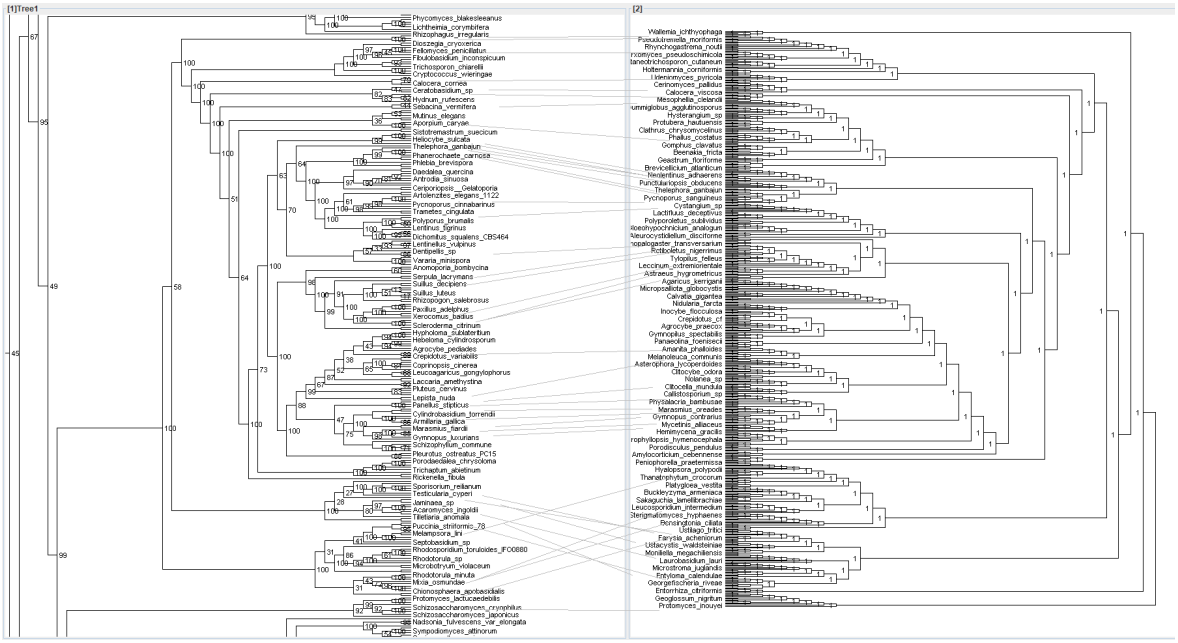

Figure 20: Tanglegram of the complete dataset phylogenetic tree (left) compared with Zhao et al six-gene phylogeny of the Basidiomycota (right). The project tree is focused on the Basidiomycota region. Bootstrap values in the project tree are indicated at branch splits.

The rest of the phylogenetic comparisons are focused around smaller phylogenies that are focused in recovering the placement of specific orders, so they encompass smaller sections of the fungal taxonomy. The next study used for the assessment of our phylogeny is a phylogeny proposed by *Davis et al (54)*, focused on the resolution of the *Zoopagales* order while using a similar maximum-likelihood approach to the analysis conducted in this

project but with a different dataset. The tanglegram comparing both phylogenies can be observed in Figure 21. There are disagreements in 3 out of the 22 taxa that are commonly shared bewteen both phylogenies, and all disagreeing taxa belong to the order *Mucorales*. Of the 3 disagreeing taxa, 2 taxa show low branch split support in our phylogenetic hypothesis, *Phycomyces blakesleeanus* and *Hesseltinella vesiculosa* (57% and 55% bootstrap support respectively), so it is likely that the branch split in our phylogeny for these species are incorrect. On the contrary, the remaining taxon, *Lichtheimia corymbifera,* is robustly supported in our phylogenetic hypothesis. This last taxon is part of a clade that is more populated in our phylogeny than in *Davis et al* phylogeny. Therefore, this increased taxa density in our phylogeny's branch could mean that our model is able to recover a well supported clade.

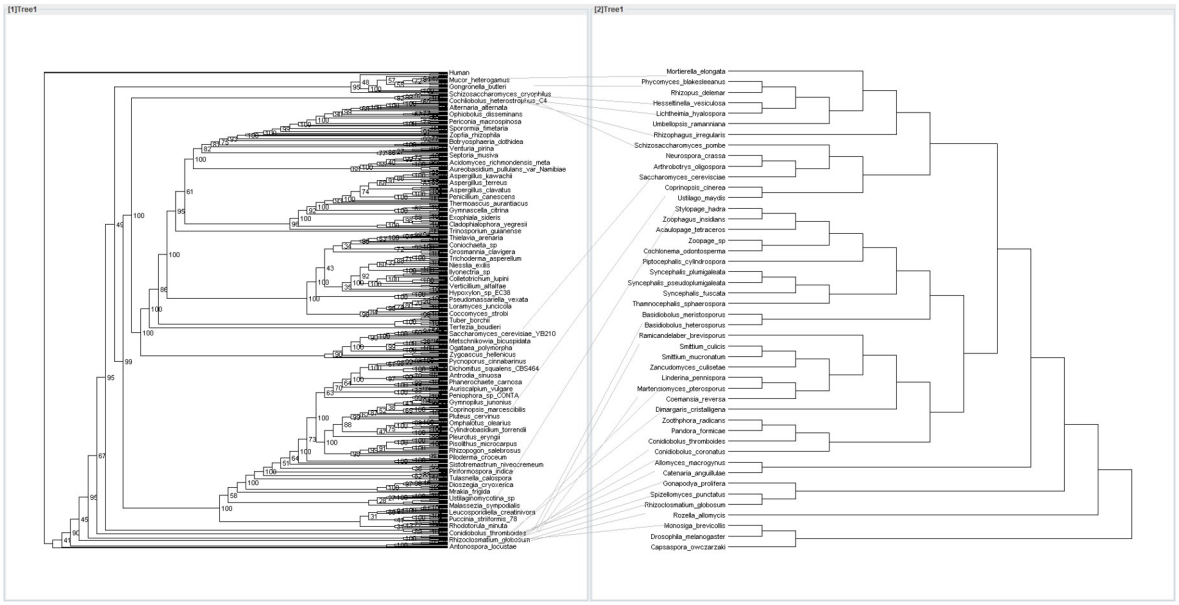

Figure 21: Tanglegram of the complete dataset phylogenetic tree (left) compared with Davis et al phylogeny (right). Bootstrap values in the project tree are indicated at branch splits.

Afterwards, we proceeded to compare our tree with a phylogeny focused in the Gloeophyllales order in the Agaricomycetes class proposed by *Chen et al (55)*. The tanglegram representing this comparison can be seen in Figure 22. In this case, 16 taxa are shared between both phylogenies, three of which are in disagreement. The first source of disagreement is found in the placement of two taxa, *Punctularia* and *Trametes,* within the Agaricomycetes class. This disagreement could be explained by the difference in representation in the Gloeophyllales order, which has barely any representation in our dataset and is evidently well represented in *Chen et al* phylogeny. As for the last disagreeing taxa, *Heterobasidion* (a member of the *incertae sedis* groups within the Agaricomycetes class), can be explained by the opposite phenomenon, where the

surrounding clades are better represented in our phylogeny, particularly other *incertae sedis* groups. Therefore, our phylogenetic hypothesis can provide a more robust placement for *Heterobasidion*. Lastly, the general structure of both phylogenies is similar despite the disagreeing taxa.

The last comparison used to validate our tree was carried out with the phylogeny proposed by *De Crop et al* (56), focused around the resolution of the Russulales order and the tanglegram comparing both phylogenies can be seen in Figure 23. The disagreements between both phylogenies are of similar nature with *Chen et al* comparison, where there is a difference between the population of the clades where the disagreeing taxa are located. Specifically for this comparison, all the disagreeing taxa are part of the Mucorales order. As an example, one of the disagreeing taxa is the *Mucor* family, which is represented by three members in our tree and only one member in *De Crop et al* phylogeny. The rest of the disagreeing taxa have similar issues, being part of well represented clades in our phylogeny and comparatively less dense clades in *De Crop et al* phylogeny. Still, both trees nodal and split distances are low, so the general structure of the trees is mostly in agreement.

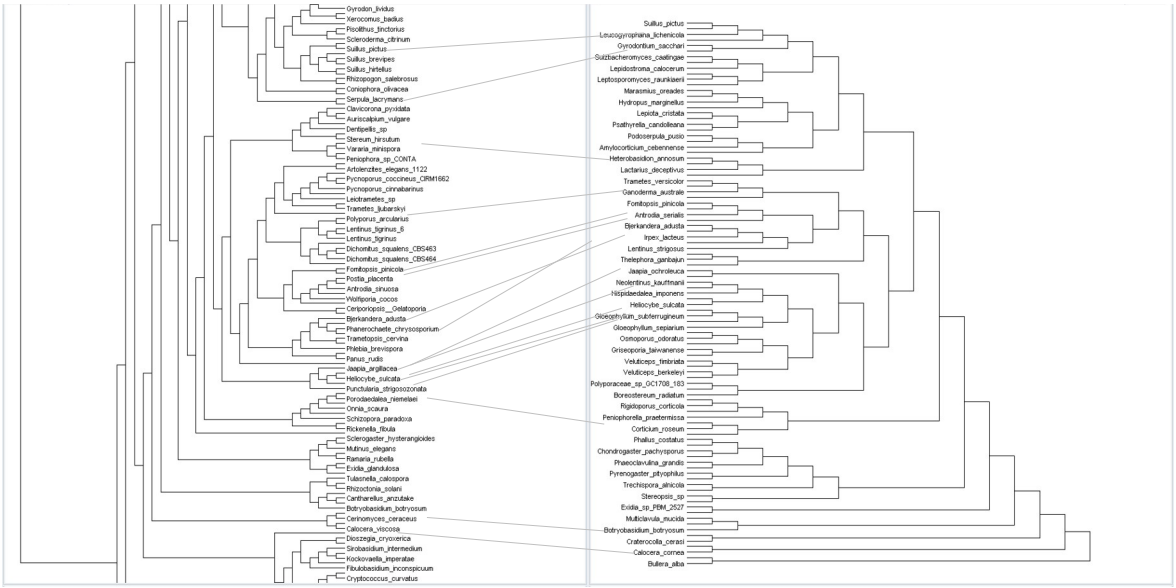

Figure 22: Tanglegram comparing this project phylogenetic tree (left) with Chen et al phylogeny (right). The project tree is focused on the Basidiomycota division. Bootstrap values in the project tree are indicated at branch splits.

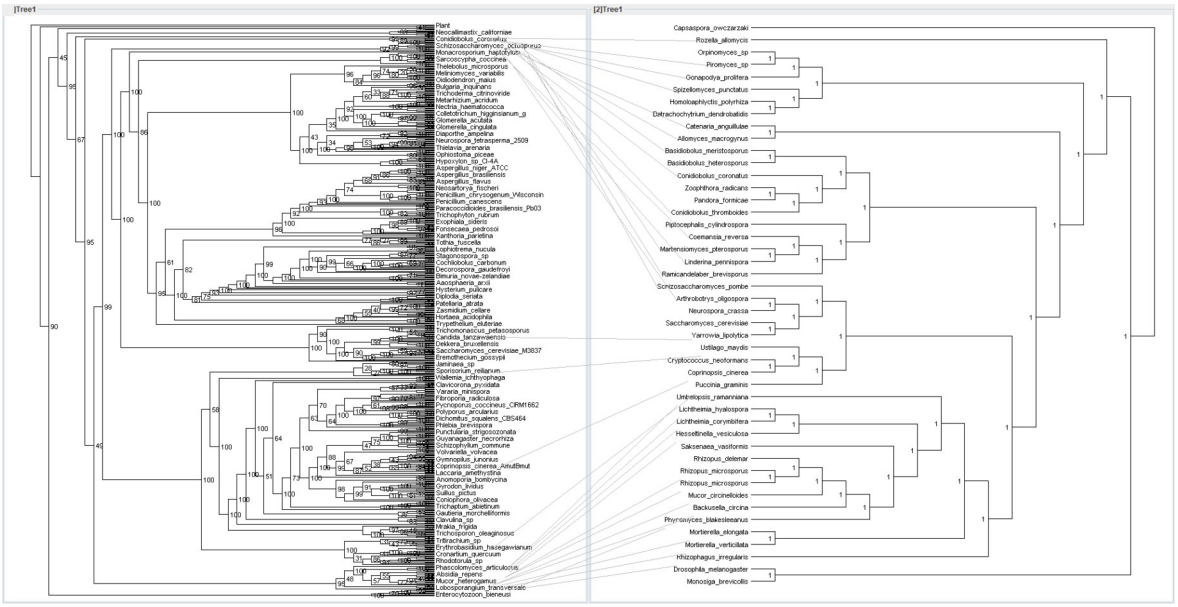

Figure 23: Tanglegram comparing this project phylogenetic tree (left) with De Crop et al phylogeny (right). Bootstrap values in the project tree are indicated at branch splits.

To conclude, our proposed phylogenetic hypothesis general structure is mostly in agreement with previously proposed phylogenies in the literature, but there are a number of branch splits with low support that are probably incorrect in our phylogeny. Many of the disagreements between phylogenies could be explained due to the different density of the clades where the disagreeing taxa are located, but for the most part our phylogeny is able to procure an insight in how the smaller groups of fungi are evolutionary related. Still, further research is needed to help with the resolution of poorly supported branch splits or clades with low population.

Models

In the initial steps of our phylogenetic analysis, the maximum-likelihood models where agreeing in the same topology when making the tree in different runs before the out-groups were included in the alignment. However, when we included the out-groups in the analysis we noticed the shortcomings of the maximum-likelihood models when dealing with the early groups of the fungi, especially Microsporidia whose rate of evolution is different when compared with the rest of the tree. The maximum-likelihood models would often split the out-groups, not agreeing in any particular topology and usually poorly supported. Even the more complex maximum-likelihood models that account in some way for

different rates of evolution in different parts of the tree were not able to provide us with a well supported phylogeny.

In order to properly account for the phylogenetic signals present in our dataset we had to resort to heterogeneous models. Heterogeneous models can account for different rates of gene or protein evolution, which allows the model to differentiate between more conserved proteins and rapidly evolving proteins, and different rates of evolution across the tree, which allows the model to account for differing substitution rates in homologous genes or proteins in different species. When we used them with the reduced dataset we were able to obtain a topology which placed the root reliably in repeated reconstructions of the fungal phylogeny. We used two different models for this purpose, a data heterogeneous model, which was in agreement with the general topology of maximum-likelihood models, and a tree and data heterogeneous model.

The tree and data heterogeneous model has shown topological discrepancies compared to less parameter-rich models, specifically the placement of Microsporidia as a sister branch of the Dikarya subkingdom, and Zoopagomycota and Mucoromycota being a monophyletic group. We have thoroughly tested all the models and even if the data show that all the phylogenetic hypotheses can plausibly explain the data we observe, the tree and data heterogeneous model is significantly better than the data heterogeneous model, and both of these models have much better likelihood when compared to homogeneous models, so heterogeneous models are able to explain the data observed in our dataset better. One thing to notice is that in every model but the data and tree heterogeneous model all long branches of the tree appear clustered together (in our tree this is the out-groups, *Rozella*, and Microsporidia). Therefore, we propose that homogeneous and data heterogeneous models could not solve the long branch attraction problem that adding a branch with high substitution rates like Microsporidia brings, and that only when we used the data and tree heterogeneous model we could solve this problem. This open the possibility that Microsporidia are not early fungi and they instead share a common ancestor with Dikarya that specialised in intracellular parasitism. Convergent evolution towards intracellular parasitism and genome reduction could also explain why they are so often placed together with *Rozella* and other Cryptomycota, since they are also obligate intracellular parasites.

Furthermore, we have validated our methodology by construction a heterogeneous maximum-likelihood tree based on the phylogeny by *McCarthy* and *Fitzpatrick,* which can be seen in Figure 24. Our validation phylogeny, seen in Figure 18, shares a similar general structure with *McCarthy* and *Fitzpatrick.* Both phylogenies are in agreement with the

placement of the major groups of the fungi, with groups like Basidiomycota being identical in both phylogenies. There are some minor discrepancies in some smaller groups, i.e. *Magnaporthe grisea* is a sister group of *Phaeoacremonium aleophilum* in their phylogeny, while in our phylogeny *Magnaporthe grisea* is a sister group of *Ophiostoma piceae* and *Phaeoacremonium aleophilum* shares a common ancestor with them. The most blatant difference between our phylogenies is the placement of *Rhizophagus irregularis* and *Gonapodya prolifera*. *Rhizophagus irregularis* is next to Dikarya in their phylogeny, while in our phylogeny it appears within Mucoromycota. *Gonapodya prolifera* appears as a sister branch to Neocallimastigomycota, while in our tree it appears within Chytridomycota. However, the placement of these species is recovered as with our validation phylogeny by some of the other phylogenetic approaches used by *McCarthy* and *Fitzpatrick* in their study. Overall, both trees are structurally similar with few minor differences. The similarity between our validation phylogeny and *McCarthy* and *Fitzpatrick* phylogeny indicates that our phylogenetic methodology can be successfully applied to smaller and less complex datasets of fungi and still recover a similar phylogeny compared with what was already published in the literature.

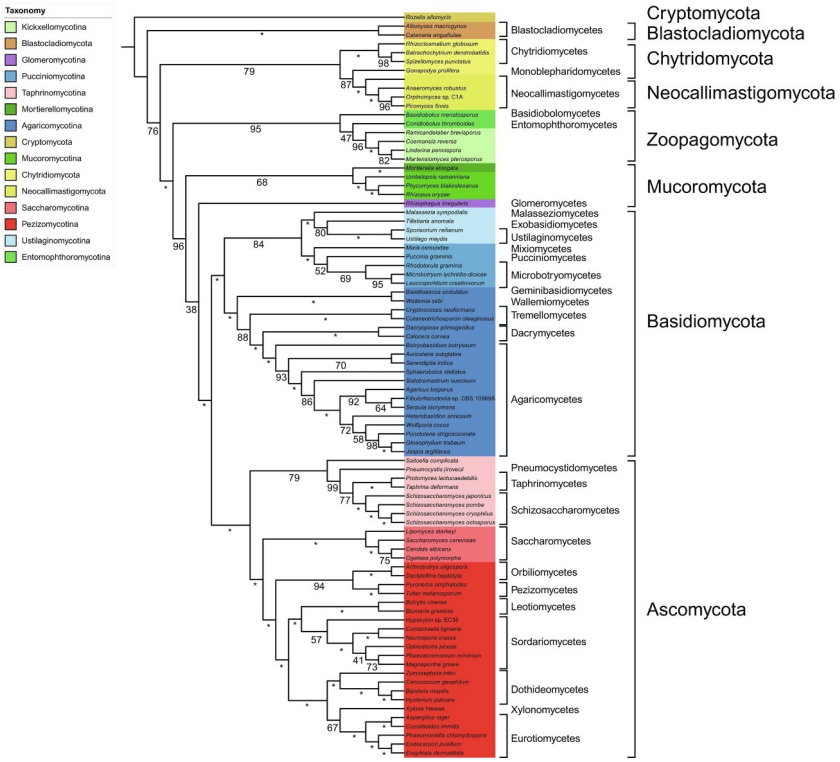

Figure 24: Maximum-likelihood phylogeny of 84 fungi with bootstrap support reconstructed by McCarthy and Fitzpatrick (49).

Conclusion

By using a combination of data heterogeneous maximum-likelihood models and data heterogeneous and tree and data heterogeneous Bayesian models we were able to construct a fungal phylogeny which has reliably placed the major groups of the fungi in the phylogeny. Many other small groups were also placed in the phylogeny in accordance with recent literature. Tree and data heterogeneous models have helped us resolve deep phylogenetic trees with very differing substitution rates and opened the possibility that Microsporidia are not early fungi but instead form a monophyletic group together with Dikarya, and that Zoopagomycota and Mucoromycota are also a monophyletic group.

Bibliography

1. M. J. Swift, The ecology of wood decomposition. *Sci. Prog. 1933-* **64**, 175–199 (1977).

2. J. W. Taylor, M. L. Berbee, Dating divergences in the Fungal Tree of Life: Review and new analyses. *Mycologia* **98**, 838–849 (2006).

3. J. E. Stajich, *et al.*, The Fungi. *Curr. Biol.* **19**, R840–R845 (2009).

4. , Home - Onnia scaura P-53A v1.0 (June 8, 2020).

5. M. Blackwell, D. S. Hibbett, J. W. Taylor, J. W. Spatafora, Research Coordination Networks: A phylogeny for kingdom Fungi (Deep Hypha). *Mycologia* **98**, 829–837 (2006).

6. D. J. McLaughlin, D. S. Hibbett, F. Lutzoni, J. W. Spatafora, R. Vilgalys, The search for the fungal tree of life. *Trends Microbiol.* **17**, 488–497 (2009).

7. S. Federhen, The NCBI Taxonomy database. *Nucleic Acids Res.* **40**, D136–D143 (2012).

8. S. Capella-Gutiérrez, M. Marcet-Houben, T. Gabaldón, Phylogenomics supports microsporidia as the earliest diverging clade of sequenced fungi. *BMC Biol.* **10**, 47 (2012).

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

10. D. Corsaro, *et al.*, Molecular identification of Nucleophaga terricolae sp. nov. (Rozellomycota), and new insights on the origin of the Microsporidia. *Parasitol. Res.* **115**, 3003–3011 (2016).

11. B. Han, L. M. Weiss, Microsporidia: Obligate Intracellular Pathogens within the Fungal Kingdom. *Microbiol. Spectr.* **5** (2017).

12. P. J. Keeling, M. A. Luker, J. D. Palmer, Evidence from Beta-Tubulin Phylogeny that Microsporidia Evolved from Within the Fungi. *Mol. Biol. Evol.* **17**, 23–31 (2000).

13. T. Y. James, *et al.*, Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature* **443**, 818 (2006).

14. M. D. M. Jones, *et al.*, Discovery of novel intermediate forms redefines the fungal tree of life. *Nature* **474**, 200–203 (2011).

15. , 1000 Fungal Genomes Project. *1000 Fungal Genomes Proj.* (October 24, 2016).

16. N. Corradi, P. J. Keeling, Microsporidia: a journey through radical taxonomical revisions. *Fungal Biol. Rev.* **23**, 1–8 (2009).

17. P. J. Keeling, N. M. Fast, Microsporidia: Biology and Evolution of Highly Reduced Intracellular Parasites. *Annu. Rev. Microbiol.* **56**, 93–116 (2002).

18. T. Y. James, *et al.*, Shared Signatures of Parasitism and Phylogenomics Unite Cryptomycota and Microsporidia. *Curr. Biol.* **23**, 1548–1553 (2013).

19. S. Nakjang, *et al.*, Reduction and Expansion in Microsporidian Genome Evolution: New Insights from Comparative Genomics. *Genome Biol. Evol.* **5**, 2285–2303 (2013).

20. S. Klopfstein, T. Massingham, N. Goldman, More on the Best Evolutionary Rate for Phylogenetic Analysis. *Syst. Biol.* **66**, 769–785 (2017).

21. J. Bergsten, A review of long-branch attraction. *Cladistics* **21**, 163–193 (2005).

22. H. Brinkmann, M. van der Giezen, Y. Zhou, G. P. de Raucourt, H. Philippe, An Empirical Assessment of Long-Branch Attraction Artefacts in Deep Eukaryotic Phylogenomics. *Syst. Biol.* **54**, 743–757 (2005).

23. S. C. K. Straub, *et al.*, Phylogenetic signal detection from an ancient rapid radiation: Effects of noise reduction, long-branch attraction, and model selection in crown clade Apocynaceae. *Mol. Phylogenet. Evol.* **80**, 169–185 (2014).

24. N. Lartillot, H. Brinkmann, H. Philippe, Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol. Biol.* **7**, S4 (2007).

25. P. G. Foster, T. Schultz, Modeling Compositional Heterogeneity. *Syst. Biol.* **53**, 485–495 (2004).

26. R. Moran, C. Morgan, M. O'Connell, A guide to phylogenetic reconstruction using heterogeneous models—a case study from the root of the placental mammal tree. *Computation* **3**, 177–196 (2015).

27. I. Foster, "Globus Toolkit Version 4: Software for Service-Oriented Systems" in *SpringerLink*, (Springer Berlin Heidelberg, 2005), pp. 2–13.

