CENTRO DE INVESTIGACION Y DE ESTUDIOS AVANZADOS DEL
INSTITUTO POLITECNICO NACIONAL

UNIDAD IRAPUATO

"Explorando los mecanismos moleculares que mantienen la simbiosis *Rhizopus microsporus – Burkholderia rhizoxinica"*

Presenta

LCG José Roberto Bermúdez Barrientos

Para obtener el grado de

Maestro en Ciencias en Biología Integrativa

Directores de tesis:

Dra. Laila Pamela Partida Martínez
Dr. Cei Leander Gastón Abreu Goodger

Irapuato, Guanajuato Febrero de 2016

CENTRO DE INVESTIGACION Y DE ESTUDIOS AVANZADOS DEL
INSTITUTO POLITECNICO NACIONAL

UNIDAD IRAPUATO

"Exploring the molecular mechanisms maintaining the *Rhizopus microsporus* - *Burkholderia rhizoxinica* symbiosis"

Presents
LCG José Roberto Bermúdez Barrientos

To obtain the grade of
Master in Sciences in Integrative Biology

Thesis directors:

Dr. Laila Pamela Partida Martínez
Dr. Cei Leander Gastón Abreu Goodger

Irapuato, Guanajuato February de 2016

Agracecimientos

Este trabajo se realizó en el Centro de Investigación y de Estudios Avanzados del IPN - Unidad Irapuato.

Bajo la dirección de Dra. Laila Pamela Partida Martínez y de Dr. Cei Abreu Goodger

En el Laboratorio de Interacciones Microbianas perteneciente al Departamento de Ingeniería Genética y en el Laboratorio de Genómica computacional del RNA perteneciente a la Unidad de Genómica Avanzada.

En colaboración con la Dra Teresa Pawlowska, Plant Pathology & Plant-Microbe Biology Lab, Plant Science Department, Cornell University, Ithaca, NY.

Gracias al apoyo del Consejo Nacional de Ciencia y Tecnología (CONACyT) a través de los Proyectos CB-2009-01-133388, CB-2010-01-151007, INFR-2012-01-187799 y la beca 358014

Este trabajo se realizó del 1 de Septiembre de 2013 al 1 de Octubre de 2015

Miembros del comité de asesores:
Dr. Alfredo Herrera Estrella, Unidad de Genómica Avanzada, CINVESTAV Irapuato
Dr. Luis Delaye Arredondo, Departamento de Ingeniería Genética, CINVESTAV Irapuato

Acknowledgements

This project was realized at Centro de Investigación y de Estudios Avanzados del IPN – Unidad Irapuato

Under the direction of Dr. Laila Pamera Partida Martínez and Dr. Cei Abreu Goodger

At the Laboratory of Microbial Interactions belonging to Departamento de Ingeniería Genética and at the RNA Computational Genomics Laboratory belonging to the Unidad de Genómica Avanzada

In colaboration with Dr. Teresa Pawlowska, Plant Pathology & Plant-Microbe Biology Lab, Plant Science Department, Cornell University, Ithaca, NY.

Thanks to the support given by Consejo Nacional de Ciencia y Tecnología (CONACyT) through projects CB-2009-01-133388, CB-2010-01-151007, INFR-2012-01-187799 and the scholarship 358014

This project was realized from September 1st of 2013 to October 1st of 2016

Committee members:
Dr. Alfredo Herrera Estrella, Unidad de Genómica Avanzada, CINVESTAV Irapuato
Dr. Luis Delaye Arredondo, Departamento de Ingeniería Genética, CINVESTAV Irapuato

Thanks:
To my advisors, Dr. Laila Partida and Dr. Cei Abreu: for their guidance and being crucial players in my formation

To my committee members: Dr. Alfredo Herrera Estrella and Dr. Luis Delaye Arredondo for their valuable comments, discussions and suggestions

To Stephen Jay Mondo and Teresa E. Pawlowska for sharing their RNA-Seq data with us, and being such great hosts to me in Ithaca

To my family for their unconditional love and support

To my all friends for the moments we shared in this adventure

To my lab partners for their support

Contents

LIST OF ABREVIATIONS 7

ABSTRACT 8

RESUMEN 9

CHAPTER I: INTRODUCTION & BACKGROUND 11

- Endosymbiosis: concepts & historical background 11
- Mycoviruses 14
  - *The Narnaviridae family of Mycovirus* 16
- Eukaryotic chromatin manipulation by bacteria 17
  - *Chromatin* 17
  - *Chromatin structure and regulation* 18
  - *Histone modification language* 19
  - *Histone variants* 21
  - *Host chromatin manipulation by pathogens and symbionts* 21
- Fungal immunity and symbiont perception 23
  - *Leucine rich repeats* 23
  - *Lysin motif* 24
- Toxin-antitoxin systems 24
- Non-Ribosomal Peptide Synthetases 26
- The *Rhizopus microsporus* – *Burkholderia rhizoxinica* symbiosis 29