28. M. Ashburner, *et al.*, Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).

29. M. Kanehisa, S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

30. A. Mitchell, *et al.*, The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.*, gku1243 (2014).

31. R. L. Tatusov, *et al.*, The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).

32. T. N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786 (2011).

33. O. H. Cisse, J. E. Stajich, FGMP: assessing fungal genome completeness and gene content. *bioRxiv*, 049619 (2016).

34. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

35. S. McGinnis, T. L. Madden, BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* **32**, W20–W25 (2004).

36. S. Van Dongen, Graph clustering by flow simulation [phd thesis].[Utrecht (The Netherlands)]: University of Utrecht (2000).

37. K. Katoh, H. Toh, Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.* **9**, 286–298 (2008).

38. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinforma. Oxf. Engl.* **25**, 1972–1973 (2009).

39. D. Darriba, G. L. Taboada, R. Doallo, D. Posada, ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinforma. Oxf. Engl.* **27**, 1164–1165 (2011).

40. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

41. W. P. Maddison, Gene Trees in Species Trees. *Syst. Biol.* **46**, 523–536 (1997).

42. J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).

43. J. Felsenstein, Confidence limits on phylogenies: an approach using the bootstrap. *evolution* **39**, 783–791 (1985).

44. R. Lanfear, P. B. Frandsen, A. M. Wright, T. Senfeld, B. Calcott, PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. *Mol. Biol. Evol.* **34**, 772–773 (2017).

45. P. Puigbò, S. Garcia-Vallvé, J. O. McInerney, TOPD/FMTS: a new software to compare phylogenetic trees. *Bioinforma. Oxf. Engl.* **23**, 1556–1558 (2007).

46. W. Piel, *et al.*, *TreeBASE v. 2: A Database of Phylogenetic Knowledge. e-BioSphere 2009* (London, 2009).

47. R. A. Vos, *et al.*, NeXML: rich, extensible, and verifiable representation of comparative data and metadata. *Syst. Biol.* **61**, 675–689 (2012).

48. D. F. Robinson, L. R. Foulds, Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).

49. C. G. P. McCarthy, D. A. Fitzpatrick, “Chapter Six - Multiple Approaches to Phylogenomic Reconstruction of the Fungal Kingdom” in *Advances in Genetics*, Fungal Phylogenetics and Phylogenomics., J. P. Townsend, Z. Wang, Eds. (Academic Press, 2017), pp. 211–266.

50. C. J. Creevey, J. O. McInerney, Clann: investigating phylogenetic information through supertree analyses. *Bioinforma. Oxf. Engl.* **21**, 390–392 (2005).

51. J. Paszek, P. Górecki, Genomic duplication problems for unrooted gene trees. *BMC Genomics* **17**, 15 (2016).

52. F. Thomarat, C. P. Vivares, M. Gouy, Phylogenetic analysis of the complete genome sequence of Encephalitozoon cuniculi supports the fungal origin of microsporidia and reveals a high frequency of fast-evolving genes. *J. Mol. Evol.* **59**, 780–791 (2004).

53. N. Zhang, S. Zhao, Q. Shen, A six-gene phylogeny reveals the evolution of mode of infection in the rice blast fungus and allied species. *Mycologia* **103**, 1267–1276 (2011).

54. W. J. Davis, *et al.*, Genome-scale phylogenetics reveals a monophyletic Zoopagales (Zoopagomycota, Fungi). *Mol. Phylogenet. Evol.* **133**, 152–163 (2019).

55. C.-C. Chen, *et al.*, Phylogenetic placement of Paratrichaptum and reconsideration of Gloeophyllales. *Fungal Syst. Evol.* **5**, 119–129 (2020).

56. E. De Crop, *et al.*, A multi-gene phylogeny of Lactifluus (Basidiomycota, Russulales) translated into a new infrageneric classification of the genus. *Persoonia* **38**, 58–80 (2017).

57. S. V. Melnikov, *et al.*, Muller's Ratchet and Ribosome Degeneration in the Obligate Intracellular Parasites Microsporidia. *Int. J. Mol. Sci.* **19** (2018).

58. J. Barandun, M. Hunziker, C. R. Vossbrinck, S. Klinge, Evolutionary compaction and adaptation visualized by the structure of the dormant microsporidian ribosome. *Nat. Microbiol.* **4**, 1798–1804 (2019).

59. I. Ebersberger, *et al.*, A Consistent Phylogenetic Backbone for the Fungi. *Mol. Biol. Evol.* **29**, 1319–1334 (2012).

60. I. Morgenstern, *et al.*, A molecular phylogeny of thermophilic fungi. *Fungal Biol.* **116**, 489–502 (2012).

61. R.-L. Zhao, *et al.*, A six-gene phylogenetic overview of Basidiomycota and allied phyla with estimated divergence times of higher taxa and a phyloproteomics perspective. *Fungal Divers.* **84**, 43–74 (2017).

62. H. Mao, H. Wang, Resolution of deep divergence of club fungi (phylum Basidiomycota). *Synth. Syst. Biotechnol.* **4**, 225–231 (2019).

Exploring fungal ecotype associations using convergent evolution and co-occurrence networks

Ignacio Riquelme-Medina[1], James O. McInerney[1,2,#]

[1.] Division of Evolution and Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester Academic Health Science Centre, Manchester M13 9PL, United Kingdom.

[2.] . School of Life Sciences, The University of Nottingham, Nottingham NG7 2UH, United Kingdom.

[#] To whom correspondence should be addressed.

Phone: +44 (0)161 2751843

Email: james.mcinerney@manchester.ac.uk; james.mcinerney@nottingham.ac.uk

Website: http://mcinerneylab.com/

Abstract

Convergent evolution is the process by which a similar trait emerges in two different species that do not share a recent ancestor. Convergent evolution is ubiquitous of life, appearing throughout all life domains. One group of organisms where convergent evolution events are likely to occurs is fungi, since they have adapted to the same habitats independently repeated times in their phylogeny.

In this project we have used co-occurrence networks to detect convergent evolution events within the fungal kingdom. Co-occurence networks detect groups of genes or proteins that appear together more than it is expected by random chance, and that can usually be linked to particular functions or processes. To avoid detecting co-occurring events where a trait is shared between two species due to recent common ancestry, the fungal phylogeny has to be considered. Additionally, by investigating the species habitat and lifestyle distribution within a co-occurring set of proteins we can draw connections between the protein set and adaptation to a particular environment.

By using co-occurrence network to detect phylogenetically independent convergence events we were able to identify two co-occurring sets of proteins that can be linked to particular lifestyles, parasitism and formation of mycorrhiza. Moreover, we detected a large co-occurring set of proteins that can be linked to the fungal core genome, proteins shared by most species within the fungal kingdom that are needed to carry out basic life functions.

Introduction

Convergent evolution can be defined as the process by which several lineages develop the same adaptation, from different ancestral starting points. Convergences in genome evolution can happen for entirely independent reasons or there might be a common process based on adaptation to similar environments or similar ways of life, for these different lineages (1). Convergent evolution is a widespread phenomenon through the history of life and we can find evidence in all domains on life, from molecular to phenotypic level (2–7). This process has been used as a possible explanation for homoplasy, where species that do not share a recent common ancestor evolving, nonetheless, to where they have similar traits. Certain traits have evolved consistently and independently in very differing organisms, some of the well-know include the eyes (8) and complex brains with high intelligence (9), both of which have evolved independently several times throughout the animal kingdom. Eyes develop similar structures independently of the lineage in which they appear, like lenses, and the same happens with high intelligence, which is linked to multimodal centres in the brain structure (9). The process of convergence in evolution can be the result of genes undergoing similar changes in distantly-related species but where these species experience similar evolutionary pressure. Therefore, we can detect convergent evolution at the molecular level in gene and protein sequences since the sequences will experience similar changes to adapt to the same evolutionary pressure (10). Sequence similarity has been measured and used in several methods to detect and quantify convergent evolution along with phenotypic methods (11).

One of the groups where convergent evolution is likely to manifest itself is the fungal kingdom. Fungi are characterized by being very variable, morphologically simple heterotrophic organisms (typically filamentous or unicellular) with relatively small genomes and they commonly have a diploid phase and haploid phase in their life cycle (12). Fungi are considered variable organisms because they have very different ways of life; they can live as saprotrophs, mutualists like in mycorrhizal communities, parasites of plants, animals and other fungi. Many of them are opportunist parasites so they can go from living in soil to infecting other organisms given the chance. These different ways of life are not particular to any specific clade of fungi, but they have appeared repeatedly throughout the fungal phylogeny (12). For example, amongst the species of the order *Hymenochaetales* we can find saprotrophs, strong and weak parasites, and mycorrhizal species. Furthermore, fungi make good model organisms and are relatively easy to grow

and study, for which there is extensive research done in many fungal proteins and pathways. In addition, many fungal species have a huge impact on humans and human activities, both positive and negative. Fungi can be human pathogens and also infect many crop plants, while also providing significant benefits in certain industries like in fermentation or research (13–16). Recently there has also been an effort to sequence a wider variety of fungal species like the 1000 Fungal Genomes Project (17), providing with easily accessible genomes along with an extensive amount of functional annotations for each species genes, proteins and pathways. On account of repeated fungal adaptations to the same environments throughout their phylogeny, we can hypothesise that there is a high chance for convergent evolution events within the fungi. The recent amount and quality of fungal genomes that have become available means that fungi make ideal candidate organisms to test convergent evolution hypotheses.

One of the ways we could use to detect convergence is by using association networks, which detect gene co-occurrence or co-avoidance through various genomes (18, 19). These networks are similar to genome-wide association studies in their purpose, where the objective is to link particular phenotypes with particular genotypic variants. In association networks it is assumed that genes that are found together more often than expected by random chance alone, might have a synergistic effect on the phenotype. We would therefore explore whether these significant co-occurrences are linked to similar functions. In a similar way, genes that avoid each other more than expected by chance could have negative effects on each other if they were cloned into the same genome. Genes that tend to avoid being in the same genome would include genes whose combined protein products would induce toxicity, or indeed might encode almost identical functions and therefore would produce redundant effects that may be costly for the organism. Co-incidence networks have been successfully used previously to detect functionally associated proteins (18, 20, 21) and to link the presence of certain functions with environments (22, 23) in prokaryotes. However, considering co-occurrence or co-avoidance of genes alone is not enough to detect homoplasy events in the fungal phylogeny. To detect homoplasy we have to consider the phylogenetic relationships of every fungus in the dataset, in order to identify instances where co-occurrence or co-avoidance occurs because the species share a recent common ancestor. In this case, the association of genes could be explained by the acquisition or loss of the genes in a common ancestor and the subsequent transmission of this trait to descendants from this ancestor. So, to detect homoplasy in the fungal phylogeny we have to find genes that co-occur or co-avoid each other more than we would expect by random change in species that do not share a recent common ancestor. One of

the recent packages of software that can accomplish this task is CoinFinder (24). CoinFinder is able to detect homoplasy events in a gene or protein network (usually the result of a sequence similarity analysis using a program such as BLAST (25)) and a phylogeny of the species present in the network. CoinFinder is scalable so it can be used on large datasets and allows to detect both co-occurrence and co-avoidance in a given network. The output of the program is several phylogenetically independent co-occurrence or co-avoidance components of the network, which can be later linked to particular functions or environments by investigating the genes present in every component. Although CoinFinder was originally designed to identify association networks in prokaryote pan-genomes, it can also be used for eukaryote genomes. Other alternatives that can also detect phylogenetically independent gene associations like Copap (26) have been considered, however Copap has some software limitations and is computationally more expensive, which makes it unsuitable for large datasets. Copap is also unable to detect co-avoidance events. There have also been some previous attempts to use convergence and divergence to study the evolution of fungal pathogenicity (27). However, this study could not take in account if the divergence and convergence was due to shared common ancestry, an issue that should be solved with the methods that will be used in this project.

Methods

We first obtained our data by downloading all the publicly available genome protein sequences and their associated metadata from the 1000 Fungal Genomes Project (FGP) database (17) by using the Globus tool (28) (Genomes = 690, Accessed 10/2016). To asses the quality of the genomes we used the Fungal Genome Mapping Project (FGMP) (29), which is a program that checks whether a list of previously determined conserved and ultra conserved fungal proteins are present in the genome. The tool then reports a percentage of proteins that are present in each genome, which correlates with the quality of the sequencing because most of these proteins should be present in every genome. Genomes that had less than 75% of the conserved gene set were considered of poor quality and were not used further in the analysis. However, the genomes of some fungal species that have undergone a genome reduction process, and thus are much less likely to have the conserved proteins present, were exempt from this quality assessment step. The quality assessment step left us with a total of 671 genomes and 8,410,690 unique proteins. But firstly, we decided to test CoinFinder by using a smaller, more manageable and well supported fungal phylogeny, which does not include the problematic groups present in our complete fungal phylogeny, as a primer dataset. This primer dataset was based on the phylogeny proposed by *McCarthy* and *Fitzpatrick (30)*, and is formed of 84 genomes obtained from the FGP similarly to our fungal dataset. However, some of these genomes used in this study did not pass the quality control step, so when possible they were substituted by another species in the same family. Specifically, *Zymoseptoria tritici* was changed for *Zymoseptoria ardabiliae, Candida albicans* for *Candida tanzawaensis, Microbotryum lychnidis-dioicae* for *Microbotryum violaceum* and *Rhizopus oryzae* for *Rhizopus microsporus.* Otherwise, Species that did not have a close relative where discarded from the analysis, specifically *Endocarpon pusillum, Orpinomyces sp. C1A,* and *Batrachochytrium dendrobatidis.* Instead of using an external source genome for *Allomyces macrogynus* that was used in the paper dataset, we used a sequence genome from the FGP instead. In the end, the total number of genomes used in the validation dataset was 81.

Next we proceeded to use the BLAST tool BLASTp v2.4.0 (25, 31) (e value = 1*e-6) by using the combined protein sequences of all 671 genomes to obtain a similarity measure between all proteins in our dataset. For the primer dataset, we extracted only the edges that include species in the 81 primer genomes from the BLAST output. The following steps were applied identically to both datasets. We used the clustering algorithm MCL 14-137

(32) in order to split the BLAST outputs into clusters of proteins that form communities of sequence similarity – a proxy for protein family identification. These clusters allow us to group proteins originating from different genomes into groups where the level of connectivity within the group is higher than connectivity from a member of the group to a non-member. MCL requires an inflation parameter, which determines how large or small the cluster will be after it partitions the original graph into connected components. The higher the value, the smaller the clusters. To optimise the inflation value for our dataset we used the MCL algorithm several times with different inflation values. Afterwards, we checked which inflation value was able to return complete clusters of ultra-conserved proteins (Ribosomal proteins were the type of proteins checked for cluster completion). We finally settled on an inflation value of 1.4, which was the highest inflation value where MCL returned clusters of ultra-conserved proteins consistently, giving a final number of 155,969 protein clusters for the primer dataset and 965,827 protein clusters for the complete dataset. For each dataset these protein clusters where named protein families, and each family is given a number in a descending order according to the number of proteins it includes in a descending order. E.g. the protein family that has the highest number of proteins is Family 1, the second largest is Family 2 and so on. We then constructed an edge list for each dataset by connecting genomes to protein families by using the proteins, e.g. if protein p from genome g was present in cluster c, we say that there is a connection between genome g and cluster c. Since clusters are groups of proteins with relatively high sequence similarity to one another, we can have a rough idea of the distribution of protein functions in the genomes.

Finally, we used CoinFinder (accessed 05/2020) in order to construct coincidence networks. CoinFinder takes an edge list and a phylogenetic tree as input and returns either a coincidence network or an avoidance network along with several figures and statistics about the network. For the primer dataset, CoinFinder was used with the previously created edge list for the primer dataset and the validation phylogenetic tree obtained in the previous chapter. As for the complete dataset, Coinfinder was used with the complete dataset edge list and the complete phylogeny of the 671 genomes obtained in the previous chapter. The following parameters were used to create the association networks: detect association, Bonferroni correction and Greater than alternative hypothesis. Also several filtering options that allow filtering of high and low abundance data as a percentage of the dataset were tested (from not filtering to different values for the high and low abundance filters). Finally, the cut-off value was set to 0.9 and 0.05 respectively, which filters the top 10% families with more members and the bottom 5% families with the lowest number of

members. These settings gave an output with the highest number of components. For the complete association network, we further analysed the resulting big component by splitting it into partitions of highly interconnected families by using the Louvain community detection algorithm implemented in NetworkX in python 3.7 (33). Since communities in networks tend to be related to specific functions or processes (34–36), we may be able to find some parts of the big component where a particular function is specifically present and find related functions or species where that function is more relevant.

Additionally, a literature search was carried out for every species in the filtered genomes to classify species according to five categories: their way of life (Biofilm, Lichen, Mycorrhizal, Saprotrophs, Pathogens or Unknown), their main habitat (Animal, Animal organic matter, Fungi, Gut, Plant, Plant organic matter, Soil, Substrate (mostly species used in fermentation), Water and Unknown), whether they can infect other organisms, and whether they are considered or not to be extremophiles. For each species a small note was made with information that did not fit with this categorization but could be important to determine the ecotype (e.g. if saprotroph fungi that decompose wood produce white or brown rot). This metadata is not intended to be fully accurate, since many species have multiple ways of life and some species have yet to be studied. However, it should still provide us with an initial classification of our results that will allow us to discern the components that are potentially more consequential for our research.

Results

CoinFinder failed to retrieve any significant associations for either protein co-occurrence or protein co-avoidance for the primer dataset, so no components were found.

As for the CoinFinder output for the complete dataset we obtained a total of 19 co-occurring components. There is one big component that makes up most of the data (that encompasses 9897 families) and 18 small components (that encompass between 13 and 2 families each). The small components can be seen in Figure 25. As for the co-avoidance network, CoinFinder only detected one big co-avoiding component.

Figure 25: Heat-map of the family presence of the smaller 18 components from CoinFinder output. The phylogenetic tree is displayed on y axis and the families on the x axis. Each colour represents a component and every coloured square indicates the presence of a protein from a family in the corresponding species from the phylogenetic tree.

Each of the families present in the components was assigned with the metadata previously obtained from the FGP, in order to check the functions and pathways that are present in each component. In this way we can try to link the components to specific processes or ideally to particular habitats of the fungi.

Finally, the components obtained from CoinFinder were also combined with with the information obtained in the species literature review, so we could check the main ways of life, habitat and infectivity of species included in each component. This information is summarised in Figure 26.

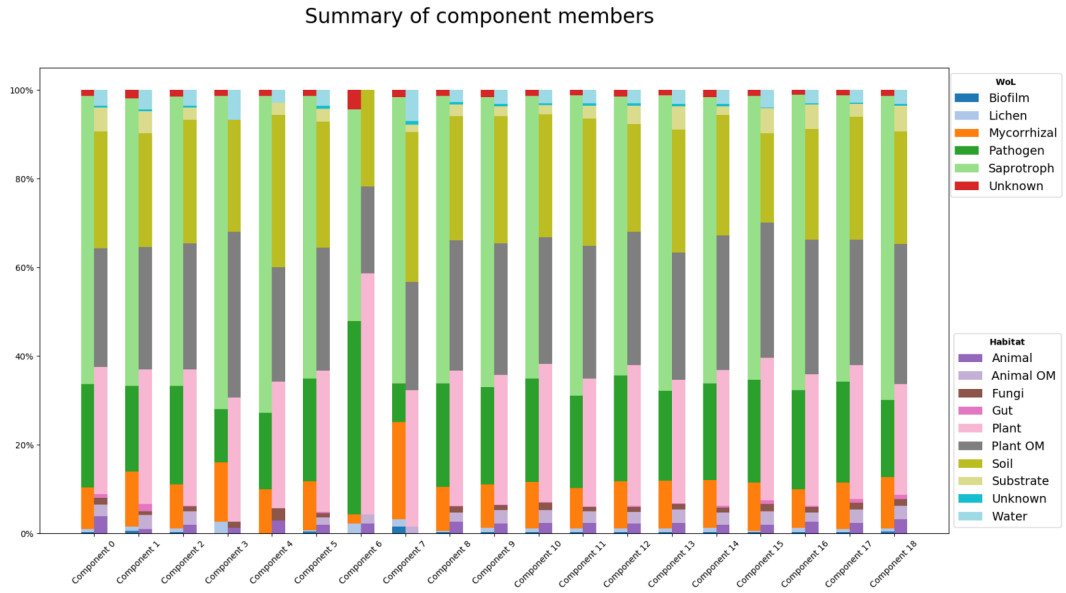

Figure 26: Summary of each component percentage of species for a particular Way of Life (WoL, left bars) and main habitat (right bars) of the total species present in each component.

Additionally, we calculated the Fold Change (FC) of each component respectively to each category average for all the components. This step was done to detect components with a higher or lower presence of species in particular categories when compared to the other components. Most of the differing groups are ways of life or habitats that have small presence in the components and are usually absent in some of the components (like Biofilm or Gut, which make up less than 1% of the species present in the components on average). However, component 6 and component 7 stand out because they have a very differing proportions in categories that encompass a big number of species. Component 6 Mycorrhizal has an FC of 0.2 and a Pathogen FC of 2.01; and component 7 Mycorrhizal has an FC of 2.06 and a Pathogen FC of 0.4. In other words, component 6 is enriched for pathogens and lacks mycorrhizal species when compared to the average, and component 7 is the opposite.

Discussion

In this paper we have tried to detect convergent evolution events in the fungal phylogeny by identifying groups of two or more genes that appear together in species not related by a recent common ancestor more than we would expect them to appear together randomly. In this way we should be able to identify genes that are required for particular phenotypes or for phenotypes that benefit each other, paying special attention to phenotypes that are specific to a habitat or a particular way of life. By using fungi, which have repeatedly adapted to similar environments and changed their ways of life several times in their phylogeny, we should increase the chances of identifying these events. We have used a smaller primer dataset as a test for the association analysis and established the need to use a big dataset to be able to discern significant protein-protein association interactions. It is likely that the number of genomes used in the validation dataset is too low to obtain any significance, as it has been tested that CoinFinder works better with bigger datasets, failing to detect any associations when the number of genomes is close to 50 (24). The number of genomes used in the primer dataset, 81, is close to this minimum, so this dataset might lack the necessary evidence to detect any significant associations. This lack of results indicated us that there more genomes are needed to discover the co-occurrence and co-avoidance patters of phylogenetically independent proteins in the fungi, so we proceeded to continue our analysis with the complete dataset. As a result of our analysis with the bigger dataset, we have obtained a total of 19 components of proteins that co-occur, separated in one big component that includes most co-occurring proteins, and 18 smaller components.

Components related to ways of life

Of the resulting components, the most interesting ones are component 6 and 7: the only components that have a discrepancy in the species distribution and could be more easily liked to a phenotype beneficial for a particular way of life. Component 6 is a component that is enriched for pathogenic species and has a low abundance of mycorrhizal species, having more than twice pathogenic species than the average of the components and five time less mycorrhizal species than other components. The distribution of species in this component could indicate that it is including mostly species that are pathogens or opportunistic pathogens, since there are many fungal species that are able to act as pathogen that spend most of their life cycles as saprotrophs. We can find two families in

this component: Family 13408 and family 16774. Family 13408 is a family of 60 proteins that are poorly characterized. There is only one protein that is characterized and has an entry in GO and EC databases, and it is DNA-directed RNA polymerase. Some other proteins have related orthologs in the KOG database, of which there are three translation initiation factors 4F, one large RNA-binding protein, one nucleolar GTPase/ATPase p130 and one collagen. The rest are either poorly characterized conserved proteins of unknown function or do not have known orthologs in the KOG database. Family 16774 is a small protein family with 37 members that only have uncharacterised hypothetical proteins in its members. Most of the proteins in Family 13408 are related to translation, since DNA-directed RNA polymerases, the RNA binding proteins, and the translation initiation factors 4F are directly involved in the translation process. The nucleolar GTPase/ATPase p130 is a protein that is related to mitosis and cell proliferation in humans (37, 38). This protein could therefore be related to an increased growth in fungi, which has been shown to be a strategy that fungi adopt during infection in several studies (39, 40). Finally collagen in fungi can be related to the formation of extracellular structures like fimbriae (41), which are structures involved in reproduction and pathogenicity composed by collagen. This family could then be related to cell proliferation and formation of extracellular structures and may be related to pathogenicity. However, due to the high number of uncharacterised proteins, especially in its co-occurring family 16774, it is impossible to have a more secure prediction of function or its relation to a particular way of life until the functions of these groups of proteins are better known.

Meanwhile Component 7 is similar to component 6 but in the opposite way: component 7 has a high abundance of mycorrhizal species while pathogens are under-represented when compared to the average of the species distribution of the components. In a similar way in which the distribution of species in component 6 could indicate that most of its species are pathogen or opportunistic pathogens, in component 7 the distribution of species can indicate that many species included in it are mycorrhizal or are able to form mycorrhizal associations. Many of the fungal species that form part of the mycorrhizal communities can also live freely in the soil as saprotrophs or have a short saprotrophic phase (42, 43). This component is composed of two protein families: family 5912 and family 13765. Family 5912 is a family of poorly characterized proteins. Out of the 405 proteins that are included in this family, only three have entries in the KOG database and only eight members have entries in the GO database. Of the proteins that are characterized in this family there seems to be four main functions: translation as part of the structure of ribosomes, transport (including ATP binding), biosynthesis of secondary metabolites as phosphopantetheine

binding related proteins and nucleic acid binding. Family 13765 is a smaller protein family but better characterized than family 5912. Of this family's 52 members, 20% have an entry in the KOG database and about half have associated terms in the GO database. The proteins in this family have functions in transcription, translation and post translation modifications, defence mechanisms (C-type lectins), as chitinases, metabolism of complex carbohydrates (starch and sucrose) and some transport and signal transduction. Some of these functions in family 13765 could be related to mycorrhizal activity. Lectins are proteins used in carbohydrate recognitions that many organisms use to detect pathogens but it can also be used to recognize mutualist organisms and has been proved to be used by plants in mycorrhizal communities (44, 45). The same mechanisms could be used by fungal species in this family in recognition of their plant hosts. Chitinases are highly expressed by plants during the beginning of mycorrhizal interaction and can be found in their roots, some of them specific to mycorrhizal fungi (46, 47). Since there seems to be a modification of the fungal wall in the first phases of mycorrhizal interaction it may be possible that fungi also produce these proteins in order to facilitate interaction. Another possibility is that fungi start using chitin as a nitrogen source when the mutualist interaction occurs (48). Finally, the proteins involved in the metabolism of complex carbohydrates, since there is an exchange of nutrients during the mycorrhizal interaction and both it has been shown that mycorrhizal fungi can use starch and sucrose as a nutrient or even induce starch accumulation in roots (49–51). In summary, family 13765 might be associated with host recognition and signalling during the initial steps of mycorrhizal interaction, while it is hard to tell the describe the function of family 5912 since it has few characterized proteins, it may be part of the transduction and translation necessary to support the previous functions activated by the signalling of family 13765. It might also be related to the production of secondary metabolites that are part of the signalling system in the plant-fungal association (52).

Other components

Other interesting components that do not have an uncommon ways of life or habitat distribution but that could help us predict related functions are components 1, 3, and 4. These components are the sparse components that appear in Figure 25 along with components 6 and 7. Component 1 is the second largest component with 14 protein families. Out of these 14 protein families, 8 of them have only uncharacterised proteins amongst its members, and the rest are poorly characterized, having only one or few

characterized proteins. These families are: Family 7137, which has a few proteins characterized as oxidoreductases and one with DNA binding function; Family 9411, that has one protein characterized as a transporter in the membrane; Family 8067, which has one protein that is predicted to be a ligase involved in peptide and co-factors synthesis; Family 8336, which has several proteins predicted to be FOG proteins; Family 8158, which has one protein involved in nucleic acid and zinc ion binding; and Family 5923, which has one protein characterized as an alpha-glucosidase and some others predicted as FOG proteins. Due to the high number of uncharacterised proteins in this component is hard to relate it to any function or pathway, however of the 6 characterized families, 3 are related to zinc. Two families have FOG proteins in their members, which have many zinc fingers in their structure, and one family is related to zinc ion binding. It may then be that this component is related to zinc metabolism or pathways that involve zinc.

Component 3 includes two families: Family 11254 and Family 12064. Family 11254 is a small family with 101 proteins that are uncharacterised. Family 12064 is another small family of 80 members with mostly characterized proteins. Most of the proteins present in this family are protein kinases that are involved in signal transduction mechanisms, although we also find proteins related to translation, transcription and cell cycle. Lastly, component 4 is composed of two small families: Family 12244, with 77 proteins, and Family 15194, with 45 proteins. Family 12244 is mostly composed of proteins involved in DNA depended regulation of transcription, one of which is predicted to be a transcription factor of the Forkhead/HNF3 family. However, Family 15194 is composed only of uncharacterised proteins. Unfortunately, because each component has one family with uncharacterised proteins is not possible to assign any common function or pathway to these components until these proteins are further investigated.

The rest of the components are problematic since they may be false positives detected by CoinFinder. These components, which can be found on the left side of Figure 25, share a similar dense structure. They include protein families with ties with most species in our dataset and lack coverage in one or some of the branches of the phylogenetic tree. With this distribution of the co-occurring families it would be hard to rule out the possibility of a shared common ancestor. Furthermore, this co-occurring families are likely the result of a deleterious event in the branches where the families are not present. Still, these families are co-occurring since species either have all the component's families or none, and should can still be linked to a particular function or pathway. For example, component 8 which

includes most species except for the subphylum Saccharomycotina. There are two protein families in component 8: family 1391 and family 4912. Family 1391 is a well characterized family with 922 members, where most of them have an entry in both GO and KOG databases. The vast majority of the proteins in this family are transferases and binding proteins related to biosynthesis. Meanwhile, family 4912 is a protein family with 582 members that are exclusively transaminases. Both of these families' proteins take part in biotin metabolism. This pathway is believed to have disappeared in one ancestor of *S. cerevisiae* and being rebuilt posteriorly (53), which would be in agreement with the species distribution seen in component 8. Component 8's protein families might have been part of *S. cerevisiae* ancestral biotin pathway.

Fungal core genome

Finally, the last component is component 0, which includes 9,897 out of the 9,946 protein families present in the co-occurrence network. Each family included in this component is spread across most species in our dataset, but not all of them since CoinFinder filters out families that are present in more than 90% of species. Since the families are found in in all the phylogeny it would be hard to consider them phylogenetically independent events. However, we can consider the protein families found in this component as the core genome of the fungi, in other words, protein families that are needed for the very basic function of a fungal organism and that most fungi have in their genomes. The core genome concept is relatively recent, being applied initially to prokaryote pan-genomes, though it is increasingly being used to describe eukaryotes including fungi and animals (30, 54). According to this concept, groups of organisms have two kinds of "genomes": the core genome, which included the basic functions needed for every organism in the group to live, and the accessory genome, which are variable genes that not all organisms in the population express, necessary for more specific phenotypes. In this study's case, the core genome would inclde functions that every member of the fungi, independently of their way of life or habitat, needs to express to be able to live. Proteins and genes found in other components like component 6 or 7, that are only expressed in some fungal species, would be part of the accessory genome of the fungi and are linked to more specific functions. If this is the case, we should be able to find an abundance of conserved and ultra-conserved proteins amongst the protein families included in component 0.

To further facilitate this component analysis, we split the component into communities or partitions, resulting in 8 partitions. Of these partitions, some seem to be related to certain functions: Partition 8 is related to translation, transduction and other nuclear processes. Partition 7 is composed by two hydrolases, one for peptide metabolism and one for nucleotides. Partition 6 is related to electron transport. Partition 5 is composed of conserved proteins, many of which are related to DNA transport, metabolism, regulation. The rest of the partitions are more mixed in their composition of protein family functions. Among them we can find structural components of the ribosome (110 families), proteins involved in translation (565 families) and transcription (124 families), DNA and RNA polymerases (around 330 families) and uncharacterised conserved proteins (382 families). From the presence of many of the proteins groups that are conserved not only in fungi, but in many other domains of life, we can assume that the protein families that belong into component 0 may indeed be part of the core genome of the fungi. Some other missing protein families might also be part of the core genome but are not included in the co-occurrence network due to filtering.

Conclusions

By using co-occurrence networks of phylogenetically independent protein families we have tried to find homoplastic events in the fungal phylogeny, and specifically events that might lead us to identify cases of convergent evolution, especially those related to particular ecotypes. We have found three different kinds of components in these networks.

First, components were homoplasy most likely happened, where groups of protein families are gained as a group in independent species across the phylogeny. In these components we were able to detect two where the species distribution of the encoded proteins is very different to the overall distribution of the dataset. One of these components is related to being pathogenic and the other to forming mycorrhizae. Although we found some functions in these families that may be related to these ways of life, the large number of poorly studied proteins makes it hard to link these protein families to phenotypes that may be preferentially expressed by either parasites or mutualists, or to link them to a particular function or process.

Second, components that are most likely the result from a deletion event in one of the branches of the phylogeny, but that we cannot assume that they are not the result of gene

loss in a common ancestor. However, these components are still able to predict a functional relation between the protein families they encompass.

Finally, a big component that includes the vast majority of the fungal protein families. This component is enriched with conserved and ultra-conserved functions, so it might be part of a fungal core genome.

All in all, association and disassociation networks are a powerful tool to detect functional relationships between groups of proteins similar in sequence. By limiting these networks to proteins that are unrelated taxonomically, we can find events of homoplasy that can be linked to ecotypes and particular functions. Even if many of these proteins are currently not characterized, the findings of these project can provide new targets to further research pathogenicity and mutualism in fungi.

Bibliography

1. C. T. Stayton, What does convergent evolution mean? The interpretation of convergence and its implications in the search for limits to evolution. *Interface Focus* **5**, 20150039 (2015).

2. G. C. Conant, A. Wagner, Convergent evolution of gene circuits. *Nat. Genet.* **34**, 264–266 (2003).

3. J. Mistry, R. D. Finn, S. R. Eddy, A. Bateman, M. Punta, Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121–e121 (2013).

4. J. J. Bull, *et al.*, Exceptional Convergent Evolution in a Virus. *Genetics* **147**, 1497–1507 (1997).

5. G. Jones, M. W. Holderied, Bat echolocation calls: adaptation and convergent evolution. *Proc. R. Soc. B Biol. Sci.* **274**, 905–912 (2007).

6. J. Parker, *et al.*, Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* **502**, 228–231 (2013).

7. E. Pichersky, E. Lewinsohn, Convergent Evolution in Plant Specialized Metabolism. *Annu. Rev. Plant Biol.* **62**, 549–566 (2011).

8. R. D. Fernald, Casting a Genetic Light on the Evolution of Eyes. *Science* **313**, 1914–1918 (2006).

9. G. Roth, Convergent evolution of complex brains and high intelligence. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20150049 (2015).

10. D. L. Stern, The genetic causes of convergent evolution. *Nat. Rev. Genet.* **14**, 751–764 (2013).

11. M. P. Speed, K. Arbuckle, Quantification provides a conceptual basis for convergent evolution. *Biol. Rev.* **92**, 815–829 (2017).

12. J. E. Stajich, *et al.*, The Fungi. *Curr. Biol.* **19**, R840–R845 (2009).

13. L. H. Hartwell, Saccharomyces cerevisiae cell cycle. *Bacteriol. Rev.* **38**, 164–198 (1974).

14. B. Qaddouri, *et al.*, The Budding Yeast “Saccharomyces cerevisiae” as a Drug Discovery Tool to Identify Plant-Derived Natural Products with Anti-Proliferative Properties. *Evid.-Based Complement. Altern. Med. ECAM* **2011** (2011).

15. A. H. Aly, A. Debbab, P. Proksch, Fifty years of drug discovery from fungi. *Fungal Divers.* **50**, 3 (2011).

16. C. Pinedo-Rivilla, J. Aleu, I. G. Collado, Pollutants Biodegradation by Fungi. *Curr. Org. Chem.* **13**, 1194–1214 (2009).

17. , 1000 Fungal Genomes Project. *1000 Fungal Genomes Proj.* (October 24, 2016).

18. P.-J. Kim, N. D. Price, Genetic Co-Occurrence Network across Sequenced Microbes. *PLOS Comput. Biol.* **7**, e1002340 (2011).

19. J. W. Shapiro, C. Putonti, Gene Co-occurrence Networks Reflect Bacteriophage Ecology and Evolution. *mBio* **9**, e01870-17 (2018).

20. Z. Wang, *et al.*, A Protein Domain Co-Occurrence Network Approach for Predicting Protein Function and Inferring Species Phylogeny. *PLOS ONE* **6**, e17906 (2011).

21. V. Libis, *et al.*, Uncovering the biosynthetic potential of rare metagenomic DNA using co-occurrence network analysis of targeted sequences. *Nat. Commun.* **10**, 3848 (2019).

22. L.-G. Li, Y. Xia, T. Zhang, Co-occurrence of antibiotic and metal resistance genes revealed in complete genome collection. *ISME J.* **11**, 651–662 (2017).

23. B. Li, *et al.*, Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. *ISME J.* **9**, 2490–2502 (2015).

24. F. J. Whelan, M. Rusilowicz, J. O. McInerney, Coinfinder: detecting significant associations and dissociations in pangenomes. *Microb. Genomics* **6**, e000338 (2020).

25. S. McGinnis, T. L. Madden, BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* **32**, W20–W25 (2004).

26. O. Cohen, H. Ashkenazy, E. Levy Karin, D. Burstein, T. Pupko, CoPAP: Coevolution of Presence–Absence Patterns. *Nucleic Acids Res.* **41**, W232–W237 (2013).

27. Y. Shang, *et al.*, Divergent and Convergent Evolution of Fungal Pathogenicity. *Genome Biol. Evol.* **8**, 1374–1387 (2016).

28. I. Foster, “Globus Toolkit Version 4: Software for Service-Oriented Systems” in *SpringerLink*, (Springer Berlin Heidelberg, 2005), pp. 2–13.

29. O. H. Cisse, J. E. Stajich, FGMP: assessing fungal genome completeness and gene content. *bioRxiv*, 049619 (2016).

30. C. G. P. McCarthy, D. A. Fitzpatrick, Pan-genome analyses of model fungal species. *Microb. Genomics* **5** (2019).

31. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

32. S. M. van Dongen, Graph clustering by flow simulation (2000) (May 29, 2019).

33. , Proceedings of the Python in Science Conference (SciPy): Exploring Network Structure, Dynamics, and Function using NetworkX (November 24, 2016).

34. M. Girvan, M. E. J. Newman, Community structure in social and biological networks. *Proc. Natl. Acad. Sci.* **99**, 7821–7826 (2002).

35. A. J. Enright, S. Van Dongen, C. A. Ouzounis, An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).

36. J. Lee, S. P. Gross, J. Lee, Improved network community structure improves function prediction. *Sci. Rep.* **3**, 2197 (2013).

37. C. Y. Pai, H. K. Chen, H. L. Sheu, N. H. Yeh, Cell-cycle-dependent alterations of a highly phosphorylated nucleolar protein p130 are associated with nucleologenesis. *J. Cell Sci.* **108**, 1911–1920 (1995).

38. C.-Y. Pai, N.-H. Yeh, Cell Proliferation-Dependent Expression of Two Isoforms of the Nucleolar Phosphoprotein P130. *Biochem. Biophys. Res. Commun.* **221**, 581–587 (1996).

39. A. Idnurm, B. J. Howlett, Pathogenicity genes of phytopathogenic fungi. *Mol. Plant Pathol.* **2**, 241–255 (2001).

40. A. P. Van De Wouw, B. J. Howlett, Fungal pathogenicity genes in the age of 'omics.' *Mol. Plant Pathol.* **12**, 507–514 (2011).

41. M. Celerin, *et al.*, Fungal fimbriae are composed of collagen. *EMBO J.* **15**, 4445–4453 (1996).

42. C. Azcón-Aguilar, B. Bago, J. M. Barea, "Saprophytic Growth of Arbuscular Mycorrhizal Fungi" in *Mycorrhiza: Structure, Function, Molecular Biology and Biotechnology*, A. Varma, B. Hock, Eds. (Springer, 1999), pp. 391–408.

43. P. Bonfante, A. Genre, Mechanisms underlying beneficial plant–fungus interactions in mycorrhizal symbiosis. *Nat. Commun.* **1**, 48 (2010).

44. P. L. De Hoff, L. M. Brill, A. M. Hirsch, Plant lectins: the ties that bind in root symbiosis and plant defense. *Mol. Genet. Genomics* **282**, 1–15 (2009).

45. Nivedita, P. K. Verma, K. C. Upadhyaya, Lectin Protein Kinase Is Induced in Plant Roots in Response to the Endophytic Fungus, Piriformospora indica. *Plant Mol. Biol. Report.* **35**, 323–332 (2017).

46. P. Spanu, *et al.*, Chitinase in roots of mycorrhizal Allium porrum: regulation and localization. *Planta* **177**, 447–455 (1989).

47. M. J. Pozo, C. Azcón-Aguilar, E. Dumas-Gaudot, J. M. Barea, Chitosanase and chitinase activities in tomato roots during interactions with arbuscular mycorrhizal fungi or Phytophthora parasitica. *J. Exp. Bot.* **49**, 1729–1739 (1998).

48. J. R. Leake, D. J. Read, Chitin as a nitrogen source for mycorrhizal fungi. *Mycol. Res.* **94**, 993–995 (1990).

49. A. Ohta, Ability of ectomycorrhizal fungi to utilize starch and related substrates. *Mycoscience* **38**, 403–408 (1997).

50. E. Hughes, D. T. Mitchell, Utilization of sucrose by Hymenoscyphus ericae (an ericoid endomycorrhizal fungus) and ectomycorrhizal fungi. *Mycol. Res.* **99**, 1233–1238 (1995).

51. C. Gutjahr, *et al.*, Presymbiotic factors released by the arbuscular mycorrhizal fungus Gigaspora margarita induce starch accumulation in Lotus japonicus roots. *New Phytol.* **183**, 53–61 (2009).

52. Y. Zhi-lin, D. Chuan-chao, C. Lian-qing, Regulation and accumulation of secondary metabolites in plant-fungus symbiotic system. *Afr. J. Biotechnol.* **6** (2007).

53. C. Hall, F. S. Dietrich, The Reacquisition of Biotin Prototrophy in Saccharomyces cerevisiae Involved Horizontal Gene Transfer, Gene Duplication and Gene Clustering. *Genetics* **177**, 2293–2307 (2007).

54. A. A. Golicz, P. E. Bayer, P. L. Bhalla, J. Batley, D. Edwards, Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications. *Trends Genet.* **36**, 132–145 (2020).

Discussion

Convergent evolution has long been studied in evolutionary biology as one of the best examples of the selective pressure and the process of adaptation and it is still extensively studied nowadays (99). Convergent evolution can produce similar phenotypes in different lineages of organisms but there can be several independent strategies or mechanisms used by the different organisms in order to achieve in these adaptations, even at the molecular level (100, 101). One of the ways in which we could check for convergent evolution is by discovering sets of associated genes that appear independently in species not related by recent common ancestry that have adapted to similar ecosystems. Across the tree of life there are sets of genes that tend to appear together, like the well-studied effect of genetic linkage in eukaryotes, where genes that participate in similar functions or pathways are close in the chromosome and tend to be inherited together (102). In prokaryotes, there have been extensive studies on how certain groups of genes, like antibiotic resistance related genes, tend to appear together (54, 103, 104). However, genetic linkage can be related to common ancestry, and in prokaryotes gene co-occurrence is studied in the context of horizontal gene transfer events in microbial communities. We wanted to explore whether these gain and loss events of groups of genes can independently appear in eukaryotes, where HGT events are not as common as in prokaryotes (105–107). Species not related by recent common ancestry in the context of convergent evolution can be particularly interesting since any evidence of convergent evolution might indicate the presence of strong selective pressure. If there are strategies that are the most optimal for adapting to a particular ecosystem, we hypothesized that we should be able to find groups of genes in different organisms that use similar strategies of adaptation and therefore can have a degree of sequence similarity.

We have sought to discover these genes that are critical for adaptation to a particular ecosystem by using a network based approach in conjunction with a phylogeny to detect cases of convergent evolution in the fungi. We hypothesized that fungi that adapt independently to similar habitats will share some sequence similarity in genes that encode essential functions to thrive in that particular habitat due to similar adaptation processes. To test this hypothesis, we have considered that genes that have evolved in similar ways to perform similar functions in species that are not related by recent common ancestry could

be considered as important for that specific function. By then exploring the species distribution within these gene groups to find which ones are enriched for the presence of species that live in a particular environment, we can ultimately link a group of genes with their importance to survival in a habitat. Not being able to find these groups of genes would instead indicate that fungi use different strategies to adapt to similar environments and therefore no significant sequence similarity can be found.

In the first instance, we have sought to gain some insight into how networks can be used as a tool to understand phylogenetic relationships between biological entities when other more traditional methods can not be applied. We have used plasmids in conjunction with bipartite networks to investigate evolutionary relationships and found that plasmids are grouped by at least two characteristics, function and taxonomy of their hosts. We further explored of the characteristics that drive the evolution of plasmids, in particular physical properties, nucleotide sequences of plasmids, and amino-acid sequences of their protein products, by using ERGMs. Still, host organism taxonomy and plasmid protein functions remain the main driving forces in shaping the network.

There have been some recent studies where plasmid evolutionary relationships have been studied with similar methods reaching similar conclusions (108, 109), which reinforces the findings we obtained in this part of the project. Thus, we found evidence that networks can indeed be used to find relevant phylogenetic information from a set of biological entities, and they could therefore be applied to the fungal dataset.

Furthermore, a similar network analysis could be applied to other datasets where phylogenetic analyses are not possible, particularly to viruses. Viruses are similar to plasmids in the sense that they are not considered living organisms, do not have common sequences shared between all viruses, and they also play an important role in HGT (110, 111). In recent years there has been an increase in the sequencing of viruses in the field of metagenomics, in order to gain insight into the virus that inhabit different regions. Viral metagenomics has shed light in many unknown viral communities, leading to a new understanding of viral ecology, taxonomy and interactions (112, 113). Viral communities have been studied in a variety of environments, like the ocean (114, 115), plants (116), and animals (117, 118). A similar approach to the one we took in this thesis could be then applied to these datasets where the relationships between the different viruses are explored. Even if in some cases the information about the virus hosts is missing, which is likely in the field of metagenomics, there is still the possibility to learn about the functional

relationships between viruses and whether the physical properties of their sequence play a role in these interactions.

We then proceeded to use networks in order to find relevant groups of associated or disassociated proteins relevant to particular phenotypes. We chose to use a big dataset, since Coinfinder has been proven to detect more associations with bigger datasets, usually requiring at least 50 genomes or more to detect significant associations (55), even if analysing a big dataset will make the analysis harder to perform. But first, to detect convergent evolution events, we needed to make a fungal phylogeny. Given that our initial attempts at making maximum-likelihood phylogenies to assess the selected genes for our dataset returned phylogenetic hypotheses with low support and in disagreement in the early groups, and that alternative methods like super-trees failed to retrieve a well supported phylogeny, we turned our attention to more complex phylogenetic methods. We then used data-heterogeneous and tree-heterogeneous models, which allowed us to consider different substitution processes for the more problematic groups like *Microspordia*, with a representative reduced dataset to get a backbone of the fungal phylogeny. The smaller dataset was used due to the excessive computational time needed for such complex analyses. The heterogeneous models returned a robust fungal phylogeny with a novel placement of *Microspordia,* which may be explained by heterogeneous models being able to split apart the long *Microspordia* branch from other long branches present in the fungal phylogeny. To asses the validity of our methodology in the reconstruction of the heterogeneous fungal phylogeny, we used a similar approach to reconstruct a previously published well supported fungal phylogeny that used a similar but smaller dataset from the FGP using various phylogenetic methodologies by *McCarthy* and *Fitzpatrick* (65). Our reconstruction of this validation phylogeny was similar to the one published by *McCarthy* and *Fitzpatrick*, with high support values as can be seen in Figure 19. As with our dataset, a tree and site heterogeneous phylogeny proved to be more robust and have a higher likelihood than a phylogenetic tree that was produced using a model that only allowed data heterogeneity in order to reconstruct the phylogeny. The difference in log likelihood between the only data heterogeneous phylogeny and the tree and site heterogeneous phylogeny is similar in both the smaller validation dataset and our original fungal dataset. By proving that our methodology is well suited to resolve both trees with and without *Microsporidia*, we have obtained more support for the potential novel placement of *Microsporidia*.

An alternative approach could have been taken to link relevant sets of genes to specific functions or phenotypes, like multivariate analysis (119). Using this approach would have allowed us to leave the complexity of the fungal phylogeny out of the analysis. However, since these methods do not consider the phylogenetic relationships between species, it would be nearly impossible to discard sets of genes related by common ancestry from independent sets of genes. Therefore, a phylogenetic reconstruction was needed for the purpose of this project. Despite the problems that data and tree heterogeneous models have, namely their complexity of use and lower scalability when compared to homogeneous models, they have proved to be a powerful tool (120). Using these models has allowed us to confidently place in the fungal phylogeny both deep branches and very divergent branches. There still remain many clades across the tree of life with a difficult placement in similar cases to *Microsporidia* like sponges and Spiralians in Animals *(121, 122)*, palms and *Amborella* in plants (123, 124) and ciliates in protozoans (125, 126). Using heterogeneous models with small but representative datasets of various groups of organisms, could be useful to resolve these deep or problematic branches that are still under discussion.