OBJECTIVE 33

- Specific objectives 33

JUSTIFICATION 34

- Differential expression analysis 34
- Comparative genomics 34

CHAPTER II: METHODS 35

- Annotations 36
  - *Pfam* 36
  - *GO* 36
  - *RAST* 37
  - *AntiSMASH* 37
- RNA-Seq analysis 38
  - *Experimental design, extraction & sequencing* 38
  - *Quality control* 40
  - *Mapping* 40
  - *Gene counts* 41
  - *Differential Expression Analysis* 41
  - *B. rhizoxinica expression inside its host* 42
  - *Tracing the origin of a high-GC peak* 42
- Fungal comparative genomics 44
  - *Defining protein families in Mucorales* 46
  - *Mucorales phylogeny reconstruction* 46
  - *Comparison of protein domains in fungi* 47
- Bacterial comparative genomics 49
  - *Defining protein families in Burkholderia* 51
  - *Burkholderia core genome & phylogeny* 51
  - *Functional categories comparison* 52
  - *Search for Burkholderia toxin-antitoxin proteins in Mucorales* 54

CHAPTER III: RNA-SEQ ANALYSIS RESULTS 55
NARNAVIRUSES IN *RHIZOPUS* 55
A strange peak of reads with high-GC content 55
Tracing peak origin and discovery of Narnavirus sequences 59
GENERAL LIBRARIES DESCRIPTION 60
HOUSEKEEPING GENES PROPOSAL OF *RHIZOPUS MICROSPORUS* AND *BURKHOLDERIA RHIZOXINICA* 61
DIFFERENTIAL EXPRESSION ANALYSIS 66
Nuclear DE proteins: TF, Histones & SET 73

CHAPTER IV: RESULTS OF FUNGAL COMPARATIVE GENOMICS & EXPRESSION. 81
MUCORALES PHYLOGENY 81
RHIZOPUS GENOME DESCRIPTION 82
PROTEIN DOMAINS COMPARISON: *RHIZOPUS* HOST VS NON-HOST 84
Losses in host strains: PIF1 helicases 84
PROTEIN DOMAINS COMPARISON: MUCORALES VS DYKARIA 85
Reductions in Mucorales 86
Expansions in Mucorales 87

CHAPTER V: RESULTS OF *B. RHIZOXINICA* COMPARATIVE GENOMICS & EXPRESSION INSIDE HOST 91
AIMS 91
BURKHOLDERIA PHYLOGENY AND GENOMIC FEATURES 91
EXPANDED FUNCTIONS IN *B. RHIZOXINICA* VS FREE-LIVING RELATIVES 94
Non-ribosomal peptide synthetases 96
Toxin-antitoxin systems 99

CHAPTER VI: DISCUSSION 104
NARNAVIRUSES IN *RHIZOPUS MICROSPORUS* 104
Narnavirus distribution 104
Narnavirus phenotype 104
DIFFERENTIAL EXPRESSION ANALYSIS 105
FUNGAL COMPARATIVE GENOMICS DISCUSSION 107
PIF1 helicases 108
Reported differences between Mucorales and Dykaria 108
NON-RIBOSOMAL PEPTIDE SYNTHETASES IN MUCORALES AND *B. RHIZOXINICA* 109
TOXIN-ANTITOXIN SYSTEMS IN *R. MICROSPORUS* – *B. RHIZOXINICA* 109

CONCLUSIONS 111

PERSPECTIVES 113
NARNAVIRUS 113
SET PROTEINS 113
BURKHOLDERIA ENDOFUNGAL BACTERIA EVOLUTION 113
TOXIN-ANTITOXIN 114

REFERENCES 115

List of Abreviations

BrSET. *Burkholderia rhizoxinica* SETdomain containing protein
DE. Differential expression
FDR. False Discovery Rate
GO. Gene Ontology
H3K4. Histone 3, lysine 4
HMM. Hidden Markov Model
ITS. Internal transcribed spacer
logFC. Logarithm fold change
LPS. Bacterial lipopolysaccharide
LRR. Leucine rich repeats
LSMT. Large subunit methyltransferase
NRPS. Non-ribosomal peptide synthetase
PKS. Polyketide synthase
rDNA. Ribosomal RNA coding DNA
RdRp. RNA dependent RNA polymerase
RIP. Repeat-induced point mutation
RNA-Seq. RNA sequencing
SET. Su(var)3-9, Enhancer of zeste, Trithorax
+ssRNA*.* Single molecule positive strand RNA
T2SS. Type II secretion system
T3SS. Type III secretion system
T4SS. Type IV secretion system
TA. Toxin-antitoxin
TE. Transposable Elements
TEF. Translation Elongation Factor
TF. Transcription Factor