Finally, with the support from the heterogeneous tree for early branches in the fungal phylogenies, we then proceeded to obtain a phylogeny for the complete set of species in our dataset. We used the phylogeny of all the fungal species present in our dataset in conjunction with association and disassociation networks to explore significant patterns of presence and absence of genes in genomes. We found 19 components of coinciding proteins that were then further analysed by exploring the genes present in each one of them in order to link them to a function. Out of these components, we found two components that are enriched for species that have a particular way of life, specifically parasitic and mycorrhizal fungi. Furthermore, the functions linked to these components are relevant to their respective way of life. The pathogen component is linked to growth rate and infective structures and the mycorrhizal component is linked to host recognition, cell wall modifications and plant related metabolism. Additionally, we have found that the largest component found in the association analysis seems to be related to the core genome of the fungi, or at least part of it. It is composed mainly of conserved functions that are needed for fungal species to perform their basic life functions. Other components are related to the accessory genome, that have more specialized functions that are not found in every fungal species and that represent more specialized functions in the adaptation to environments.

Pan-genomes have been mostly explored in prokaryotes (92), but there have been a few studies of pan-genomes in eukaryotes (127–129), including fungi (130, 131). Even if it was not in our initial objectives, we have thus found further evidence of the structure of the fungal pan-genome.

By obtaining these results we have accomplished the objectives we set out to do in the beginning of the project, obtaining evidence of selection of similar molecular strategies in convergent evolution that give rise to sequence similarity between independent sets of genes. However, this thesis results involve only computational work that only provides candidate genes with potential involvement in pathogenicity or mutualism, and further research would need to be done in order to test whether the candidate genes are valid, particularly with experimental support. One possible way we could test our findings in convergent evolution is with knockout experiments. Knockout experiments can be used to test the effect that the loss of a gene or a set of genes affect an organism. Knockout experiments have been extensively used in mice to test the influence of genes in diseases, immunity and even behaviour (132–134). Fungi are also a common target for knockout experiments, many of which are focused on genes that affect fungal pathogenicity (135–137). Hence, we could use a similar approach with knockout experiments to test how the loss of one of the genes present in a particular component affects one of the fungal species that have these genes present. For example, we could test in pathogens whether loosing the genes in the component affects their ability to infect their host or if they are less successful doing so. For mycorrhizal species, we could check whether removing these genes affects their capability to form mutualistic associations. It could also be interesting to test whether loss of function (i.e. loosing the ability to infect or form mycorrhizal associations) also happen if we just remove one of the genes present in the component or it just happens when both are removed. If it is the case that there is only a reduced effectiveness of the organism without all the genes in a component, we could also test whether loosing only one of the genes reduces effectiveness to a lesser degree. Similar experiments have been thoroughly performed to test gene-gene interactions in yeast metabolism, where the effects on the phenotype of a single knockout strain are compared with the effects on the phenotype of double knockout strains (138) or multiple knockout strains (139). There has also been research in the effect of knockouts of different combination of pathogenicity related genes in fungal pathogens of wheat (140), which are similar experiments to what could be applied to the results of this thesis. These experiments are also prevalent using mouse as a model, where some experiments have found that single knockout mice do not have negative effects on the phenotype but double knockout mice have a negative

phenotype (141). Furthermore, some experiments have found that single gene knockouts to have a positive effect on the phenotype while double knockouts have a negative effect (142), which can be interesting to explore in the case of disassociation networks. Additionally, co-occurrence networks could be applied to other groups of organisms to check for convergent evolution events to identify interesting sets of genes or proteins. It would be particularly helpful to test this methodology with clades with an already well supported phylogeny in the literature, as this would remove some of the problems we encountered in this project.

Lastly, we validated our project by repeating our analysis with a smaller and more manageable dataset with a well-supported phylogeny, that did not include the problematic clades that made our analysis more difficult. However, when we applied the same methodology to this validation fungal phylogeny, we were unable to find any significant associations or disassociations. This lack of findings when we analysed a bigger dataset has served to validate our initial stance that a large dataset was needed to extract the information we required. By increasing the dataset size from 81 species to 671 ("only" eight times bigger) we are able to find evidence of convergent evolution in the fungi. As it has been happening to the field of big data in biology, there has been numerous efforts to further increase the availability of data. Projects like the FGP, many of the human sequencing projects or the previously mentioned metagenomic projects have increased the quantity of sequenced genomes, from only a few genomes a decade ago to several tens of thousands nowadays, an example of which can be checked in JGI GOLD statists website (143). This ever-increasing availability of bigger and bigger datasets is allowing us to use more demanding methods that require large quantities of data to obtain informative conclusions. If by increasing by less than tenfold our dataset size has allowed us to detect evidence of convergent evolution in fungi, by feeding more data to newer and more demanding methods we may be able to discover many features that as of today have eluded us.

Conclusion

In this project we have tried to use network methods to extract information about the evolutionary relationships between members of the fungi, with particular attention to convergent evolution and its ties to shared protein functionality.

First, we tested network methods by using a more manageable plasmid dataset, where we used bipartite graphs and ERGMs to identify properties of the evolutionary relationships between plasmids and their hosts, where a traditional phylogenetic approach can not be used. We found that plasmid tend to organize around two types of communities: related to protein function and related to taxonomy. Communities related to taxonomy seem to have emerged mostly as a result of isolation of the hosts, either by extreme habitats or other features like atypical plasmid structure. With ERGMs we proved that physical properties of the plasmids sequences and the sequence of the proteins they encode can impact the evolutionary relationships with their hosts. In particular, proteins and plasmids adapted to extreme conditions are more shared in a plasmid network that it is expected if plasmids where shared randomly. Therefore, there seems to be an increased sharing of stable proteins in extremophile environments, which could be the result of a positive fitness effect in hosts carrying these plasmids in extreme habitats.

Then, we proceeded to analyse our fungal dataset. As a first step we needed to produce a phylogenetic hypothesis of the fungal taxonomy to account for common ancestry when investigating protein association networks. However, long branch attraction was proving to be an issue to resolve deep branches in the fungal phylogeny, especially considering the presence of intracellular parasites that have very rapid amino-acid substitution rates. To get around this issue, we used a reduced dataset with data and tree heterogeneous models, which provided with a robust general structure of the fungal clades as well as a novel placement of *Micropsoridia* in the fungal phylogeny, seemingly resulting from solving the long branch attraction issues. As a last step we used a complete data-heterogeneous model with all taxa in our fungal dataset to be used as a phylogeny for association networks.

The last step was detecting convergent evolution in the fungal kingdom by using protein co-occurrence networks along with the fungal phylogeny obtained in the previous step. We detected a total of 19 co-occurring components, two of which can be tied to parasitical fungi and mycorrhizal fungi respectively. Other components are hard to associate with

particular phenotypes but usually a common function or pathway can be found in each component. Nonetheless, the low characterization of proteins that are found in the components presents and issue when trying to link components to functions, so further investigation is required in these groups of proteins before they can be confidently linked to a particular function. Lastly, we tied the component that included the vast majority of the co-occurring protein families to the fungal core genome, functions that are found in most fungal species and are necessary for to maintain basic life, as many protein families are found in this large component.

Bibliography

1. C. T. Stayton, What does convergent evolution mean? The interpretation of convergence and its implications in the search for limits to evolution. *Interface Focus* **5**, 20150039 (2015).

2. G. C. Conant, A. Wagner, Convergent evolution of gene circuits. *Nat. Genet.* **34**, 264–266 (2003).

3. J. Mistry, R. D. Finn, S. R. Eddy, A. Bateman, M. Punta, Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121–e121 (2013).

4. J. J. Bull, *et al.*, Exceptional Convergent Evolution in a Virus. *Genetics* **147**, 1497–1507 (1997).

5. G. Jones, M. W. Holderied, Bat echolocation calls: adaptation and convergent evolution. *Proc. R. Soc. B Biol. Sci.* **274**, 905–912 (2007).

6. J. Parker, *et al.*, Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* **502**, 228–231 (2013).

7. E. Pichersky, E. Lewinsohn, Convergent Evolution in Plant Specialized Metabolism. *Annu. Rev. Plant Biol.* **62**, 549–566 (2011).

8. D. L. Stern, The genetic causes of convergent evolution. *Nat. Rev. Genet.* **14**, 751–764 (2013).

9. D. B. Wake, M. H. Wake, C. D. Specht, Homoplasy: From Detecting Pattern to Determining Process and Mechanism of Evolution. *Science* **331**, 1032–1035 (2011).

10. R. D. Fernald, Casting a Genetic Light on the Evolution of Eyes. *Science* **313**, 1914–1918 (2006).

11. Y. Liu, *et al.*, Convergent sequence evolution between echolocating bats and dolphins. *Curr. Biol.* **20**, R53–R54 (2010).

12. G. Roth, Convergent evolution of complex brains and high intelligence. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20150049 (2015).

13. M. P. Speed, K. Arbuckle, Quantification provides a conceptual basis for convergent evolution. *Biol. Rev.* **92**, 815–829 (2017).

14. J. E. Stajich, *et al.*, The Fungi. *Curr. Biol.* **19**, R840–R845 (2009).

15. , Home - Onnia scaura P-53A v1.0 (June 8, 2020).

16. R. Dean, *et al.*, The Top 10 fungal pathogens in molecular plant pathology. *Mol. Plant Pathol.* **13**, 414–430 (2012).

17. R. Horbach, A. R. Navarro-Quesada, W. Knogge, H. B. Deising, When and how to kill a plant cell: Infection strategies of plant pathogenic fungi. *J. Plant Physiol.* **168**, 51–62 (2011).

18. O. Filtenborg, J. C. Frisvad, U. Thrane, Moulds in food spoilage. *Int. J. Food Microbiol.* **33**, 85–102 (1996).

19. J. I. Pitt, A. D. Hocking, "The Ecology of Fungal Food Spoilage" in *Fungi and Food Spoilage*, J. I. Pitt, A. D. Hocking, Eds. (Springer US, 2009), pp. 3–9.

20. V. Tournas, Heat-Resistant Fungi of Importance to the Food and Beverage Industry. *Crit. Rev. Microbiol.* **20**, 243–263 (1994).

21. G. D. Brown, *et al.*, Hidden Killers: Human Fungal Infections. *Sci. Transl. Med.* **4**, 165rv13-165rv13 (2012).

22. M. C. Fisher, *et al.*, Emerging fungal threats to animal, plant and ecosystem health. *Nature* **484**, 186–194 (2012).

23. L. H. Hartwell, Saccharomyces cerevisiae cell cycle. *Bacteriol. Rev.* **38**, 164–198 (1974).

24. B. Qaddouri, *et al.*, The Budding Yeast "Saccharomyces cerevisiae" as a Drug Discovery Tool to Identify Plant-Derived Natural Products with Anti-Proliferative Properties. *Evid.-Based Complement. Altern. Med. ECAM* **2011** (2011).

25. A. H. Aly, A. Debbab, P. Proksch, Fifty years of drug discovery from fungi. *Fungal Divers.* **50**, 3 (2011).

26. C. Pinedo-Rivilla, J. Aleu, I. G. Collado, Pollutants Biodegradation by Fungi. *Curr. Org. Chem.* **13**, 1194–1214 (2009).

27. M. G. A. van der Heijden, *et al.*, Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity. *Nature* **396**, 69–72 (1998).

28. Y. Rouphael, *et al.*, Arbuscular mycorrhizal fungi act as biostimulants in horticultural crops. *Sci. Hortic.* **196**, 91–108 (2015).

29. J. Arihara, T. Karasawa, Effect of previous crops on arbuscular mycorrhizal formation and growth of succeeding maize. *Soil Sci. Plant Nutr.* **46**, 43–51 (2000).

30. , 1000 Fungal Genomes Project. *1000 Fungal Genomes Proj.* (October 24, 2016).

31. M. Blackwell, D. S. Hibbett, J. W. Taylor, J. W. Spatafora, Research Coordination Networks: A phylogeny for kingdom Fungi (Deep Hypha). *Mycologia* **98**, 829–837 (2006).

32. D. J. McLaughlin, D. S. Hibbett, F. Lutzoni, J. W. Spatafora, R. Vilgalys, The search for the fungal tree of life. *Trends Microbiol.* **17**, 488–497 (2009).

33. T. Y. James, *et al.*, A molecular phylogeny of the flagellated fungi (Chytridiomycota) and description of a new phylum (Blastocladiomycota). *Mycologia* **98**, 860–871 (2006).

34. D. Redecker, P. Raab, Phylogeny of the Glomeromycota (arbuscular mycorrhizal fungi): Recent developments and new gene markers. *Mycologia* **98**, 885–895 (2006).

35. A.-L. Barabási, Scale-Free Networks: A Decade and Beyond. *Science* **325**, 412–413 (2009).

36. P. Hawe, L. Ghali, Use of social network analysis to map the social relationships of staff and teachers at school. *Health Educ. Res.* **23**, 62–69 (2008).

37. S. Halary, J. W. Leigh, B. Cheaib, P. Lopez, E. Bapteste, Network analyses structure genetic diversity in independent genetic worlds. *Proc. Natl. Acad. Sci.* **107**, 127–132 (2010).

38. X. Zhu, M. Gerstein, M. Snyder, Getting connected: analysis and principles of biological networks. *Genes Dev.* **21**, 1010–1024 (2007).

39. G. A. Pavlopoulos, *et al.*, Using graph theory to analyze biological networks. *BioData Min.* **4**, 10 (2011).

40. M. Newman, The Structure and Function of Complex Networks. *SIAM Rev.* **45**, 167–256 (2003).

41. R. Albert, A.-L. Barabási, Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).

42. S. Fortunato, Community detection in graphs. *Phys. Rep.* **486**, 75–174 (2010).

43. A.-L. Barabási, Z. N. Oltvai, Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).

44. R. Albert, Scale-free networks in cell biology. *J. Cell Sci.* **118**, 4947–4957 (2005).

45. A.-L. Barabási, N. Gulbahce, J. Loscalzo, Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).

46. R. Diestel, *Graph Theory* (Springer Science & Business Media, 2006).

47. P. Zhang, *et al.*, Clustering coefficient and community structure of bipartite networks. *Phys. Stat. Mech. Its Appl.* **387**, 6869–6875 (2008).

48. P. G. Lind, M. C. González, H. J. Herrmann, Cycles and clustering in bipartite networks. *Phys. Rev. E* **72**, 056127 (2005).

49. G. Robins, M. Alexander, Small Worlds Among Interlocking Directors: Network Structure and Distance in Bipartite Graphs. *Comput. Math. Organ. Theory* **10**, 69–94 (2004).

50. Z. Wang, *et al.*, A Protein Domain Co-Occurrence Network Approach for Predicting Protein Function and Inferring Species Phylogeny. *PLOS ONE* **6**, e17906 (2011).

51. P.-J. Kim, N. D. Price, Genetic Co-Occurrence Network across Sequenced Microbes. *PLOS Comput. Biol.* **7**, e1002340 (2011).

52. V. Libis, *et al.*, Uncovering the biosynthetic potential of rare metagenomic DNA using co-occurrence network analysis of targeted sequences. *Nat. Commun.* **10**, 3848 (2019).

53. B. Li, *et al.*, Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. *ISME J.* **9**, 2490–2502 (2015).

54. L.-G. Li, Y. Xia, T. Zhang, Co-occurrence of antibiotic and metal resistance genes revealed in complete genome collection. *ISME J.* **11**, 651–662 (2017).

55. F. J. Whelan, M. Rusilowicz, J. O. McInerney, Coinfinder: detecting significant associations and dissociations in pangenomes. *Microb. Genomics* **6**, e000338 (2020).

56. O. Cohen, H. Ashkenazy, E. Levy Karin, D. Burstein, T. Pupko, CoPAP: Coevolution of Presence–Absence Patterns. *Nucleic Acids Res.* **41**, W232–W237 (2013).

57. Y. Shang, *et al.*, Divergent and Convergent Evolution of Fungal Pathogenicity. *Genome Biol. Evol.* **8**, 1374–1387 (2016).

58. M. A. Naranjo-Ortiz, T. Gabaldón, Fungal evolution: diversity, taxonomy and phylogeny of the Fungi. *Biol. Rev.* **94**, 2101–2137 (2019).

59. R. J. O'Hara, Homage to Clio, or, Toward an Historical Philosophy for Evolutionary Biology. *Syst. Biol.* **37**, 142–155 (1988).

60. Z. Yang, B. Rannala, Molecular phylogenetics: principles and practice. *Nat. Rev. Genet.* **13**, 303–314 (2012).

61. B. T. Grenfell, *et al.*, Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. *Science* **303**, 327–332 (2004).

62. R. D. Gray, A. J. Drummond, S. J. Greenhill, Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement. *Science* **323**, 479–483 (2009).

63. A. D. Young, J. P. Gillung, Phylogenomics — principles, opportunities and pitfalls of big-data phylogenetics. *Syst. Entomol.* **45**, 225–247 (2020).

64. E. R. Mardis, A decade's perspective on DNA sequencing technology. *Nature* **470**, 198–203 (2011).

65. C. G. P. McCarthy, D. A. Fitzpatrick, "Chapter Six - Multiple Approaches to Phylogenomic Reconstruction of the Fungal Kingdom" in *Advances in Genetics*, Fungal Phylogenetics and Phylogenomics., J. P. Townsend, Z. Wang, Eds. (Academic Press, 2017), pp. 211–266.

66. R. W. Holley, *et al.*, Structure of a Ribonucleic Acid. *Science* **147**, 1462–1465 (1965).

67. A. Goffeau, *et al.*, Life with 6000 Genes. *Science* **274**, 546–567 (1996).

68. J. E. Galagan, *et al.*, The genome sequence of the filamentous fungus Neurospora crassa. *Nature* **422**, 859–868 (2003).

69. J. E. Galagan, M. R. Henn, L.-J. Ma, C. A. Cuomo, B. Birren, Genomics of the fungal kingdom: Insights into eukaryotic biology. *Genome Res.* **15**, 1620–1631 (2005).

70. V. Wood, *et al.*, The genome sequence of Schizosaccharomyces pombe. *Nature* **415**, 871–880 (2002).

71. C. A. Cuomo, B. W. Birren, “Chapter 34 - The Fungal Genome Initiative and Lessons Learned from Genome Sequencing” in *Methods in Enzymology*, Guide to Yeast Genetics: Functional Genomics, Proteomics, and Other Systems Analysis., (Academic Press, 2010), pp. 833–855.

72. M. J. Sanderson, A. Purvis, C. Henze, Phylogenetic supertrees: Assembling the trees of life. *Trends Ecol. Evol.* **13**, 105–109 (1998).

73. P. G. Foster, T. Schultz, Modeling Compositional Heterogeneity. *Syst. Biol.* **53**, 485–495 (2004).

74. N. Corradi, P. J. Keeling, Microsporidia: a journey through radical taxonomical revisions. *Fungal Biol. Rev.* **23**, 1–8 (2009).

75. P. J. Keeling, N. M. Fast, Microsporidia: Biology and Evolution of Highly Reduced Intracellular Parasites. *Annu. Rev. Microbiol.* **56**, 93–116 (2002).

76. T. Y. James, *et al.*, Shared Signatures of Parasitism and Phylogenomics Unite Cryptomycota and Microsporidia. *Curr. Biol.* **23**, 1548–1553 (2013).

77. S. Nakjang, *et al.*, Reduction and Expansion in Microsporidian Genome Evolution: New Insights from Comparative Genomics. *Genome Biol. Evol.* **5**, 2285–2303 (2013).

78. S. Klopfstein, T. Massingham, N. Goldman, More on the Best Evolutionary Rate for Phylogenetic Analysis. *Syst. Biol.* **66**, 769–785 (2017).

79. J. Bergsten, A review of long-branch attraction. *Cladistics* **21**, 163–193 (2005).

80. C. M. Thomas, *et al.*, Annotation of plasmid genes. *Plasmid* (2017).

81. C. M. Thomas, K. M. Nielsen, Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. *Nat. Rev. Microbiol.* **3**, 711 (2005).

82. H. Ochman, J. G. Lawrence, E. A. Groisman, Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299 (2000).

83. C. I. Kado, Origin and evolution of plasmids. *Antonie Van Leeuwenhoek* **73**, 117–126 (1998).

84. M. J. Bottery, A. J. Wood, M. A. Brockhurst, Temporal dynamics of bacteria-plasmid coevolution under antibiotic selection. *ISME J.* **13**, 559 (2019).

85. S. N. Cohen, Transposable genetic elements and plasmid evolution. *Nature* **263**, 731–738 (1976).

86. J. Petersen, Phylogeny and compatibility: plasmid classification in the genomics era. *Arch. Microbiol.* **193**, 313–321 (2011).

87. F. D. Ciccarelli, *et al.*, Toward Automatic Reconstruction of a Highly Resolved Tree of Life. *Science* **311**, 1283–1287 (2006).

88. G. del Solar, R. Giraldo, M. J. Ruiz-Echevarría, M. Espinosa, R. Díaz-Orejas, Replication and Control of Circular Bacterial Plasmids. *Microbiol. Mol. Biol. Rev.* **62**, 434–464 (1998).

89. T. Dimitriu, L. Marchant, A. Buckling, B. Raymond, Bacteria from natural populations transfer plasmids mostly towards their kin. *Proc. R. Soc. B Biol. Sci.* **286**, 20191110 (2019).

90. R. Jain, M. C. Rivera, J. E. Moore, J. A. Lake, Horizontal Gene Transfer Accelerates Genome Innovation and Evolution. *Mol. Biol. Evol.* **20**, 1598–1602 (2003).

91. H. Heuer, K. Smalla, Plasmids foster diversification and adaptation of bacterial populations in soil. *FEMS Microbiol. Rev.* **36**, 1083–1104 (2012).

92. J. O. McInerney, A. McNally, M. J. O'Connell, Why prokaryotes have pangenomes. *Nat. Microbiol.* **2**, 17040 (2017).

93. V. Burrus, G. Pavlovic, B. Decaris, G. Guédon, Conjugative transposons: the tip of the iceberg. *Mol. Microbiol.* **46**, 601–610 (2002).

94. J. P. Hall, M. A. Brockhurst, C. Dytham, E. Harrison, The evolution of plasmid stability: Are infectious transmission and compensatory evolution competing evolutionary trajectories? *Plasmid* **91**, 90–95 (2017).

95. E. Bosi, R. Fani, M. Fondi, The mosaicism of plasmids revealed by atypical genes detection and analysis. *BMC Genomics* **12**, 403 (2011).

96. M. Kohiyama, S. Hiraga, I. Matic, M. Radman, Bacterial Sex: Playing Voyeurs 50 Years Later. *Science* **301**, 802–803 (2003).

97. M. Morris, M. S. Handcock, D. R. Hunter, Specification of Exponential-Family Random Graph Models: Terms and Computational Aspects. *J. Stat. Softw.* **24**, 1548–7660 (2008).

98. D. R. Hunter, M. S. Handcock, C. T. Butts, S. M. Goodreau, M. Morris, ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks. *J. Stat. Softw.* **24**, nihpa54860 (2008).

99. T. B. Sackton, N. Clark, Convergent evolution in the genomics era: new insights and directions. *Philos. Trans. R. Soc. B Biol. Sci.* **374**, 20190102 (2019).

100. K. A. Mackin, R. A. Roy, D. L. Theobald, An Empirical Test of Convergent Evolution in Rhodopsins. *Mol. Biol. Evol.* **31**, 85–95 (2014).

101. T. Ashfield, L. E. Ong, K. Nobuta, C. M. Schneider, R. W. Innes, Convergent Evolution of Disease Resistance Gene Specificity in Two Flowering Plant Families. *Plant Cell* **16**, 309–318 (2004).

102. M. Foulongne-Oriol, Genetic linkage mapping in fungi: current state, applications, and future trends. *Appl. Microbiol. Biotechnol.* **95**, 891–904 (2012).

103. Z. Yu, P. He, L. Shao, H. Zhang, F. Lü, Co-occurrence of mobile genetic elements and antibiotic resistance genes in municipal solid waste landfill leachates: A preliminary insight into the role of landfill age. *Water Res.* **106**, 583–592 (2016).

104. C. Pal, J. Bengtsson-Palme, E. Kristiansson, D. G. J. Larsson, Co-occurrence of resistance genes to antibiotics, biocides and metals reveals novel insights into their co-selection potential. *BMC Genomics* **16**, 964 (2015).

105. C. G. Kurland, B. Canback, O. G. Berg, Horizontal gene transfer: a critical view. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 9658–9662 (2003).

106. R. Bock, The give-and-take of DNA: horizontal gene transfer in plants. *Trends Plant Sci.* **15**, 11–22 (2010).

107. J. Huang, Horizontal gene transfer in eukaryotes: The weak-link model. *BioEssays* **35**, 868–875 (2013).

108. M. Acman, L. van Dorp, J. M. Santini, F. Balloux, Large-scale network analysis captures biological features of bacterial plasmids. *Nat. Commun.* **11**, 2452 (2020).

109. M. Tamminen, M. Virta, R. Fani, M. Fondi, Large-Scale Analysis of Plasmid Relationships through Gene-Sharing Networks. *Mol. Biol. Evol.* **29**, 1225–1240 (2012).

110. E. V. Koonin, T. G. Senkevich, V. V. Dolja, The ancient Virus World and evolution of cells. *Biol. Direct* **1**, 29 (2006).

111. E. Sano, S. Carlson, L. Wegley, F. Rohwer, Movement of viruses between biomes. *Appl. Environ. Microbiol.* **70**, 5842–5846 (2004).

112. F. Schulz, *et al.*, Giant virus diversity and host interactions through global metagenomics. *Nature* **578**, 432–436 (2020).

113. P. Simmonds, *et al.*, Virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol.* **15**, 161–168 (2017).

114. D. M. Kristensen, A. R. Mushegian, V. V. Dolja, E. V. Koonin, New dimensions of the virus world discovered through metagenomics. *Trends Microbiol.* **18**, 11–19 (2010).

115. A. I. Culley, A. S. Lang, C. A. Suttle, Metagenomic Analysis of Coastal RNA Virus Communities. *Science* **312**, 1795–1798 (2006).

116. A. H. Stobbe, M. J. Roossinck, Plant virus metagenomics: what we know and why we need to know more. *Front. Plant Sci.* **5** (2014).

117. K. T. T. Kwok, D. F. Nieuwenhuijse, M. V. T. Phan, M. P. G. Koopmans, Virus Metagenomics in Farm Animals: A Systematic Review. *Viruses* **12**, 107 (2020).

118. J. M. Day, L. L. Ballard, M. V. Duke, B. E. Scheffler, L. Zsak, Metagenomic analysis of the turkey gut RNA virus community. *Virol. J.* **7**, 313 (2010).

119. D. Montaner, J. Dopazo, Multidimensional Gene Set Analysis of Genomic Data. *PLOS ONE* **5**, e10348 (2010).

120. P. G. Foster, Modeling Compositional Heterogeneity. *Syst. Biol.* **53**, 485–495 (2004).

121. C. E. Laumer, *et al.*, Spiralian Phylogeny Informs the Evolution of Microscopic Lineages. *Curr. Biol.* **25**, 2000–2006 (2015).

122. G. Wörheide, *et al.*, “Chapter One - Deep Phylogeny and Evolution of Sponges (Phylum Porifera)” in *Advances in Marine Biology*, Advances in Sponge Science:

Phylogeny, Systematics, Ecology., M. A. Becerro, M. J. Uriz, M. Maldonado, X. Turon, Eds. (Academic Press, 2012), pp. 1–78.

123. C. F. Barrett, *et al.*, Plastid genomes reveal support for deep phylogenetic relationships and extensive rate variation among palms and other commelinid monocots. *New Phytol.* **209**, 855–870 (2016).

124. Z. Xi, L. Liu, J. S. Rest, C. C. Davis, Coalescent versus concatenation methods and the placement of Amborella as sister to water lilies. *Syst. Biol.* **63**, 919–932 (2014).

125. E. Gentekaki, *et al.*, Large-scale phylogenomic analysis reveals the phylogenetic position of the problematic taxon Protocruzia and unravels the deep phylogenetic affinities of the ciliate lineages. *Mol. Phylogenet. Evol.* **78**, 36–42 (2014).

126. J. Jiang, J. Huang, S. A. Al-Farraj, X. Lin, X. Hu, Morphology and Molecular Phylogeny of Two Poorly Known Species of Protocruzia (Ciliophora: Protocruziida). *J. Eukaryot. Microbiol.* **64**, 144–152 (2017).

127. A. A. Golicz, P. E. Bayer, P. L. Bhalla, J. Batley, D. Edwards, Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications. *Trends Genet.* **36**, 132–145 (2020).

128. A. A. Golicz, J. Batley, D. Edwards, Towards plant pangenomics. *Plant Biotechnol. J.* **14**, 1099–1105 (2016).

129. R. Li, *et al.*, Building the sequence map of the human pan-genome. *Nat. Biotechnol.* **28**, 57–63 (2010).

130. C. Plissonneau, F. E. Hartmann, D. Croll, Pangenome analyses of the wheat pathogen Zymoseptoria tritici reveal the structural basis of a highly plastic eukaryotic genome. *BMC Biol.* **16**, 5 (2018).

131. C. G. P. McCarthy, D. A. Fitzpatrick, Pan-genome analyses of model fungal species. *Microb. Genomics* **5** (2019).

132. M. R. Picciotto, K. Wickman, Using Knockout and Transgenic Mice to Study Neurophysiology and Behavior. *Physiol. Rev.* **78**, 1131–1163 (1998).

133. J. Zhang, *et al.*, Serotonin Exhibits Accelerated Bleomycin-Induced Pulmonary Fibrosis through TPH1 Knockout Mouse Experiments. *Mediators Inflamm.* **2018**, e7967868 (2018).

134. W. Wu, *et al.*, Impairment of Immune Response against Dematiaceous Fungi in Card9 Knockout Mice. *Mycopathologia* **181**, 631–642 (2016).

135. Z. Wang, *et al.*, Global Gene Expression and Focused Knockout Analysis Reveals Genes Associated with Fungal Fruiting Body Development in Neurospora crassa. *Eukaryot. Cell* **13**, 154–169 (2014).

136. C. S. Moffat, P. T. See, R. P. Oliver, Generation of a ToxA knockout strain of the wheat tan spot pathogen Pyrenophora tritici-repentis. *Mol. Plant Pathol.* **15**, 918–926 (2014).

137. J. Lu, H. Cao, L. Zhang, P. Huang, F. Lin, Systematic Analysis of Zn2Cys6 Transcription Factors Required for Development and Pathogenicity by High-Throughput Gene Knockout in the Rice Blast Fungus. *PLOS Pathog.* **10**, e1004432 (2014).

138. D. Segrè, A. DeLuna, G. M. Church, R. Kishony, Modular epistasis in yeast metabolism. *Nat. Genet.* **37**, 77–83 (2005).

139. D. Deutscher, I. Meilijson, M. Kupiec, E. Ruppin, Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nat. Genet.* **38**, 993–998 (2006).

140. M. C. McDonald, P. S. Solomon, Just the surface: advances in the discovery and characterization of necrotrophic wheat effectors. *Curr. Opin. Microbiol.* **46**, 14–18 (2018).

141. J. W. Wu, *et al.*, Epistatic interaction between the lipase-encoding genes Pnpla2 and Lipe causes liposarcoma in mice. *PLOS Genet.* **13**, e1006716 (2017).

142. F. Papaleo, M. C. Burdick, J. H. Callicott, D. R. Weinberger, Epistatic interaction between COMT and DTNBP1 modulates prefrontal function in mice and in humans. *Mol. Psychiatry* **19**, 311–316 (2014).

143. S. Mukherjee, *et al.*, Genomes OnLine Database (GOLD) v.8: overview and updates. *Nucleic Acids Res.* **49**, D723–D733 (2021).

144. T. Y. James, *et al.*, Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature* **443**, 818 (2006).

145. M. Bastian, S. Heymann, M. Jacomy, others, Gephi: an open source software for exploring and manipulating networks. *ICWSM* **8**, 361–362 (2009).

Supplementary information

Table 2: Identifier, species name and database accession link of each of the 671 fungal genomes from the FGP used in this project.

FGP ID	Species Name	FGP Accession
Aaoar1	*Aaosphaeria arxii*	https://mycocosm.jgi.doe.gov/Aaoar1/Aaoar1.home.html
Absrep1	*Absidia repens*	https://mycocosm.jgi.doe.gov/Absrep1/Absrep1.home.html
Acain1	*Acaromyces ingoldii*	https://mycocosm.jgi.doe.gov/Acain1/Acain1.home.html
Acema1	*Acephala macrosclerotiorum*	https://mycocosm.jgi.doe.gov/Acema1/Acema1.home.html
Aciaci1	*Acidothrix acidophila*	https://mycocosm.jgi.doe.gov/Aciaci1/Aciaci1.home.html
Aciri1_iso	*Acidomyces richmondensis iso*	https://mycocosm.jgi.doe.gov/Aciri1_iso/Aciri1_iso.home.html
Aciri1_meta	*Acidomyces richmondensis meta*	https://mycocosm.jgi.doe.gov/Aciri1_meta/Aciri1_meta.home.html
Acral2	*Acremonium alcalophilum*	https://mycocosm.jgi.doe.gov/Acral2/Acral2.home.html
Acrst1	*Acremonium strictum*	https://mycocosm.jgi.doe.gov/Acrst1/Acrst1.home.html
Agabi_varbisH97_2	*Agaricus bisporus var bisporus*	https://mycocosm.jgi.doe.gov/Agabi_varbisH97_2/Agabi_varbisH97_2.home.html
Agahy1	*Agaricostilbum hyphaenes*	https://mycocosm.jgi.doe.gov/Agahy1/Agahy1.home.html
Agrped1	*Agrocybe pediades*	https://mycocosm.jgi.doe.gov/Agrped1/Agrped1.home.html
Allma1	*Allomyces macrogynus*	https://mycocosm.jgi.doe.gov/Allma1/Allma1.home.html
Altal1	*Alternaria alternata*	https://mycocosm.jgi.doe.gov/Altal1/Altal1.home.html
Amamu1	*Amanita muscaria*	https://mycocosm.jgi.doe.gov/Amamu1/Amamu1.home.html
Amath1	*Amanita thiersii*	https://mycocosm.jgi.doe.gov/Amath1/Amath1.home.html
Amnli1	*Amniculicola lignicola*	https://mycocosm.jgi.doe.gov/Amnli1/Amnli1.home.html
Amore1	*Amorphotheca resinae*	https://mycocosm.jgi.doe.gov/Amore1/Amore1.home.html
Ampqui1	*Ampelomyces quisqualis*	https://mycocosm.jgi.doe.gov/Ampqui1/Ampqui1.home.html
Anasp1	*Anaeromyces robustus*	https://mycocosm.jgi.doe.gov/Anasp1/Anasp1.home.html

Anobom1	*Anomoporia bombycina*	https://mycocosm.jgi.doe.gov/Anobom1/Anobom1.home.html
Antav1	*Anthostoma avocetta*	https://mycocosm.jgi.doe.gov/Antav1/Antav1.home.html
Antlo1	*Antonospora locustae*	https://mycocosm.jgi.doe.gov/Antlo1/Antlo1.home.html
Antsi1	*Antrodia sinuosa*	https://mycocosm.jgi.doe.gov/Antsi1/Antsi1.home.html
Apimo1	*Apiospora montagnei*	https://mycocosm.jgi.doe.gov/Apimo1/Apimo1.home.html
Aplpr1	*Aplosporella prunicola*	https://mycocosm.jgi.doe.gov/Aplpr1/Aplpr1.home.html
Armga1	*Armillaria gallica*	https://mycocosm.jgi.doe.gov/Armga1/Armga1.home.html
Armme1_1	*Armillaria mellea*	https://mycocosm.jgi.doe.gov/Armme1_1/Armme1_1.home.html
Armost1	*Armillaria ostoyae*	https://mycocosm.jgi.doe.gov/Armost1/Armost1.home.html
Artbe1	*Arthroderma benhamiae*	https://mycocosm.jgi.doe.gov/Artbe1/Artbe1.home.html
Artel1	*Artolenzites elegans 1663*	https://mycocosm.jgi.doe.gov/Artel1/Artel1.home.html
Artele1122_1	*Artolenzites elegans 1122*	https://mycocosm.jgi.doe.gov/Artele1122_1/Artele1122_1.home.html
Artfe1_2	*Arthroascus fermentans*	https://mycocosm.jgi.doe.gov/Artfe1_2/Artfe1_2.home.html
Artol1	*Arthrobotrys oligospora*	https://mycocosm.jgi.doe.gov/Artol1/Artol1.home.html
Arxad1	*Blastobotrys Arxula*	https://mycocosm.jgi.doe.gov/Arxad1/Arxad1.home.html
Ascim1	*Ascobolus immersus*	https://mycocosm.jgi.doe.gov/Ascim1/Ascim1.home.html
Ascni1	*Ascodesmis nigricans*	https://mycocosm.jgi.doe.gov/Ascni1/Ascni1.home.html
Ascru1	*Ascoidea rubescens*	https://mycocosm.jgi.doe.gov/Ascru1/Ascru1.home.html
Ascsa1	*Ascocoryne sarcoides*	https://mycocosm.jgi.doe.gov/Ascsa1/Ascsa1.home.html
Ashgo1_1	*Eremothecium gossypii*	https://mycocosm.jgi.doe.gov/Ashgo1_1/Ashgo1_1.home.html
Aspbr1	*Aspergillus brasiliensis*	https://mycocosm.jgi.doe.gov/Aspbr1/Aspbr1.home.html
Aspcam1	*Aspergillus campestris*	https://mycocosm.jgi.doe.gov/Aspcam1/Aspcam1.home.html
Aspcl1	*Aspergillus clavatus*	https://mycocosm.jgi.doe.gov/Aspcl1/Aspcl1.home.html
Aspfl1	*Aspergillus flavus*	https://mycocosm.jgi.doe.gov/Aspfl1/Aspfl1.home.html
Aspfo1	*Aspergillus luchuensis*	https://mycocosm.jgi.doe.gov/Aspfo1/Aspfo1.home.html

Aspfu_A1163_1	*Aspergillus fumigatus A1163*	https://mycocosm.jgi.doe.gov/Aspfu_A1163_1/Aspfu_A1163_1.home.html
Aspfu1	*Aspergillus fumigatus Af293*	https://mycocosm.jgi.doe.gov/Aspfu1/Aspfu1.home.html
Aspgl1	*Aspergillus glaucus*	https://mycocosm.jgi.doe.gov/Aspgl1/Aspgl1.home.html
Aspka1_1	*Aspergillus kawachii*	https://mycocosm.jgi.doe.gov/Aspka1_1/Aspka1_1.home.html
Aspni_bvT_1	*Aspergillus niger van Tieghem*	https://mycocosm.jgi.doe.gov/Aspni_bvT_1/Aspni_bvT_1.home.html
Aspni_DSM_1	*Aspergillus niger CBS*	https://mycocosm.jgi.doe.gov/Aspni_DSM_1/Aspni_DSM_1.home.html
Aspni_NRRL3_1	*Aspergillus niger NRRL3*	https://mycocosm.jgi.doe.gov/Aspni_NRRL3_1/Aspni_NRRL3_1.home.html
Aspni7	*Aspergillus niger ATCC*	https://mycocosm.jgi.doe.gov/Aspni7/Aspni7.home.html
Aspnid1	*Aspergillus nidulans*	https://mycocosm.jgi.doe.gov/Aspnid1/Aspnid1.home.html
Aspnov1	*Aspergillus novofumigatus*	https://mycocosm.jgi.doe.gov/Aspnov1/Aspnov1.home.html
Aspoch1	*Aspergillus ochraceoroseus*	https://mycocosm.jgi.doe.gov/Aspoch1/Aspoch1.home.html
Aspor1	*Aspergillus oryzae*	https://mycocosm.jgi.doe.gov/Aspor1/Aspor1.home.html
Aspph1	*Aspergillus phoenicis*	https://mycocosm.jgi.doe.gov/Aspph1/Aspph1.home.html
Aspste1	*Aspergillus steynii*	https://mycocosm.jgi.doe.gov/Aspste1/Aspste1.home.html
Aspsy1	*Aspergillus sydowii*	https://mycocosm.jgi.doe.gov/Aspsy1/Aspsy1.home.html
Aspte1	*Aspergillus terreus*	https://mycocosm.jgi.doe.gov/Aspte1/Aspte1.home.html
Asptu1	*Aspergillus tubingensis*	https://mycocosm.jgi.doe.gov/Asptu1/Asptu1.home.html
Aspve1	*Aspergillus versicolor*	https://mycocosm.jgi.doe.gov/Aspve1/Aspve1.home.html
Aspwe1	*Aspergillus wentii*	https://mycocosm.jgi.doe.gov/Aspwe1/Aspwe1.home.html
Aspzo1	*Aspergillus zonatus*	https://mycocosm.jgi.doe.gov/Aspzo1/Aspzo1.home.html
Atrsp2	*Atractiellales sp.*	https://mycocosm.jgi.doe.gov/Atrsp2/Atrsp2.home.html
Aulhe2	*Aulographum hederae*	https://mycocosm.jgi.doe.gov/Aulhe2/Aulhe2.home.html
Auramp1	*Auriculariopsis ampla*	https://mycocosm.jgi.doe.gov/Auramp1/Auramp1.home.html

Aurde3_1	*Auricularia subglabra*	https://mycocosm.jgi.doe.gov/Aurde3_1/Aurde3_1.home.html
Aurpu_var_mel1	*Aureobasidium pullulans var. Melanogenum*	https://mycocosm.jgi.doe.gov/Aurpu_var_mel1/Aurpu_var_mel1.home.html
Aurpu_var_nam1	*Aureobasidium pullulans var. Namibiae*	https://mycocosm.jgi.doe.gov/Aurpu_var_nam1/Aurpu_var_nam1.home.html
Aurpu_var_pul1	*Aureobasidium pullulans var. Pullulans*	https://mycocosm.jgi.doe.gov/Aurpu_var_pul1/Aurpu_var_pul1.home.html
Aurpu_var_sub1	*Aureobasidium pullulans var. Subglaciale*	https://mycocosm.jgi.doe.gov/Aurpu_var_sub1/Aurpu_var_sub1.home.html
Aurvu1	*Auriscalpium vulgare*	https://mycocosm.jgi.doe.gov/Aurvu1/Aurvu1.home.html
Babin1	*Babjeviella inositovora*	https://mycocosm.jgi.doe.gov/Babin1/Babin1.home.html
Bacci1	*Backusella circina*	https://mycocosm.jgi.doe.gov/Bacci1/Bacci1.home.html
Basme2finSC	*Basidiobolus meristosporus*	https://mycocosm.jgi.doe.gov/Basme2finSC/Basme2finSC.home.html
Basun1	*Basidioascus undulatus*	https://mycocosm.jgi.doe.gov/Basun1/Basun1.home.html
Bauco1	*Baudoinia compniacensis*	https://mycocosm.jgi.doe.gov/Bauco1/Bauco1.home.html
Beaba1	*Beauveria bassiana*	https://mycocosm.jgi.doe.gov/Beaba1/Beaba1.home.html
Bimnz1	*Bimuria novae-zelandiae*	https://mycocosm.jgi.doe.gov/Bimnz1/Bimnz1.home.html
Bjead1_1	*Bjerkandera adusta*	https://mycocosm.jgi.doe.gov/Bjead1_1/Bjead1_1.home.html
Blade1	*Blastomyces dermatitidis*	https://mycocosm.jgi.doe.gov/Blade1/Blade1.home.html
Blatri1	*Blakeslea trispora*	https://mycocosm.jgi.doe.gov/Blatri1/Blatri1.home.html
Blugr1	*Blumeria graminis*	https://mycocosm.jgi.doe.gov/Blugr1/Blugr1.home.html
Boled1	*Boletus edulis*	https://mycocosm.jgi.doe.gov/Boled1/Boled1.home.html
Bolvit1	*Bolbitius vitellinus*	https://mycocosm.jgi.doe.gov/Bolvit1/Bolvit1.home.html
Botbo1	*Botryobasidium botryosum*	https://mycocosm.jgi.doe.gov/Botbo1/Botbo1.home.html
Botci1	*Botrytis cinerea*	https://mycocosm.jgi.doe.gov/Botci1/Botci1.home.html
Botdo1_1	*Botryosphaeria dothidea*	https://mycocosm.jgi.doe.gov/Botdo1_1/Botdo1_1.home.html
Bulin1	*Bulgaria inquinans*	https://mycocosm.jgi.doe.gov/Bulin1/Bulin1.home.html

Bysci1	*Byssothecium circinans*	https://mycocosm.jgi.doe.gov/Bysci1/Bysci1.home.html
Cadsp1	*Cadophora sp.*	https://mycocosm.jgi.doe.gov/Cadsp1/Cadsp1.home.html
Calco1	*Calocera cornea*	https://mycocosm.jgi.doe.gov/Calco1/Calco1.home.html
Calful1	*Caloscypha fulgens*	https://mycocosm.jgi.doe.gov/Calful1/Calful1.home.html
Calor1	*Caliciopsis orientalis*	https://mycocosm.jgi.doe.gov/Calor1/Calor1.home.html
Calpu1	*Calosphaeria pulchella*	https://mycocosm.jgi.doe.gov/Calpu1/Calpu1.home.html
Calvi1	*Calocera viscosa*	https://mycocosm.jgi.doe.gov/Calvi1/Calvi1.home.html
Cananz1	*Cantharellus anzutake*	https://mycocosm.jgi.doe.gov/Cananz1/Cananz1.home.html
Canar1	*Candida arabinofermentans*	https://mycocosm.jgi.doe.gov/Canar1/Canar1.home.html
Canca1	*Tortispora caseinolytica*	https://mycocosm.jgi.doe.gov/Canca1/Canca1.home.html
Canta1	*Candida tanzawaensis*	https://mycocosm.jgi.doe.gov/Canta1/Canta1.home.html
Capcor1	*Capronia coronata*	https://mycocosm.jgi.doe.gov/Capcor1/Capcor1.home.html
Capep1	*Capronia epimyces*	https://mycocosm.jgi.doe.gov/Capep1/Capep1.home.html
Capse1	*Capronia semiimmersa*	https://mycocosm.jgi.doe.gov/Capse1/Capse1.home.html
Catan2	*Catenaria anguillulae*	https://mycocosm.jgi.doe.gov/Catan2/Catan2.home.html
Cenge3	*Cenococcum geophilum*	https://mycocosm.jgi.doe.gov/Cenge3/Cenge3.home.html
Cepal1_1	*Cephaloascus albidus*	https://mycocosm.jgi.doe.gov/Cepal1_1/Cepal1_1.home.html
Cepfr1_1	*Cephaloascus fragrans*	https://mycocosm.jgi.doe.gov/Cepfr1_1/Cepfr1_1.home.html
CerAGI	*Ceratobasidium sp.*	https://mycocosm.jgi.doe.gov/CerAGI/CerAGI.home.html
Cercer1	*Cerinomyces ceraceus*	https://mycocosm.jgi.doe.gov/Cercer1/Cercer1.home.html
Cersp1	*Ceraceosorus sp.*	https://mycocosm.jgi.doe.gov/Cersp1/Cersp1.home.html
Cersu1	*Ceriporiopsis Gelatoporia*	https://mycocosm.jgi.doe.gov/Cersu1/Cersu1.home.html
Cerun2	*Cerrena unicolor*	https://mycocosm.jgi.doe.gov/Cerun2/Cerun2.home.html
Cerzm1	*Cercospora zeae-maydis*	https://mycocosm.jgi.doe.gov/Cerzm1/Cerzm1.home.html
Chagl_1	*Chaetomium globosum*	https://mycocosm.jgi.doe.gov/Chagl_1/Chagl_1.home.html