Abstract

The β-proteobacterium *Burkholderia rhizoxinica* is an intracellular symbiont of the Mucoralean fungus *Rhizopus microsporus*. The endosymbiont produces the toxin Rhizoxin, which its resistant fungal host uses to infect rice seedlings. Furthermore, *R. microsporus* is dependent of *B. rhizoxinica* to achieve sporulation. Unusual for such a close symbiotic relationship, both organisms can be grown independently, potentially modified and brought back together. There are also several *Rhizopus microsporus* strains that don't harbor bacteria and don't rely on *B. rhizoxinica* to sporulate, representing an interesting comparison set for the symbiotic system. To date, genome sequences of *Burkholderia rhizoxinica* and *R. microsporus* are available. All of this makes the *R. microsporus – B. rhizoxinica* an attractive model system to study symbiosis.

In this work we aimed to propose candidate genes relevant for the maintenance of *R. microsporus – B. rhizoxinica* symbiosis

To achieve this goal, we analyzed RNA-Seq data of two *R. microsporus* host strains growing alone or together, and with or without their corresponding endosymbionts. We found evidence for the presence of Narnaviruses in one of these strains; suggesting a third player in this symbiotic system. We found that nuclear proteins, particularly transcription factors, are over-expressed in absence of the endosymbiont. Fungal histone methyl-transferases are up-regulated when the endosymbiont is present, contrary to the majority of fungal nuclear genes. Interestingly, *B. rhizoxinica* also has a histone methyl-transferase that is expressed inside its host along with a type III secretion system (T3SS). We propose that the bacterial histone methyl-transferase is secreted via T3SS and modifies *R. microsporus* chromatin.

We compared protein domain contents of Mucoralean and Dykarial genomes. We identified previously reported expansions in Mucorales such as chitin synthases, Ras proteins and their regulators. Additionally we found expansion of Septins, B lectins and LRR in Mucorales relative to Dykaria. Interestingly, many of these expanded domains are differentially expressed in response to the presence of *B. rhizoxinica*.

With the same approach we compared protein domain contents of endofungal bacteria *B. rhizoxinica* and *Ca.* Glomeribacter gigasporarum relative to free-living *Burkholderia*. We found an enrichment of toxin-antitoxin system proteins in endofungal bacteria. Toxin-antitoxins were first described as a plasmid and integron maintenance system. We propose that these proteins could be relevant in endofungal bacterial inheritance. Additionally, we found more non-ribosomal peptide synthases (NRPS) in *B. rhizoxinica* than in any other genus member. These NRPS are expressed inside its *R. microsporus* along with the Rhizoxin cluster. This suggests a role for NRPS in *R. microsporus* - *B. rhizoxinica* symbiosis, like the cluster for Rhizoxin synthesis.

We propose candidate symbiosis genes both in the fungus and the bacterium and propose future research lines in this intriguing symbiosis system.

Resumen

La bacteria *Burkholderia rhizoxinica* (Betaproteobacteria) es un simbionte intracelular del hongo Mucoral *Rhizopus microsporus*. El endosimbionte produce la toxina Rhizoxin, la cual es usada por su hospedero resistente para infectar plántulas de arroz. Además, *R. microsporus* es dependiente de *B. rhizoxinica* para poder esporular. Inusualmente para una relación tan cercana, ambos organismos pueden crecerse independientemente, potencialmente modificarse y reestablecer la interacción en condiciones de laboratorio. Existen varias cepas de *R. microsporus* que no poseen bacterias que no dependen de *B. rhizoxinica* para esporular, representando un conjunto de comparación interesante para este sistema simbiótico. A la fecha se cuenta con secuencias genómicas de *B. rhizoxinica* y *R. microsporus*. Todo esto hace de *R. microsporus* y *B. rhizoxinica* un modelo atractivo para estudiar simbiosis.

En este trabajo nos planteamos proponer genes relevantes para la manutención de la simbiosis entre *R. microsporus* y *B. rhizoxinica*.