Chalo1	*Chalara longipes*	https://mycocosm.jgi.doe.gov/Chalo1/Chalo1.home.html
Chiap1	*Chionosphaera apobasidialis*	https://mycocosm.jgi.doe.gov/Chiap1/Chiap1.home.html
Chicu1	*Chionosphaera cuniculicola*	https://mycocosm.jgi.doe.gov/Chicu1/Chicu1.home.html
Chlpad1	*Absidia padenii*	https://mycocosm.jgi.doe.gov/Chlpad1/Chlpad1.home.html
Chove1	*Choiromyces venosus*	https://mycocosm.jgi.doe.gov/Chove1/Chove1.home.html
Claba1	*Cladophialophora bantiana*	https://mycocosm.jgi.doe.gov/Claba1/Claba1.home.html
Claca1	*Cladophialophora carrionii*	https://mycocosm.jgi.doe.gov/Claca1/Claca1.home.html
Clael1	*Clathrospora elynae*	https://mycocosm.jgi.doe.gov/Clael1/Clael1.home.html
Clafu1	*Cladosporium fulvum*	https://mycocosm.jgi.doe.gov/Clafu1/Clafu1.home.html
Clagr3	*Cladonia grayi*	https://mycocosm.jgi.doe.gov/Clagr3/Clagr3.home.html
Claim1	*Cladophialophora immunda*	https://mycocosm.jgi.doe.gov/Claim1/Claim1.home.html
Clalu1_2	*Clavispora lusitaniae*	https://mycocosm.jgi.doe.gov/Clalu1_2/Clalu1_2.home.html
ClaPMI390	*Clavulina sp.*	https://mycocosm.jgi.doe.gov/ClaPMI390/ClaPMI390.home.html
Claps1	*Cladophialophora psammophila*	https://mycocosm.jgi.doe.gov/Claps1/Claps1.home.html
Clapy1	*Clavicorona pyxidata*	https://mycocosm.jgi.doe.gov/Clapy1/Clapy1.home.html
Claye1	*Cladophialophora yegresii*	https://mycocosm.jgi.doe.gov/Claye1/Claye1.home.html
Cloaq1	*Clohesyomyces aquaticus*	https://mycocosm.jgi.doe.gov/Cloaq1/Cloaq1.home.html
Cloro1	*Clonostachys rosea*	https://mycocosm.jgi.doe.gov/Cloro1/Cloro1.home.html
Cocca1	*Cochliobolus carbonum*	https://mycocosm.jgi.doe.gov/Cocca1/Cocca1.home.html
CocheC4_1	*Cochliobolus heterostrophus C4*	https://mycocosm.jgi.doe.gov/CocheC4_1/CocheC4_1.home.html
CocheC5_3	*Cochliobolus heterostrophus C5*	https://mycocosm.jgi.doe.gov/CocheC5_3/CocheC5_3.home.html
Cocim1	*Coccidioides immitis*	https://mycocosm.jgi.doe.gov/Cocim1/Cocim1.home.html
Coclu2	*Cochliobolus lunatus*	https://mycocosm.jgi.doe.gov/Coclu2/Coclu2.home.html
Cocmi1	*Cochliobolus miyabeanus*	https://mycocosm.jgi.doe.gov/Cocmi1/Cocmi1.home.html
Cocsa1	*Cochliobolus sativus*	https://mycocosm.jgi.doe.gov/Cocsa1/Cocsa1.home.html

Cocst1	*Coccomyces strobi*	https://mycocosm.jgi.doe.gov/Cocst1/Cocst1.home.html
Cocvi1	*Cochliobolus victoriae*	https://mycocosm.jgi.doe.gov/Cocvi1/Cocvi1.home.html
Coere1	*Coemansia reversa*	https://mycocosm.jgi.doe.gov/Coere1/Coere1.home.html
Colca1	*Colletotrichum caudatum*	https://mycocosm.jgi.doe.gov/Colca1/Colca1.home.html
Coler1	*Colletotrichum eremochloae*	https://mycocosm.jgi.doe.gov/Coler1/Coler1.home.html
Colfa1	*Colletotrichum falcatum*	https://mycocosm.jgi.doe.gov/Colfa1/Colfa1.home.html
Colfi1	*Colletotrichum fioriniae*	https://mycocosm.jgi.doe.gov/Colfi1/Colfi1.home.html
Colgo1	*Colletotrichum godetiae*	https://mycocosm.jgi.doe.gov/Colgo1/Colgo1.home.html
Colgr1	*Colletotrichum graminicola*	https://mycocosm.jgi.doe.gov/Colgr1/Colgr1.home.html
Colhi1	*Colletotrichum higginsianum*	https://mycocosm.jgi.doe.gov/Colhi1/Colhi1.home.html
Colhig1	*Colletotrichum higginsianum g*	https://mycocosm.jgi.doe.gov/Colhig1/Colhig1.home.html
Collu1	*Colletotrichum lupini*	https://mycocosm.jgi.doe.gov/Collu1/Collu1.home.html
Colny1	*Colletotrichum nymphaeae*	https://mycocosm.jgi.doe.gov/Colny1/Colny1.home.html
Colph1	*Colletotrichum phormii*	https://mycocosm.jgi.doe.gov/Colph1/Colph1.home.html
Colsa1	*Colletotrichum salicis*	https://mycocosm.jgi.doe.gov/Colsa1/Colsa1.home.html
Colsi1	*Colletotrichum simmondsii*	https://mycocosm.jgi.doe.gov/Colsi1/Colsi1.home.html
Colso1	*Colletotrichum somersetensis*	https://mycocosm.jgi.doe.gov/Colso1/Colso1.home.html
Colsu1	*Colletotrichum sublineola*	https://mycocosm.jgi.doe.gov/Colsu1/Colsu1.home.html
Colzo1	*Colletotrichum zoysiae*	https://mycocosm.jgi.doe.gov/Colzo1/Colzo1.home.html
Conap1	*Coniosporium apollinis*	https://mycocosm.jgi.doe.gov/Conap1/Conap1.home.html
Conco1	*Conidiobolus coronatus*	https://mycocosm.jgi.doe.gov/Conco1/Conco1.home.html
Conli1	*Coniochaeta ligniaria CBS*	https://mycocosm.jgi.doe.gov/Conli1/Conli1.home.html
Conlig1	*Coniochaeta ligniaria NRRL*	https://mycocosm.jgi.doe.gov/Conlig1/Conlig1.home.html
Conol1	*Coniophora olivacea*	https://mycocosm.jgi.doe.gov/Conol1/Conol1.home.html
ConPMI546	*Coniochaeta sp.*	https://mycocosm.jgi.doe.gov/ConPMI546/ConPMI546.home.html

Conpu1	*Coniophora puteana*	https://mycocosm.jgi.doe.gov/Conpu1/Conpu1.home.html
Conth1	*Conidiobolus thromboides*	https://mycocosm.jgi.doe.gov/Conth1/Conth1.home.html
Copci_AmutBmut1	*Coprinopsis cinerea AmutBmut*	https://mycocosm.jgi.doe.gov/Copci_AmutBmut1/Copci_AmutBmut1.home.html
Copci1	*Coprinopsis cinerea*	https://mycocosm.jgi.doe.gov/Copci1/Copci1.home.html
Copmar1	*Coprinopsis marcescibilis*	https://mycocosm.jgi.doe.gov/Copmar1/Copmar1.home.html
Copmic2	*Coprinellus micaceus*	https://mycocosm.jgi.doe.gov/Copmic2/Copmic2.home.html
Coppel1	*Coprinellus pellucidus*	https://mycocosm.jgi.doe.gov/Coppel1/Coppel1.home.html
Copscl1	*Coprinopsis sclerotiger*	https://mycocosm.jgi.doe.gov/Copscl1/Copscl1.home.html
Corca1	*Corynespora cassiicola*	https://mycocosm.jgi.doe.gov/Corca1/Corca1.home.html
Corgl3	*Cortinarius glaucopus*	https://mycocosm.jgi.doe.gov/Corgl3/Corgl3.home.html
Corma2	*Corollospora maritima*	https://mycocosm.jgi.doe.gov/Corma2/Corma2.home.html
Cormi1	*Cordyceps militaris*	https://mycocosm.jgi.doe.gov/Cormi1/Cormi1.home.html
Crevar1	*Crepidotus variabilis*	https://mycocosm.jgi.doe.gov/Crevar1/Crevar1.home.html
Croqu1	*Cronartium quercuum*	https://mycocosm.jgi.doe.gov/Croqu1/Croqu1.home.html
Crula1	*Crucibulum laeve*	https://mycocosm.jgi.doe.gov/Crula1/Crula1.home.html
Crycu1	*Cryptococcus curvatus*	https://mycocosm.jgi.doe.gov/Crycu1/Crycu1.home.html
Cryne_JEC21_1	*Cryptococcus neoformans var neoformans*	https://mycocosm.jgi.doe.gov/Cryne_JEC21_1/Cryne_JEC21_1.home.html
Crypa2	*Cryphonectria parasitica*	https://mycocosm.jgi.doe.gov/Crypa2/Crypa2.home.html
Cryte1	*Cryptococcus terricola*	https://mycocosm.jgi.doe.gov/Cryte1/Cryte1.home.html
Cryvi1	*Cryptococcus vishniacii*	https://mycocosm.jgi.doe.gov/Cryvi1/Cryvi1.home.html
Crywi1	*Cryptococcus wieringae*	https://mycocosm.jgi.doe.gov/Crywi1/Crywi1.home.html
Cucbe1	*Cucurbitaria berberidis*	https://mycocosm.jgi.doe.gov/Cucbe1/Cucbe1.home.html
Cyastr2	*Cyathus striatus*	https://mycocosm.jgi.doe.gov/Cyastr2/Cyastr2.home.html
Cybja1	*Cyberlindnera jadinii*	https://mycocosm.jgi.doe.gov/Cybja1/Cybja1.home.html

Cylto1	*Cylindrobasidium torrendii*	https://mycocosm.jgi.doe.gov/Cylto1/Cylto1.home.html
Cypeu1	*Cyphellophora europaea*	https://mycocosm.jgi.doe.gov/Cypeu1/Cypeu1.home.html
Dacsp1	*Dacryopinax primogenitus*	https://mycocosm.jgi.doe.gov/Dacsp1/Dacsp1.home.html
Daequ1	*Daedalea quercina*	https://mycocosm.jgi.doe.gov/Daequ1/Daequ1.home.html
DalEC12_1	*Daldinia eschscholzii*	https://mycocosm.jgi.doe.gov/DalEC12_1/DalEC12_1.home.html
Debha1	*Debaryomyces hansenii*	https://mycocosm.jgi.doe.gov/Debha1/Debha1.home.html
Decga1	*Decorospora gaudefroyi*	https://mycocosm.jgi.doe.gov/Decga1/Decga1.home.html
Dekbr2	*Dekkera bruxellensis*	https://mycocosm.jgi.doe.gov/Dekbr2/Dekbr2.home.html
Delco1	*Delitschia confertaspora*	https://mycocosm.jgi.doe.gov/Delco1/Delco1.home.html
Denbi1	*Dendrothele bispora*	https://mycocosm.jgi.doe.gov/Denbi1/Denbi1.home.html
Densp1	*Dentipellis sp.*	https://mycocosm.jgi.doe.gov/Densp1/Densp1.home.html
Diaam1	*Diaporthe ampelina*	https://mycocosm.jgi.doe.gov/Diaam1/Diaam1.home.html
Dicsq1	*Dichomitus squalens LYAD-421*	https://mycocosm.jgi.doe.gov/Dicsq1/Dicsq1.home.html
Dicsqu18370_1	*Dichomitus squalens OM18370.1*	https://mycocosm.jgi.doe.gov/Dicsqu18370_1/Dicsqu18370_1.home.html
Dicsqu463_1	*Dichomitus squalens CBS463*	https://mycocosm.jgi.doe.gov/Dicsqu463_1/Dicsqu463_1.home.html
Dicsqu464_1	*Dichomitus squalens CBS464*	https://mycocosm.jgi.doe.gov/Dicsqu464_1/Dicsqu464_1.home.html
Didex1	*Didymella exigua*	https://mycocosm.jgi.doe.gov/Didex1/Didex1.home.html
Didma1	*Didymella zeae-maydis*	https://mycocosm.jgi.doe.gov/Didma1/Didma1.home.html
Diocr1	*Dioszegia cryoxerica*	https://mycocosm.jgi.doe.gov/Diocr1/Diocr1.home.html
Dipse1	*Diplodia seriata*	https://mycocosm.jgi.doe.gov/Dipse1/Dipse1.home.html
Disac1	*Dissoconium aciculare*	https://mycocosm.jgi.doe.gov/Disac1/Disac1.home.html
Dotse1	*Dothistroma septosporum*	https://mycocosm.jgi.doe.gov/Dotse1/Dotse1.home.html
Dotsy1	*Dothidotthia symphoricarpi*	https://mycocosm.jgi.doe.gov/Dotsy1/Dotsy1.home.html
Elmca1	*Aporpium caryae*	https://mycocosm.jgi.doe.gov/Elmca1/Elmca1.home.html

Elsamp1	*Elsinoe ampelina*	https://mycocosm.jgi.doe.gov/Elsamp1/Elsamp1.home.html
Enccu1	*Encephalitozoon cuniculi*	https://mycocosm.jgi.doe.gov/Enccu1/Enccu1.home.html
Enche1	*Encephalitozoon hellem*	https://mycocosm.jgi.doe.gov/Enche1/Enche1.home.html
Encin1	*Encephalitozoon intestinalis*	https://mycocosm.jgi.doe.gov/Encin1/Encin1.home.html
Encro1	*Encephalitozoon romaleae*	https://mycocosm.jgi.doe.gov/Encro1/Encro1.home.html
Entbi1	*Enterocytozoon bieneusi*	https://mycocosm.jgi.doe.gov/Entbi1/Entbi1.home.html
Erebi1	*Eremomyces bilateralis*	https://mycocosm.jgi.doe.gov/Erebi1/Erebi1.home.html
Eryha1	*Erythrobasidium hasegawianum*	https://mycocosm.jgi.doe.gov/Eryha1/Eryha1.home.html
Eurhe1	*Eurotium rubrum*	https://mycocosm.jgi.doe.gov/Eurhe1/Eurhe1.home.html
Eutla1	*Eutypa lata*	https://mycocosm.jgi.doe.gov/Eutla1/Eutla1.home.html
Exigl1	*Exidia glandulosa*	https://mycocosm.jgi.doe.gov/Exigl1/Exigl1.home.html
Exoaq1	*Exophiala aquamarina*	https://mycocosm.jgi.doe.gov/Exoaq1/Exoaq1.home.html
Exode1	*Exophiala dermatitidis*	https://mycocosm.jgi.doe.gov/Exode1/Exode1.home.html
Exome1	*Exophiala mesophila*	https://mycocosm.jgi.doe.gov/Exome1/Exome1.home.html
Exool1	*Exophiala oligosperma*	https://mycocosm.jgi.doe.gov/Exool1/Exool1.home.html
Exosi1	*Exophiala sideris*	https://mycocosm.jgi.doe.gov/Exosi1/Exosi1.home.html
Exosp1	*Exophiala spinifera*	https://mycocosm.jgi.doe.gov/Exosp1/Exosp1.home.html
Exova1	*Exobasidium vaccinii*	https://mycocosm.jgi.doe.gov/Exova1/Exova1.home.html
Exoxe1	*Exophiala xenobiotica*	https://mycocosm.jgi.doe.gov/Exoxe1/Exoxe1.home.html
Felpe1	*Fellomyces penicillatus*	https://mycocosm.jgi.doe.gov/Felpe1/Felpe1.home.html
Fibin1	*Fibulobasidium inconspicuum*	https://mycocosm.jgi.doe.gov/Fibin1/Fibin1.home.html
Fibra1	*Fibroporia radiculosa*	https://mycocosm.jgi.doe.gov/Fibra1/Fibra1.home.html
Fibsp1	*Fibulorhizoctonia sp.*	https://mycocosm.jgi.doe.gov/Fibsp1/Fibsp1.home.html
Fishe1	*Fistulina hepatica*	https://mycocosm.jgi.doe.gov/Fishe1/Fishe1.home.html
Fomme1	*Fomitiporia mediterranea*	https://mycocosm.jgi.doe.gov/Fomme1/Fomme1.home.html

Fompi3	*Fomitopsis pinicola*	https://mycocosm.jgi.doe.gov/Fompi3/Fompi3.home.html
Fonmu1	*Fonsecaea multimorphosa*	https://mycocosm.jgi.doe.gov/Fonmu1/Fonmu1.home.html
Fonpe1	*Fonsecaea pedrosoi*	https://mycocosm.jgi.doe.gov/Fonpe1/Fonpe1.home.html
Fusfu1	*Fusarium fujikuroi*	https://mycocosm.jgi.doe.gov/Fusfu1/Fusfu1.home.html
Fusgr1	*Fusarium graminearum*	https://mycocosm.jgi.doe.gov/Fusgr1/Fusgr1.home.html
Fusox2	*Fusarium oxysporum*	https://mycocosm.jgi.doe.gov/Fusox2/Fusox2.home.html
Fusve2	*Fusarium verticillioides*	https://mycocosm.jgi.doe.gov/Fusve2/Fusve2.home.html
Gaegr1	*Gaeumannomyces graminis var tritici*	https://mycocosm.jgi.doe.gov/Gaegr1/Gaegr1.home.html
Galma1	*Galerina marginata*	https://mycocosm.jgi.doe.gov/Galma1/Galma1.home.html
Ganpr1	*Gonapodya prolifera*	https://mycocosm.jgi.doe.gov/Ganpr1/Ganpr1.home.html
Gansp1	*Ganoderma sp.*	https://mycocosm.jgi.doe.gov/Gansp1/Gansp1.home.html
Gaumor1_1	*Gautieria morchelliformis*	https://mycocosm.jgi.doe.gov/Gaumor1_1/Gaumor1_1.home.html
Glalo1	*Glarea lozoyensis*	https://mycocosm.jgi.doe.gov/Glalo1/Glalo1.home.html
Gloac1	*Glomerella acutata*	https://mycocosm.jgi.doe.gov/Gloac1/Gloac1.home.html
Gloci1	*Glomerella cingulata*	https://mycocosm.jgi.doe.gov/Gloci1/Gloci1.home.html
Gloin1	*Rhizophagus irregularis*	https://mycocosm.jgi.doe.gov/Gloin1/Gloin1.home.html
Glopol1	*Globomyces pollinis-pini*	https://mycocosm.jgi.doe.gov/Glopol1/Glopol1.home.html
Glost2	*Glonium stellatum*	https://mycocosm.jgi.doe.gov/Glost2/Glost2.home.html
Glotr1_1	*Gloeophyllum trabeum*	https://mycocosm.jgi.doe.gov/Glotr1_1/Glotr1_1.home.html
Gonbut1	*Gongronella butleri*	https://mycocosm.jgi.doe.gov/Gonbut1/Gonbut1.home.html
Grocl1	*Grosmannia clavigera*	https://mycocosm.jgi.doe.gov/Grocl1/Grocl1.home.html
Guyne1	*Guyanagaster necrorhiza*	https://mycocosm.jgi.doe.gov/Guyne1/Guyne1.home.html
Gyman1	*Gymnopus androsaceus*	https://mycocosm.jgi.doe.gov/Gyman1/Gyman1.home.html
Gymau1	*Gymnascella aurantiaca*	https://mycocosm.jgi.doe.gov/Gymau1/Gymau1.home.html
Gymch1	*Gymnopilus chrysopellus*	https://mycocosm.jgi.doe.gov/Gymch1/Gymch1.home.html

Gymci1_1	*Gymnascella citrina*	https://mycocosm.jgi.doe.gov/Gymci1_1/Gymci1_1.home.html
Gymjun1	*Gymnopilus junonius*	https://mycocosm.jgi.doe.gov/Gymjun1/Gymjun1.home.html
Gymlu1	*Gymnopus luxurians*	https://mycocosm.jgi.doe.gov/Gymlu1/Gymlu1.home.html
Gyresc1	*Gyromitra esculenta*	https://mycocosm.jgi.doe.gov/Gyresc1/Gyresc1.home.html
Gyrli1	*Gyrodon lividus*	https://mycocosm.jgi.doe.gov/Gyrli1/Gyrli1.home.html
Hanpo2	*Ogataea polymorpha*	https://mycocosm.jgi.doe.gov/Hanpo2/Hanpo2.home.html
Hebcy2	*Hebeloma cylindrosporum*	https://mycocosm.jgi.doe.gov/Hebcy2/Hebcy2.home.html
Helsul1	*Heliocybe sulcata*	https://mycocosm.jgi.doe.gov/Helsul1/Helsul1.home.html
Hesve2finisherSC	*Hesseltinella vesiculosa*	https://mycocosm.jgi.doe.gov/Hesve2finisherSC/Hesve2finisherSC.home.html
Hetan2	*Heterobasidion annosum*	https://mycocosm.jgi.doe.gov/Hetan2/Hetan2.home.html
Hetpy1	*Heterogastridium pycnidioideum*	https://mycocosm.jgi.doe.gov/Hetpy1/Hetpy1.home.html
Hisca1	*Histoplasma capsulatum*	https://mycocosm.jgi.doe.gov/Hisca1/Hisca1.home.html
Horac1	*Hortaea acidophila*	https://mycocosm.jgi.doe.gov/Horac1/Horac1.home.html
Hyabl1	*Hyalopycnis blepharistoma*	https://mycocosm.jgi.doe.gov/Hyabl1/Hyabl1.home.html
Hydfim1	*Hydnopolyporus fimbriatus*	https://mycocosm.jgi.doe.gov/Hydfim1/Hydfim1.home.html
Hydpi2	*Hydnomerulius pinastri*	https://mycocosm.jgi.doe.gov/Hydpi2/Hydpi2.home.html
Hydru2	*Hydnum rufescens*	https://mycocosm.jgi.doe.gov/Hydru2/Hydru2.home.html
Hymrad1	*Hymenopellis radicata*	https://mycocosm.jgi.doe.gov/Hymrad1/Hymrad1.home.html
Hypbu1	*Hyphopichia burtonii*	https://mycocosm.jgi.doe.gov/Hypbu1/Hypbu1.home.html
HypCI4A_1	*Hypoxylon sp. CI-4A*	https://mycocosm.jgi.doe.gov/HypCI4A_1/HypCI4A_1.home.html
HypCO275_1	*Hypoxylon sp. CO27*	https://mycocosm.jgi.doe.gov/HypCO275_1/HypCO275_1.home.html
HypEC38_3	*Hypoxylon sp. EC38*	https://mycocosm.jgi.doe.gov/HypEC38_3/HypEC38_3.home.html
Hypsu1	*Hypholoma sublateritium*	https://mycocosm.jgi.doe.gov/Hypsu1/Hypsu1.home.html
Hyspu1_1	*Hysterium pulicare*	https://mycocosm.jgi.doe.gov/Hyspu1_1/Hyspu1_1.home.html

Ilysp1	*Ilyonectria sp.*	https://mycocosm.jgi.doe.gov/Ilysp1/Ilysp1.home.html
Jaaar1	*Jaapia argillacea*	https://mycocosm.jgi.doe.gov/Jaaar1/Jaaar1.home.html
Jamsp1	*Jaminaea sp.*	https://mycocosm.jgi.doe.gov/Jamsp1/Jamsp1.home.html
Kalpfe1	*Kalaharituber pfeilii*	https://mycocosm.jgi.doe.gov/Kalpfe1/Kalpfe1.home.html
Karrh1	*Karstenula rhodostoma*	https://mycocosm.jgi.doe.gov/Karrh1/Karrh1.home.html
Khuory1	*Khuskia oryzae*	https://mycocosm.jgi.doe.gov/Khuory1/Khuory1.home.html
Kircor1	*Mucor cordense*	https://mycocosm.jgi.doe.gov/Kircor1/Kircor1.home.html
Klula1	*Kluyveromyces lactis*	https://mycocosm.jgi.doe.gov/Klula1/Klula1.home.html
Kocim1	*Kockovaella imperatae*	https://mycocosm.jgi.doe.gov/Kocim1/Kocim1.home.html
Kurca1	*Kuraishia capsulata*	https://mycocosm.jgi.doe.gov/Kurca1/Kurca1.home.html
Lacam2	*Laccaria amethystina*	https://mycocosm.jgi.doe.gov/Lacam2/Lacam2.home.html
Lacbi2	*Laccaria bicolor*	https://mycocosm.jgi.doe.gov/Lacbi2/Lacbi2.home.html
Lacqui1	*Lactarius quietus*	https://mycocosm.jgi.doe.gov/Lacqui1/Lacqui1.home.html
Laesu1	*Laetiporus sulphureus var sulphureus*	https://mycocosm.jgi.doe.gov/Laesu1/Laesu1.home.html
Leisp1	*Leiotrametes sp.*	https://mycocosm.jgi.doe.gov/Leisp1/Leisp1.home.html
Lenfl1	*Lentithecium fluviatile*	https://mycocosm.jgi.doe.gov/Lenfl1/Lenfl1.home.html
Lenti6_1	*Lentinus tigrinus 6*	https://mycocosm.jgi.doe.gov/Lenti6_1/Lenti6_1.home.html
Lenti7_1	*Lentinus tigrinus 7*	https://mycocosm.jgi.doe.gov/Lenti7_1/Lenti7_1.home.html
Lenvul1	*Lentinellus vulpinus*	https://mycocosm.jgi.doe.gov/Lenvul1/Lenvul1.home.html
Lepmu1	*Leptosphaeria maculans*	https://mycocosm.jgi.doe.gov/Lepmu1/Lepmu1.home.html
Lepnud1	*Lepista nuda*	https://mycocosm.jgi.doe.gov/Lepnud1/Lepnud1.home.html
Leppa1	*Lepidopterella palustris*	https://mycocosm.jgi.doe.gov/Leppa1/Leppa1.home.html
Leptod1	*Leptodontium sp.*	https://mycocosm.jgi.doe.gov/Leptod1/Leptod1.home.html
Leucr1	*Leucosporidiella creatinivora*	https://mycocosm.jgi.doe.gov/Leucr1/Leucr1.home.html
Leugo1_1	*Leucoagaricus gongylophorus*	https://mycocosm.jgi.doe.gov/Leugo1_1/Leugo1_1.home.html

Leumo1	*Leucogyrophana mollusca*	https://mycocosm.jgi.doe.gov/Leumo1/Leumo1.home.html
Liccor1	*Lichtheimia corymbifera*	https://mycocosm.jgi.doe.gov/Liccor1/Liccor1.home.html
Lichy1	*Lichtheimia hyalospora*	https://mycocosm.jgi.doe.gov/Lichy1/Lichy1.home.html
Linin1	*Lindgomyces ingoldianus*	https://mycocosm.jgi.doe.gov/Linin1/Linin1.home.html
Linpe1	*Linderina pennispora*	https://mycocosm.jgi.doe.gov/Linpe1/Linpe1.home.html
Linrh1	*Lineolata rhizophorae*	https://mycocosm.jgi.doe.gov/Linrh1/Linrh1.home.html
Linth1	*Lindra thalassiae*	https://mycocosm.jgi.doe.gov/Linth1/Linth1.home.html
Lipst1_1	*Lipomyces starkeyi*	https://mycocosm.jgi.doe.gov/Lipst1_1/Lipst1_1.home.html
Lizem1	*Lizonia empirigonia*	https://mycocosm.jgi.doe.gov/Lizem1/Lizem1.home.html
Lobtra1	*Lobosporangium transversale*	https://mycocosm.jgi.doe.gov/Lobtra1/Lobtra1.home.html
Lolmi1	*Lollipopaia minuta*	https://mycocosm.jgi.doe.gov/Lolmi1/Lolmi1.home.html
Lopma1	*Lophiostoma macrostomum*	https://mycocosm.jgi.doe.gov/Lopma1/Lopma1.home.html
Lopmy1	*Lophium mytilinum*	https://mycocosm.jgi.doe.gov/Lopmy1/Lopmy1.home.html
Lopni1	*Peniophora sp. CONTA*	https://mycocosm.jgi.doe.gov/Lopni1/Lopni1.home.html
Lopnu1	*Lophiotrema nucula*	https://mycocosm.jgi.doe.gov/Lopnu1/Lopnu1.home.html
Lorju1	*Loramyces juncicola*	https://mycocosm.jgi.doe.gov/Lorju1/Lorju1.home.html
Lorma1	*Loramyces macrosporus*	https://mycocosm.jgi.doe.gov/Lorma1/Lorma1.home.html
Macan1	*Macroventuria anomochaeta*	https://mycocosm.jgi.doe.gov/Macan1/Macan1.home.html
Macfu1	*Macrolepiota fuliginosa*	https://mycocosm.jgi.doe.gov/Macfu1/Macfu1.home.html
Macph1	*Macrophomina phaseolina*	https://mycocosm.jgi.doe.gov/Macph1/Macph1.home.html
Maggr1	*Magnaporthe grisea*	https://mycocosm.jgi.doe.gov/Maggr1/Maggr1.home.html
Magpo1	*Magnaporthiopsis poae*	https://mycocosm.jgi.doe.gov/Magpo1/Magpo1.home.html
Malgl1	*Malassezia globosa*	https://mycocosm.jgi.doe.gov/Malgl1/Malgl1.home.html
Malsy1_1	*Malassezia sympodialis*	https://mycocosm.jgi.doe.gov/Malsy1_1/Malsy1_1.home.html
Marbr1	*Marssonina brunnea*	https://mycocosm.jgi.doe.gov/Marbr1/Marbr1.home.html

Marfi1	*Marasmius fiardii*	https://mycocosm.jgi.doe.gov/Marfi1/Marfi1.home.html
Marpt1	*Martensiomyces pterosporus*	https://mycocosm.jgi.doe.gov/Marpt1/Marpt1.home.html
Maseb1	*Massarina eburnea*	https://mycocosm.jgi.doe.gov/Maseb1/Maseb1.home.html
Meimi1	*Meira miltonrushii*	https://mycocosm.jgi.doe.gov/Meimi1/Meimi1.home.html
Melap1finSC_191	*Melampsora allii-populina*	https://mycocosm.jgi.doe.gov/Melap1finSC_191/Melap1finSC_191.home.html
Melbi2	*Meliniomyces bicolor*	https://mycocosm.jgi.doe.gov/Melbi2/Melbi2.home.html
Melen1	*Melanotaenium endogenum*	https://mycocosm.jgi.doe.gov/Melen1/Melen1.home.html
Melli1	*Melampsora lini*	https://mycocosm.jgi.doe.gov/Melli1/Melli1.home.html
Mellp2_3	*Melampsora laricis-populina*	https://mycocosm.jgi.doe.gov/Mellp2_3/Mellp2_3.home.html
Melpu1	*Melanomma pulvis-pyrius*	https://mycocosm.jgi.doe.gov/Melpu1/Melpu1.home.html
Melsp1	*Melanconium sp.*	https://mycocosm.jgi.doe.gov/Melsp1/Melsp1.home.html
Melti1	*Melanospora tiffanyae*	https://mycocosm.jgi.doe.gov/Melti1/Melti1.home.html
Melva1	*Meliniomyces variabilis*	https://mycocosm.jgi.doe.gov/Melva1/Melva1.home.html
Mereb1	*Meredithblackwellia eburnea*	https://mycocosm.jgi.doe.gov/Mereb1/Mereb1.home.html
Metac1	*Metarhizium acridum*	https://mycocosm.jgi.doe.gov/Metac1/Metac1.home.html
Metan1	*Metarhizium robertsii*	https://mycocosm.jgi.doe.gov/Metan1/Metan1.home.html
Metbi1	*Metschnikowia bicuspidata*	https://mycocosm.jgi.doe.gov/Metbi1/Metbi1.home.html
Meygui1	*Meyerozyma guilliermondii*	https://mycocosm.jgi.doe.gov/Meygui1/Meygui1.home.html
Micbo1	*Microdochium bolleyi*	https://mycocosm.jgi.doe.gov/Micbo1/Micbo1.home.html
Micca1	*Microsporum canis*	https://mycocosm.jgi.doe.gov/Micca1/Micca1.home.html
Micmi1	*Microthyrium microscopicum*	https://mycocosm.jgi.doe.gov/Micmi1/Micmi1.home.html
Mictr1	*Microascus trigonosporus*	https://mycocosm.jgi.doe.gov/Mictr1/Mictr1.home.html
Micvi1	*Microbotryum violaceum*	https://mycocosm.jgi.doe.gov/Micvi1/Micvi1.home.html
Mixos1	*Mixia osmundae*	https://mycocosm.jgi.doe.gov/Mixos1/Mixos1.home.html

Monha1	*Monacrosporium haptotylum*	https://mycocosm.jgi.doe.gov/Monha1/Monha1.home.html
Monpe1_1	*Moniliophthora perniciosa*	https://mycocosm.jgi.doe.gov/Monpe1_1/Monpe1_1.home.html
Monpu1	*Monascus purpureus*	https://mycocosm.jgi.doe.gov/Monpu1/Monpu1.home.html
Monru1	*Monascus ruber*	https://mycocosm.jgi.doe.gov/Monru1/Monru1.home.html
Morco1	*Morchella conica*	https://mycocosm.jgi.doe.gov/Morco1/Morco1.home.html
Morel2	*Mortierella elongata*	https://mycocosm.jgi.doe.gov/Morel2/Morel2.home.html
Morimp1	*Morchella importuna*	https://mycocosm.jgi.doe.gov/Morimp1/Morimp1.home.html
Mormul1	*Mortierella multidivaricata*	https://mycocosm.jgi.doe.gov/Mormul1/Mormul1.home.html
Morve1	*Mortierella verticillata*	https://mycocosm.jgi.doe.gov/Morve1/Morve1.home.html
Mrafri1	*Mrakia frigida*	https://mycocosm.jgi.doe.gov/Mrafri1/Mrafri1.home.html
Mucci2	*Mucor circinelloides*	https://mycocosm.jgi.doe.gov/Mucci2/Mucci2.home.html
Mutel1	*Mutinus elegans*	https://mycocosm.jgi.doe.gov/Mutel1/Mutel1.home.html
Mycafr1	*Mycotypha africana*	https://mycocosm.jgi.doe.gov/Mycafr1/Mycafr1.home.html
Mycfi2	*Pseudocercospora Mycosphaerella fijiensis*	https://mycocosm.jgi.doe.gov/Mycfi2/Mycfi2.home.html
Mycgal1	*Mycena galopus*	https://mycocosm.jgi.doe.gov/Mycgal1/Mycgal1.home.html
Myche1	*Myceliophthora heterothallica*	https://mycocosm.jgi.doe.gov/Myche1/Myche1.home.html
Myrdu1	*Myriangium duriaei*	https://mycocosm.jgi.doe.gov/Myrdu1/Myrdu1.home.html
Myrin1	*Myrothecium inundatum*	https://mycocosm.jgi.doe.gov/Myrin1/Myrin1.home.html
Mytre1	*Mytilinidion resinicola*	https://mycocosm.jgi.doe.gov/Mytre1/Mytre1.home.html
Nadfu1	*Nadsonia fulvescens var elongata*	https://mycocosm.jgi.doe.gov/Nadfu1/Nadfu1.home.html
Naifl1	*Naiadella fluitans*	https://mycocosm.jgi.doe.gov/Naifl1/Naifl1.home.html
Naose1	*Naohidea sebacea*	https://mycocosm.jgi.doe.gov/Naose1/Naose1.home.html
Necha2	*Nectria haematococca*	https://mycocosm.jgi.doe.gov/Necha2/Necha2.home.html
Nempa1	*Nematocida parisii*	https://mycocosm.jgi.doe.gov/Nempa1/Nempa1.home.html
Neofi1	*Neosartorya fischeri*	https://mycocosm.jgi.doe.gov/Neofi1/Neofi1.home.html

Neole1	*Neolentinus lepideus*	https://mycocosm.jgi.doe.gov/Neole1/Neole1.home.html
Neopa1	*Neofusicoccum parvum*	https://mycocosm.jgi.doe.gov/Neopa1/Neopa1.home.html
Neosp1	*Neocallimastix californiae*	https://mycocosm.jgi.doe.gov/Neosp1/Neosp1.home.html
Neucr_trp3_1	*Neurospora crassa FGSC*	https://mycocosm.jgi.doe.gov/Neucr_trp3_1/Neucr_trp3_1.home.html
Neucr2	*Neurospora crassa OR74A*	https://mycocosm.jgi.doe.gov/Neucr2/Neucr2.home.html
Neudi1	*Neurospora discreta*	https://mycocosm.jgi.doe.gov/Neudi1/Neudi1.home.html
Neute_mat_a1	*Neurospora tetrasperma 2509*	https://mycocosm.jgi.doe.gov/Neute_mat_a1/Neute_mat_a1.home.html
Nieex1	*Niesslia exilis*	https://mycocosm.jgi.doe.gov/Nieex1/Nieex1.home.html
Nosce1	*Nosema ceranae*	https://mycocosm.jgi.doe.gov/Nosce1/Nosce1.home.html
Obbri1	*Obba rivulosa*	https://mycocosm.jgi.doe.gov/Obbri1/Obbri1.home.html
Oidma1	*Oidiodendron maius*	https://mycocosm.jgi.doe.gov/Oidma1/Oidma1.home.html
Ompol1	*Omphalotus olearius*	https://mycocosm.jgi.doe.gov/Ompol1/Ompol1.home.html
Onnsc1	*Onnia scaura*	https://mycocosm.jgi.doe.gov/Onnsc1/Onnsc1.home.html
Ophdi1	*Ophiobolus disseminans*	https://mycocosm.jgi.doe.gov/Ophdi1/Ophdi1.home.html
Ophpi1	*Ophiostoma piliferum*	https://mycocosm.jgi.doe.gov/Ophpi1/Ophpi1.home.html
Ophpic1	*Ophiostoma piceae*	https://mycocosm.jgi.doe.gov/Ophpic1/Ophpic1.home.html
Oudmuc1	*Oudemansiella mucida*	https://mycocosm.jgi.doe.gov/Oudmuc1/Oudmuc1.home.html
Pacta1_2	*Pachysolen tannophilus*	https://mycocosm.jgi.doe.gov/Pacta1_2/Pacta1_2.home.html
Panru1	*Panus rudis*	https://mycocosm.jgi.doe.gov/Panru1/Panru1.home.html
Panst_KUC8834_1_1	*Panellus stipticus*	https://mycocosm.jgi.doe.gov/Panst_KUC8834_1_1/Panst_KUC8834_1_1.home.html
Panst_LUM_1_1	*Panellus stipticus LUM*	https://mycocosm.jgi.doe.gov/Panst_LUM_1_1/Panst_LUM_1_1.home.html
Parbr1	*Paracoccidioides brasiliensis Pb03*	https://mycocosm.jgi.doe.gov/Parbr1/Parbr1.home.html
Parbra1	*Paracoccidioides brasiliensis Pb18*	https://mycocosm.jgi.doe.gov/Parbra1/Parbra1.home.html

Parpar1	*Parasitella parasitica*	https://mycocosm.jgi.doe.gov/Parpar1/Parpar1.home.html
Parsp1	*Paraconiothyrium sporulosum*	https://mycocosm.jgi.doe.gov/Parsp1/Parsp1.home.html
Patat1	*Patellaria atrata*	https://mycocosm.jgi.doe.gov/Patat1/Patat1.home.html
Paxam1	*Paxillus ammoniavirescens*	https://mycocosm.jgi.doe.gov/Paxam1/Paxam1.home.html
Paxin1	*Paxillus involutus*	https://mycocosm.jgi.doe.gov/Paxin1/Paxin1.home.html
Paxru2	*Paxillus adelphus*	https://mycocosm.jgi.doe.gov/Paxru2/Paxru2.home.html
Penac1	*Talaromyces aculeatus*	https://mycocosm.jgi.doe.gov/Penac1/Penac1.home.html
Penbi1	*Penicillium bilaiae*	https://mycocosm.jgi.doe.gov/Penbi1/Penbi1.home.html
Penbr2	*Penicillium brevicompactum*	https://mycocosm.jgi.doe.gov/Penbr2/Penbr2.home.html
PenbrAgRF18_1	*Penicillium brevicompactum AgRF18*	https://mycocosm.jgi.doe.gov/PenbrAgRF18_1/PenbrAgRF18_1.home.html
Penca1	*Penicillium canescens*	https://mycocosm.jgi.doe.gov/Penca1/Penca1.home.html
Pench1	*Penicillium chrysogenum*	https://mycocosm.jgi.doe.gov/Pench1/Pench1.home.html
PenchWisc1_1	*Penicillium chrysogenum Wisconsin*	https://mycocosm.jgi.doe.gov/PenchWisc1_1/PenchWisc1_1.home.html
Pendi1	*Penicillium digitatum*	https://mycocosm.jgi.doe.gov/Pendi1/Pendi1.home.html
Penex1	*Penicillium expansum*	https://mycocosm.jgi.doe.gov/Penex1/Penex1.home.html
Penfe1	*Penicillium fellutanum*	https://mycocosm.jgi.doe.gov/Penfe1/Penfe1.home.html
Pengl1	*Penicillium glabrum*	https://mycocosm.jgi.doe.gov/Pengl1/Pengl1.home.html
Penja1	*Penicillium janthinellum*	https://mycocosm.jgi.doe.gov/Penja1/Penja1.home.html
Penla1	*Penicillium lanosocoeruleum*	https://mycocosm.jgi.doe.gov/Penla1/Penla1.home.html
Penox1	*Penicillium oxalicum*	https://mycocosm.jgi.doe.gov/Penox1/Penox1.home.html
Penra1	*Penicillium raistrickii*	https://mycocosm.jgi.doe.gov/Penra1/Penra1.home.html
Penth1	*Penicillium thymicola*	https://mycocosm.jgi.doe.gov/Penth1/Penth1.home.html
Perma1	*Periconia macrospinosa*	https://mycocosm.jgi.doe.gov/Perma1/Perma1.home.html
Phaal1	*Phaeoacremonium aleophilum*	https://mycocosm.jgi.doe.gov/Phaal1/Phaal1.home.html

Phaart1	*Phascolomyces articulosus*	https://mycocosm.jgi.doe.gov/Phaart1/Phaart1.home.html
Phaca1	*Phanerochaete carnosa*	https://mycocosm.jgi.doe.gov/Phaca1/Phaca1.home.html
Phach1	*Phaeomoniella chlamydospora*	https://mycocosm.jgi.doe.gov/Phach1/Phach1.home.html
PhaPMI808	*Phaeosphaeriaceae sp.*	https://mycocosm.jgi.doe.gov/PhaPMI808/PhaPMI808.home.html
Phchr2	*Phanerochaete chrysosporium*	https://mycocosm.jgi.doe.gov/Phchr2/Phchr2.home.html
Phisc1	*Phialocephala scopiformis*	https://mycocosm.jgi.doe.gov/Phisc1/Phisc1.home.html
Phlbr1	*Phlebia brevispora*	https://mycocosm.jgi.doe.gov/Phlbr1/Phlbr1.home.html
Phlgi1	*Phlebiopsis gigantea*	https://mycocosm.jgi.doe.gov/Phlgi1/Phlgi1.home.html
Phoaln1	*Pholiota alnicola*	https://mycocosm.jgi.doe.gov/Phoaln1/Phoaln1.home.html
Phocon1	*Pholiota conissans*	https://mycocosm.jgi.doe.gov/Phocon1/Phocon1.home.html
Photr1	*Phoma tracheiphila*	https://mycocosm.jgi.doe.gov/Photr1/Photr1.home.html
Phybl2	*Phycomyces blakesleeanus*	https://mycocosm.jgi.doe.gov/Phybl2/Phybl2.home.html
Phycit1	*Phyllosticta citriasiana*	https://mycocosm.jgi.doe.gov/Phycit1/Phycit1.home.html
Picme2	*Pichia membranifaciens*	https://mycocosm.jgi.doe.gov/Picme2/Picme2.home.html
Picpa1	*Pichia pastoris*	https://mycocosm.jgi.doe.gov/Picpa1/Picpa1.home.html
Pieho1_1	*Piedraia hortae*	https://mycocosm.jgi.doe.gov/Pieho1_1/Pieho1_1.home.html
Pilcr1	*Piloderma croceum*	https://mycocosm.jgi.doe.gov/Pilcr1/Pilcr1.home.html
PirE2_1	*Piromyces sp.*	https://mycocosm.jgi.doe.gov/PirE2_1/PirE2_1.home.html
Pirfi3	*Piromyces finnis*	https://mycocosm.jgi.doe.gov/Pirfi3/Pirfi3.home.html
Pirin1	*Piriformospora indica*	https://mycocosm.jgi.doe.gov/Pirin1/Pirin1.home.html
Pismi1	*Pisolithus microcarpus*	https://mycocosm.jgi.doe.gov/Pismi1/Pismi1.home.html
Pisti1	*Pisolithus tinctorius*	https://mycocosm.jgi.doe.gov/Pisti1/Pisti1.home.html
Plecu1	*Plectosphaerella cucumerina*	https://mycocosm.jgi.doe.gov/Plecu1/Plecu1.home.html
Pleery1	*Pleurotus eryngii*	https://mycocosm.jgi.doe.gov/Pleery1/Pleery1.home.html
Plemel1	*Plectania melastoma*	https://mycocosm.jgi.doe.gov/Plemel1/Plemel1.home.html

PleosPC15_2	*Pleurotus ostreatus PC15*	https://mycocosm.jgi.doe.gov/PleosPC15_2/PleosPC15_2.home.html
Plesi1	*Pleomassaria siparia*	https://mycocosm.jgi.doe.gov/Plesi1/Plesi1.home.html
Plicr1	*Plicaturopsis crispa*	https://mycocosm.jgi.doe.gov/Plicr1/Plicr1.home.html
Plucer1	*Pluteus cervinus*	https://mycocosm.jgi.doe.gov/Plucer1/Plucer1.home.html
Pneji1	*Pneumocystis jirovecii*	https://mycocosm.jgi.doe.gov/Pneji1/Pneji1.home.html
Podan2	*Podospora anserina*	https://mycocosm.jgi.doe.gov/Podan2/Podan2.home.html
Podcur1	*Podospora curvicolla*	https://mycocosm.jgi.doe.gov/Podcur1/Podcur1.home.html
Polar1	*Polyporus arcularius*	https://mycocosm.jgi.doe.gov/Polar1/Polar1.home.html
Polbr1	*Polyporus brumalis*	https://mycocosm.jgi.doe.gov/Polbr1/Polbr1.home.html
Polci1	*Polychaeton citri*	https://mycocosm.jgi.doe.gov/Polci1/Polci1.home.html
Polfu1	*Polyplosphaeria fusca*	https://mycocosm.jgi.doe.gov/Polfu1/Polfu1.home.html
Porchr1	*Porodaedalea chrysoloma*	https://mycocosm.jgi.doe.gov/Porchr1/Porchr1.home.html
Pornie1_2	*Porodaedalea niemelaei*	https://mycocosm.jgi.doe.gov/Pornie1_2/Pornie1_2.home.html
Pospl1	*Postia placenta*	https://mycocosm.jgi.doe.gov/Pospl1/Pospl1.home.html
PosplRSB12_1	*Postia placenta SB12*	https://mycocosm.jgi.doe.gov/PosplRSB12_1/PosplRSB12_1.home.html
Proin1	*Protomyces inouyei*	https://mycocosm.jgi.doe.gov/Proin1/Proin1.home.html
Prola1	*Protomyces lactucaedebilis*	https://mycocosm.jgi.doe.gov/Prola1/Prola1.home.html
Psean1_1	*Pseudozyma antarctica*	https://mycocosm.jgi.doe.gov/Psean1_1/Psean1_1.home.html
Pseel1	*Pseudographis elatina*	https://mycocosm.jgi.doe.gov/Pseel1/Pseel1.home.html
Psehu1	*Pseudozyma hubeiensis*	https://mycocosm.jgi.doe.gov/Psehu1/Psehu1.home.html
Psehy1	*Pseudovirgaria hyperparasitica*	https://mycocosm.jgi.doe.gov/Psehy1/Psehy1.home.html
Pseve2	*Pseudomassariella vexata*	https://mycocosm.jgi.doe.gov/Pseve2/Pseve2.home.html
Ptegra1	*Pterula gracilis*	https://mycocosm.jgi.doe.gov/Ptegra1/Ptegra1.home.html
Pucgr2	*Puccinia graminis*	https://mycocosm.jgi.doe.gov/Pucgr2/Pucgr2.home.html
Pucst_PST78_1	*Puccinia striiformis 78*	https://mycocosm.jgi.doe.gov/

		Pucst_PST78_1/ Pucst_PST78_1.home.html
Pucst1	*Puccinia striiformis 130*	https://mycocosm.jgi.doe.gov/Pucst1/ Pucst1.home.html
Puctr1	*Puccinia triticina*	https://mycocosm.jgi.doe.gov/Puctr1/ Puctr1.home.html
Punst1	*Punctularia strigosozonata*	https://mycocosm.jgi.doe.gov/Punst1/ Punst1.home.html
Pursp1	*Purpureocillium sp.*	https://mycocosm.jgi.doe.gov/Pursp1/ Pursp1.home.html
Pycci1	*Pycnoporus cinnabarinus*	https://mycocosm.jgi.doe.gov/Pycci1/ Pycci1.home.html
Pycco1	*Pycnoporus coccineus BRFM*	https://mycocosm.jgi.doe.gov/Pycco1/ Pycco1.home.html
Pycco1662_1	*Pycnoporus coccineus CIRM1662*	https://mycocosm.jgi.doe.gov/ Pycco1662_1/Pycco1662_1.home.html
Pycpun1	*Pycnoporus puniceus*	https://mycocosm.jgi.doe.gov/Pycpun1/ Pycpun1.home.html
Pycsa1	*Pycnoporus sanguineus*	https://mycocosm.jgi.doe.gov/Pycsa1/ Pycsa1.home.html
Pyrco1	*Pyronema confluens*	https://mycocosm.jgi.doe.gov/Pyrco1/ Pyrco1.home.html
Pyrsp1	*Pyrenochaeta sp.*	https://mycocosm.jgi.doe.gov/Pyrsp1/ Pyrsp1.home.html
Pyrtt1	*Pyrenophora teres*	https://mycocosm.jgi.doe.gov/Pyrtt1/ Pyrtt1.home.html
Ramac1	*Ramaria rubella*	https://mycocosm.jgi.doe.gov/Ramac1/ Ramac1.home.html
Rambr1	*Ramicandelaber brevisporus*	https://mycocosm.jgi.doe.gov/Rambr1/ Rambr1.home.html
Rhich1	*Rhizopus microsporus var chinensis*	https://mycocosm.jgi.doe.gov/Rhich1/ Rhich1.home.html
Rhier1	*Rhizoscyphus ericae*	https://mycocosm.jgi.doe.gov/Rhier1/ Rhier1.home.html
Rhihy1	*Rhizoclosmatium globosum*	https://mycocosm.jgi.doe.gov/Rhihy1/ Rhihy1.home.html
Rhili1	*Rhizodiscina lignyota*	https://mycocosm.jgi.doe.gov/Rhili1/ Rhili1.home.html
Rhimi_ATCC11559 _1	*Rhizopus microsporus*	https://mycocosm.jgi.doe.gov/ Rhimi_ATCC11559_1/ Rhimi_ATCC11559_1.home.html
Rhimi_ATCC52814 _1	*Rhizopus microsporus var microsporus 4*	https://mycocosm.jgi.doe.gov/ Rhimi_ATCC52814_1/ Rhimi_ATCC52814_1.home.html
Rhimi1_1	*Rhizopus microsporus var microsporus 3*	https://mycocosm.jgi.doe.gov/ Rhimi1_1/Rhimi1_1.home.html
Rhisa1	*Rhizopogon salebrosus*	https://mycocosm.jgi.doe.gov/Rhisa1/ Rhisa1.home.html
Rhiso1	*Rhizoctonia solani*	https://mycocosm.jgi.doe.gov/Rhiso1/ Rhiso1.home.html

Rhivi1	*Rhizopogon vinicolor*	https://mycocosm.jgi.doe.gov/Rhivi1/Rhivi1.home.html
Rhoba1_1	*Rhodotorula graminis*	https://mycocosm.jgi.doe.gov/Rhoba1_1/Rhoba1_1.home.html
Rhobut1_1	*Rhodocollybia butyracea*	https://mycocosm.jgi.doe.gov/Rhobut1_1/Rhobut1_1.home.html
Rhodsp1	*Microstromatales sp.*	https://mycocosm.jgi.doe.gov/Rhodsp1/Rhodsp1.home.html
Rhomi1	*Rhodotorula minuta*	https://mycocosm.jgi.doe.gov/Rhomi1/Rhomi1.home.html
Rhosp1	*Rhodotorula sp.*	https://mycocosm.jgi.doe.gov/Rhosp1/Rhosp1.home.html
Rhoto_IFO0559_1	*Rhodosporidium toruloides IFO0559*	https://mycocosm.jgi.doe.gov/Rhoto_IFO0559_1/Rhoto_IFO0559_1.home.html
Rhoto_IFO0880_2	*Rhodosporidium toruloides IFO0880*	https://mycocosm.jgi.doe.gov/Rhoto_IFO0880_2/Rhoto_IFO0880_2.home.html
Rhoto_IFO1236_1	*Rhodosporidium toruloides IFO1236*	https://mycocosm.jgi.doe.gov/Rhoto_IFO1236_1/Rhoto_IFO1236_1.home.html
Rhoto1	*Rhodosporidium toruloides NP11*	https://mycocosm.jgi.doe.gov/Rhoto1/Rhoto1.home.html
Rhyru1_1	*Rhytidhysteron rufulum*	https://mycocosm.jgi.doe.gov/Rhyru1_1/Rhyru1_1.home.html
Ricfib1	*Rickenella fibula*	https://mycocosm.jgi.doe.gov/Ricfib1/Ricfib1.home.html
Ricme1	*Peniophora sp.*	https://mycocosm.jgi.doe.gov/Ricme1/Ricme1.home.html
Ricmel1	*Rickenella mellea*	https://mycocosm.jgi.doe.gov/Ricmel1/Ricmel1.home.html
Rozal1_1	*Rozella allomycis*	https://mycocosm.jgi.doe.gov/Rozal1_1/Rozal1_1.home.html
Rutfi1	*Rutstroemia firma*	https://mycocosm.jgi.doe.gov/Rutfi1/Rutfi1.home.html
Sacce1	*Saccharomyces cerevisiae S288C*	https://mycocosm.jgi.doe.gov/Sacce1/Sacce1.home.html
SacceM3707_1	*Saccharomyces cerevisiae M3707*	https://mycocosm.jgi.doe.gov/SacceM3707_1/SacceM3707_1.home.html
SacceM3836_1	*Saccharomyces cerevisiae M3836*	https://mycocosm.jgi.doe.gov/SacceM3836_1/SacceM3836_1.home.html
SacceM3837_1	*Saccharomyces cerevisiae M3837*	https://mycocosm.jgi.doe.gov/SacceM3837_1/SacceM3837_1.home.html
SacceM3838_1	*Saccharomyces cerevisiae M3838*	https://mycocosm.jgi.doe.gov/SacceM3838_1/SacceM3838_1.home.html
SacceM3839_1	*Saccharomyces cerevisiae M3839*	https://mycocosm.jgi.doe.gov/SacceM3839_1/

		SacceM3839_1.home.html
SacceYB210_1	*Saccharomyces cerevisiae YB210*	https://mycocosm.jgi.doe.gov/SacceYB210_1/SacceYB210_1.home.html
Sacpr1	*Saccharata proteae*	https://mycocosm.jgi.doe.gov/Sacpr1/Sacpr1.home.html
Saico1	*Saitoella complicata*	https://mycocosm.jgi.doe.gov/Saico1/Saico1.home.html
Sarco1	*Sarcoscypha coccinea*	https://mycocosm.jgi.doe.gov/Sarco1/Sarco1.home.html
Schco_LoeD_1	*Schizophyllum commune Loenen*	https://mycocosm.jgi.doe.gov/Schco_LoeD_1/Schco_LoeD_1.home.html
Schco_TatD_1	*Schizophyllum commune Tattone*	https://mycocosm.jgi.doe.gov/Schco_TatD_1/Schco_TatD_1.home.html
Schco3	*Schizophyllum commune*	https://mycocosm.jgi.doe.gov/Schco3/Schco3.home.html
Schcy1	*Schizosaccharomyces cryophilus*	https://mycocosm.jgi.doe.gov/Schcy1/Schcy1.home.html
Schja1	*Schizosaccharomyces japonicus*	https://mycocosm.jgi.doe.gov/Schja1/Schja1.home.html
Schoc1	*Schizosaccharomyces octosporus*	https://mycocosm.jgi.doe.gov/Schoc1/Schoc1.home.html
Schpa1	*Schizopora paradoxa*	https://mycocosm.jgi.doe.gov/Schpa1/Schpa1.home.html
Schpo1	*Schizosaccharomyces pombe*	https://mycocosm.jgi.doe.gov/Schpo1/Schpo1.home.html
Sclci1	*Scleroderma citrinum*	https://mycocosm.jgi.doe.gov/Sclci1/Sclci1.home.html
Sclhys1_1	*Sclerogaster hysterangioides*	https://mycocosm.jgi.doe.gov/Sclhys1_1/Sclhys1_1.home.html
Sclsc1	*Sclerotinia sclerotiorum*	https://mycocosm.jgi.doe.gov/Sclsc1/Sclsc1.home.html
Scysp1_1	*Scytinostroma sp.*	https://mycocosm.jgi.doe.gov/Scysp1_1/Scysp1_1.home.html
Sebve1	*Sebacina vermifera*	https://mycocosm.jgi.doe.gov/Sebve1/Sebve1.home.html
Sebvebe1	*Sebacina vermifera ssp bescii*	https://mycocosm.jgi.doe.gov/Sebvebe1/Sebvebe1.home.html
Sepmu1	*Septoria musiva*	https://mycocosm.jgi.doe.gov/Sepmu1/Sepmu1.home.html
Seppo1	*Septoria populicola*	https://mycocosm.jgi.doe.gov/Seppo1/Seppo1.home.html
Sepsp1	*Septobasidium sp.*	https://mycocosm.jgi.doe.gov/Sepsp1/Sepsp1.home.html
Serla_varsha1	*Serpula lacrymans var shastensis*	https://mycocosm.jgi.doe.gov/Serla_varsha1/Serla_varsha1.home.html
SerlaS7_9_2	*Serpula lacrymans*	https://mycocosm.jgi.doe.gov/SerlaS7_9_2/SerlaS7_9_2.home.html

Setho1	*Setomelanomma holmii*	https://mycocosm.jgi.doe.gov/Setho1/Setho1.home.html
Settu1	*Setosphaeria turcica Et28A*	https://mycocosm.jgi.doe.gov/Settu1/Settu1.home.html
Settur1	*Setosphaeria turcica NY001*	https://mycocosm.jgi.doe.gov/Settur1/Settur1.home.html
Sirint1	*Sirobasidium intermedium*	https://mycocosm.jgi.doe.gov/Sirint1/Sirint1.home.html
Sisbr1	*Lentinus tigrinus*	https://mycocosm.jgi.doe.gov/Sisbr1/Sisbr1.home.html
Sisni1	*Sistotremastrum niveocremeum*	https://mycocosm.jgi.doe.gov/Sisni1/Sisni1.home.html
Sissu1	*Sistotremastrum suecicum*	https://mycocosm.jgi.doe.gov/Sissu1/Sissu1.home.html
Sodal1	*Sodiomyces alkalinus*	https://mycocosm.jgi.doe.gov/Sodal1/Sodal1.home.html
Spafl1	*Spathularia flavida*	https://mycocosm.jgi.doe.gov/Spafl1/Spafl1.home.html
Sphst1	*Sphaerobolus stellatus*	https://mycocosm.jgi.doe.gov/Sphst1/Sphst1.home.html
Spipu1	*Spizellomyces punctatus*	https://mycocosm.jgi.doe.gov/Spipu1/Spipu1.home.html
Spofi1	*Sporormia fimetaria*	https://mycocosm.jgi.doe.gov/Spofi1/Spofi1.home.html
Spola1	*Sporopachydermia lactativora*	https://mycocosm.jgi.doe.gov/Spola1/Spola1.home.html
Spoli1	*Sporobolomyces linderae*	https://mycocosm.jgi.doe.gov/Spoli1/Spoli1.home.html
Spopa1	*Sporidiobolus pararoseus Phaff*	https://mycocosm.jgi.doe.gov/Spopa1/Spopa1.home.html
Spore1	*Sporisorium reilianum*	https://mycocosm.jgi.doe.gov/Spore1/Spore1.home.html
Spoth2	*Myceliophthora thermophila*	https://mycocosm.jgi.doe.gov/Spoth2/Spoth2.home.html
Stagr1	*Stanjemonium grisellum*	https://mycocosm.jgi.doe.gov/Stagr1/Stagr1.home.html
Stano2	*Stagonospora nodorum*	https://mycocosm.jgi.doe.gov/Stano2/Stano2.home.html
Stasp1	*Stagonospora sp.*	https://mycocosm.jgi.doe.gov/Stasp1/Stasp1.home.html
Stehi1	*Stereum hirsutum*	https://mycocosm.jgi.doe.gov/Stehi1/Stehi1.home.html
Sugame1	*Sugiyamaella americana*	https://mycocosm.jgi.doe.gov/Sugame1/Sugame1.home.html
Suiame1	*Suillus americanus*	https://mycocosm.jgi.doe.gov/Suiame1/Suiame1.home.html
Suibr2	*Suillus brevipes*	https://mycocosm.jgi.doe.gov/Suibr2/Suibr2.home.html
Suidec1	*Suillus decipiens*	https://mycocosm.jgi.doe.gov/Suidec1/Suidec1.home.html

Suigr1	*Suillus granulatus*	https://mycocosm.jgi.doe.gov/Suigr1/Suigr1.home.html
Suihi1	*Suillus hirtellus*	https://mycocosm.jgi.doe.gov/Suihi1/Suihi1.home.html
Suilu3	*Suillus luteus*	https://mycocosm.jgi.doe.gov/Suilu3/Suilu3.home.html
Suipic1	*Suillus pictus*	https://mycocosm.jgi.doe.gov/Suipic1/Suipic1.home.html
Symat1	*Sympodiomyces attinorum*	https://mycocosm.jgi.doe.gov/Symat1/Symat1.home.html
Symko1	*Symbiotaphrina kochii*	https://mycocosm.jgi.doe.gov/Symko1/Symko1.home.html
Synrac1	*Syncephalastrum racemosum*	https://mycocosm.jgi.doe.gov/Synrac1/Synrac1.home.html
Talma1_2	*Talaromyces marneffei*	https://mycocosm.jgi.doe.gov/Talma1_2/Talma1_2.home.html
Talst1_2	*Talaromyces stipitatus*	https://mycocosm.jgi.doe.gov/Talst1_2/Talst1_2.home.html
Tapde1_1	*Taphrina deformans*	https://mycocosm.jgi.doe.gov/Tapde1_1/Tapde1_1.home.html
Terbo2	*Terfezia boudieri*	https://mycocosm.jgi.doe.gov/Terbo2/Terbo2.home.html
Ternu1	*Teratosphaeria nubilosa*	https://mycocosm.jgi.doe.gov/Ternu1/Ternu1.home.html
Tescy1	*Testicularia cyperi*	https://mycocosm.jgi.doe.gov/Tescy1/Tescy1.home.html
Thaele1	*Thamnidium elegans*	https://mycocosm.jgi.doe.gov/Thaele1/Thaele1.home.html
Theau1	*Thermoascus aurantiacus*	https://mycocosm.jgi.doe.gov/Theau1/Theau1.home.html
Thega1	*Thelephora ganbajun*	https://mycocosm.jgi.doe.gov/Thega1/Thega1.home.html
Themi1	*Thelebolus microsporus*	https://mycocosm.jgi.doe.gov/Themi1/Themi1.home.html
Thest1	*Thelebolus stercoreus*	https://mycocosm.jgi.doe.gov/Thest1/Thest1.home.html
Thian1	*Thielavia antarctica*	https://mycocosm.jgi.doe.gov/Thian1/Thian1.home.html
Thiap1	*Thielavia appendiculata*	https://mycocosm.jgi.doe.gov/Thiap1/Thiap1.home.html
Thiar1	*Thielavia arenaria*	https://mycocosm.jgi.doe.gov/Thiar1/Thiar1.home.html
Thihy1	*Thielavia hyrcaniae*	https://mycocosm.jgi.doe.gov/Thihy1/Thihy1.home.html
Thite2	*Thielavia terrestris*	https://mycocosm.jgi.doe.gov/Thite2/Thite2.home.html
ThoPMI491_1	*Thozetella sp.*	https://mycocosm.jgi.doe.gov/ThoPMI491_1/ThoPMI491_1.home.html

Tilalb1	*Tilletiopsis albescens*	https://mycocosm.jgi.doe.gov/Tilalb1/Tilalb1.home.html
Tilan2	*Tilletiaria anomala*	https://mycocosm.jgi.doe.gov/Tilan2/Tilan2.home.html
Tilwa1	*Tilletiopsis washingtonensis*	https://mycocosm.jgi.doe.gov/Tilwa1/Tilwa1.home.html
Torde1	*Torulaspora delbrueckii*	https://mycocosm.jgi.doe.gov/Torde1/Torde1.home.html
Torra1	*Torpedospora radiata*	https://mycocosm.jgi.doe.gov/Torra1/Torra1.home.html
Totfu1	*Tothia fuscella*	https://mycocosm.jgi.doe.gov/Totfu1/Totfu1.home.html
Trace1	*Trametopsis cervina*	https://mycocosm.jgi.doe.gov/Trace1/Trace1.home.html
Traci1	*Trametes cingulata*	https://mycocosm.jgi.doe.gov/Traci1/Traci1.home.html
Tralj1	*Trametes ljubarskyi*	https://mycocosm.jgi.doe.gov/Tralj1/Tralj1.home.html
Trave1	*Trametes versicolor*	https://mycocosm.jgi.doe.gov/Trave1/Trave1.home.html
Treen1	*Tremella encephala*	https://mycocosm.jgi.doe.gov/Treen1/Treen1.home.html
Trepe1	*Trematosphaeria pertusa*	https://mycocosm.jgi.doe.gov/Trepe1/Trepe1.home.html
Triab1_1	*Trichaptum abietinum*	https://mycocosm.jgi.doe.gov/Triab1_1/Triab1_1.home.html
Trias1	*Trichoderma asperellum*	https://mycocosm.jgi.doe.gov/Trias1/Trias1.home.html
Triasp1	*Trichoderma asperellum TR356*	https://mycocosm.jgi.doe.gov/Triasp1/Triasp1.home.html
Triat2	*Trichoderma atroviride*	https://mycocosm.jgi.doe.gov/Triat2/Triat2.home.html
Tribi1	*Trichodelitschia bisporula*	https://mycocosm.jgi.doe.gov/Tribi1/Tribi1.home.html
Trich1	*Trichosporon chiarellii*	https://mycocosm.jgi.doe.gov/Trich1/Trich1.home.html
Trici4	*Trichoderma citrinoviride*	https://mycocosm.jgi.doe.gov/Trici4/Trici4.home.html
Trigu1	*Trinosporium guianense*	https://mycocosm.jgi.doe.gov/Trigu1/Trigu1.home.html
Trigue1	*Trichosporon guehoae*	https://mycocosm.jgi.doe.gov/Trigue1/Trigue1.home.html
Triha1	*Trichoderma harzianum*	https://mycocosm.jgi.doe.gov/Triha1/Triha1.home.html
Trilo3	*Trichoderma longibrachiatum*	https://mycocosm.jgi.doe.gov/Trilo3/Trilo3.home.html
Trima3	*Tricholoma matsutake*	https://mycocosm.jgi.doe.gov/Trima3/Trima3.home.html
Triol1	*Trichosporon oleaginosus*	https://mycocosm.jgi.doe.gov/Triol1/Triol1.home.html

Tripe1	*Trichomonascus petasosporus*	https://mycocosm.jgi.doe.gov/Tripe1/Tripe1.home.html
TrireRUTC30_1	*Trichoderma reesei RUT*	https://mycocosm.jgi.doe.gov/TrireRUTC30_1/TrireRUTC30_1.home.html
Triru1	*Trichophyton rubrum*	https://mycocosm.jgi.doe.gov/Triru1/Triru1.home.html
Trisp1	*Tritirachium sp.*	https://mycocosm.jgi.doe.gov/Trisp1/Trisp1.home.html
Triver1	*Trichophyton verrucosum*	https://mycocosm.jgi.doe.gov/Triver1/Triver1.home.html
TriviGv29_8_2	*Trichoderma virens*	https://mycocosm.jgi.doe.gov/TriviGv29_8_2/TriviGv29_8_2.home.html
Tryvi1	*Trypethelium eluteriae*	https://mycocosm.jgi.doe.gov/Tryvi1/Tryvi1.home.html
Tubbor1	*Tuber borchii*	https://mycocosm.jgi.doe.gov/Tubbor1/Tubbor1.home.html
Tubme1	*Tuber melanosporum*	https://mycocosm.jgi.doe.gov/Tubme1/Tubme1.home.html
Tulca1	*Tulasnella calospora*	https://mycocosm.jgi.doe.gov/Tulca1/Tulca1.home.html
Umbra1	*Umbelopsis ramanniana*	https://mycocosm.jgi.doe.gov/Umbra1/Umbra1.home.html
Uncre1	*Uncinocarpus reesii*	https://mycocosm.jgi.doe.gov/Uncre1/Uncre1.home.html
Usnflo1	*Usnea florida*	https://mycocosm.jgi.doe.gov/Usnflo1/Usnflo1.home.html
Ustma1	*Ustilago maydis*	https://mycocosm.jgi.doe.gov/Ustma1/Ustma1.home.html
Ustsp1	*Ustilaginomycotina sp.*	https://mycocosm.jgi.doe.gov/Ustsp1/Ustsp1.home.html
Valla1	*Valetoniellopsis laxa*	https://mycocosm.jgi.doe.gov/Valla1/Valla1.home.html
Varmin1	*Vararia minispora*	https://mycocosm.jgi.doe.gov/Varmin1/Varmin1.home.html
Venin1	*Venturia inaequalis*	https://mycocosm.jgi.doe.gov/Venin1/Venin1.home.html
Venpi1	*Venturia pirina*	https://mycocosm.jgi.doe.gov/Venpi1/Venpi1.home.html
Veral1	*Verticillium alfalfae*	https://mycocosm.jgi.doe.gov/Veral1/Veral1.home.html
Verda1	*Verticillium dahliae*	https://mycocosm.jgi.doe.gov/Verda1/Verda1.home.html
Veren1	*Verruculina enalia*	https://mycocosm.jgi.doe.gov/Veren1/Veren1.home.html
Verga1	*Verruconis gallopava*	https://mycocosm.jgi.doe.gov/Verga1/Verga1.home.html
Volvo1	*Volvariella volvacea*	https://mycocosm.jgi.doe.gov/Volvo1/Volvo1.home.html

Walic1	*Wallemia ichthyophaga*	https://mycocosm.jgi.doe.gov/Walic1/Walic1.home.html
Walse1	*Wallemia sebi*	https://mycocosm.jgi.doe.gov/Walse1/Walse1.home.html
Wesor1	*Westerdykella ornata*	https://mycocosm.jgi.doe.gov/Wesor1/Wesor1.home.html
Wican1	*Wickerhamomyces anomalus*	https://mycocosm.jgi.doe.gov/Wican1/Wican1.home.html
Wicdo1	*Wickerhamiella domercqiae*	https://mycocosm.jgi.doe.gov/Wicdo1/Wicdo1.home.html
Wilmi1	*Wilcoxina mikolae*	https://mycocosm.jgi.doe.gov/Wilmi1/Wilmi1.home.html
Wolco1	*Wolfiporia cocos*	https://mycocosm.jgi.doe.gov/Wolco1/Wolco1.home.html
Xanpa2	*Xanthoria parietina*	https://mycocosm.jgi.doe.gov/Xanpa2/Xanpa2.home.html
Xenvag1	*Xenasmatella vaga*	https://mycocosm.jgi.doe.gov/Xenvag1/Xenvag1.home.html
Xerba1	*Xerocomus badius*	https://mycocosm.jgi.doe.gov/Xerba1/Xerba1.home.html
Xylhe1	*Xylona heveae*	https://mycocosm.jgi.doe.gov/Xylhe1/Xylhe1.home.html
Xylhyp1	*Xylaria hypoxylon*	https://mycocosm.jgi.doe.gov/Xylhyp1/Xylhyp1.home.html
Yarli1	*Yarrowia lipolytica*	https://mycocosm.jgi.doe.gov/Yarli1/Yarli1.home.html
Zasce1	*Zasmidium cellare*	https://mycocosm.jgi.doe.gov/Zasce1/Zasce1.home.html
Zoprh1	*Zopfia rhizophila*	https://mycocosm.jgi.doe.gov/Zoprh1/Zoprh1.home.html
Zycmex1	*Zychaea mexicana*	https://mycocosm.jgi.doe.gov/Zycmex1/Zycmex1.home.html
Zyghe1_2	*Zygoascus hellenicus*	https://mycocosm.jgi.doe.gov/Zyghe1_2/Zyghe1_2.home.html
Zyghet1	*Mucor heterogamus*	https://mycocosm.jgi.doe.gov/Zyghet1/Zyghet1.home.html
Zygro1	*Zygosaccharomyces rouxii*	https://mycocosm.jgi.doe.gov/Zygro1/Zygro1.home.html
Zymar1	*Zymoseptoria ardabiliae*	https://mycocosm.jgi.doe.gov/Zymar1/Zymar1.home.html
Zymps1	*Zymoseptoria pseudotritici*	https://mycocosm.jgi.doe.gov/Zymps1/Zymps1.home.html