Para ello, analizamos datos de RNA-Seq de dos cepas hospederas de *R. microsporus* creciendo juntas o aisladas, y con o sin sus endosimbiontes correspondientes. Encontramos evidencia para la presencia de Narnaviruses en una de las cepas; lo cual sugiere un tercer miembro en este sistema de simbiosis. Encontramos que las proteínas nucleares, particularmente factores de transcripción, tienden a estar sobre-expresadas en ausencia del endosimbionte. Por otro lado, las histona metil-transferasas del hongo aumentan su expresión cuando el endosimbionte está presente, de manera contraria a la mayoría de los genes nucleares. Interesantemente, *B. rhizoxinica* también tiene una histona metil-transferasa que es expresada dentro de su hospedero junto con un sistema de secreción de tipo III (T3SS por sus siglas en inglés). Nosotros proponemos que la histona metil-transferasa bacteriana es secretada mediante el T3SS y modifica la cromatina de *R. microsporus*.

Comparamos los contenidos de dominios de proteínas de genomas Mucorales y Dykaria. Encontramos expansiones previamente reportadas en Mucorales como quitin sintasas, proteínas Ras y sus reguladores. Adicionalmente encontramos expansiones de septinas, B lectinas y repetidos ricos en leucinas. Interesantemente, muchos de estos dominios expandidos son diferencialmente expresados en respuesta a la presencia de *B. rhizoxinica*.

Con la misma metodología comparamos la distribución de dominios de proteínas en las bacterias endosimbiontes de hongos *B. rhizoxinica* y *Ca.* Glomeribacter gigasporarum con respecto a *Burkolderias* de vida libre. Encontramos un enriquecimiento de proteínas de sistemas toxina-antitoxina en estas bacterias endosimbiontes. Los sistemas toxina-antitoxina fueron descubiertos como sistemas de mantenimiento de plásmidos e integrones. Nosotros proponemos que estas proteínas son relevantes para el mantenimiento de bacterias endosimbiontes de hongos. Adicionalmente, encontramos más sintasas de péptidos no ribosomales (NRPS por sus siglas en inglés) en *B. rhizoxinica* que en cualquier otro miembro del género. Estas NRPS se expresan dentro de *R. microsporus* junto con el cluster de síntesis de Rhizoxin. Estos hallazgos sugieren un papel de las NRPS en la simbiosis *R. microsporus – B. rhizoxinica*, de manera similar al cluster de síntesis de Rhizoxin.

En este trabajo proponemos genes candidatos para la simbiosis tanto en el hongo como la bacteria y sugerimos futuros experimentos para estudiar este sorprendente sistema simbiótico.

Chapter I: Introduction & Background

Endosymbiosis: concepts & historical background

Symbiosis was defined in 1879 by Heinrich Anton de Bary as unlike species living together. This definition embraces all different kinds of species interactions, and they can be classified based on the outcome for the involved species (Figure 1). This classification includes: mutualism, in which both interacting species benefit; commensalism, in which one of the species benefits, while the other isn't helped or harmed; parasitism, in which one member benefits at the expense of the other; amensalism, in which one species has a detrimental effect on another one without benefiting or harming itself. Finally, competition is the result of both organisms being negatively affected by their interaction (Holland & Bronstein 2008).

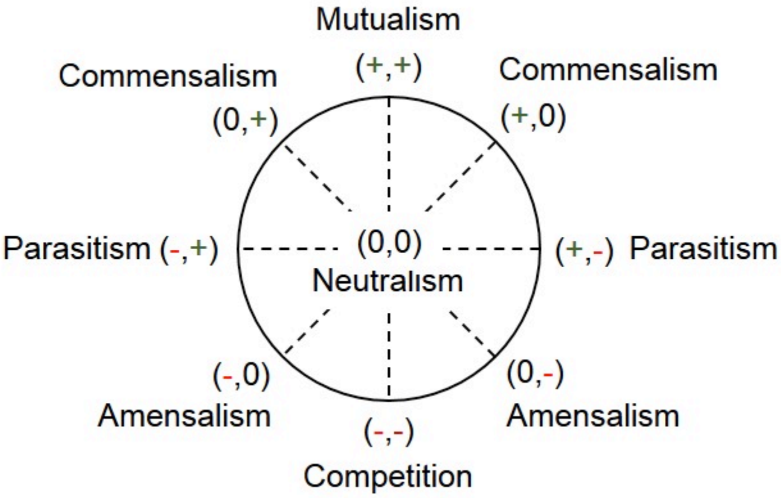


Figure 1. Classification of interspecies interactions based on the outcome of interacting populations. The center of the circle represents no interaction at all. Departing from the center increases the magnitude of the interaction with the corresponding sign. Moving along a different angle in the circumference changes the interaction outcome sign. Positive interaction (+), negative interaction (-), no interaction (0). Based on Holland and Bronstein 2008.

The association of arbuscular mycorrhiza with land plants represents an excellent example of mutualism. Approximately 80% of land plant species form associations with arbuscular mycorrhiza. The fungus provides water, phosphate and nitrogen, in exchange of up to 20% of the carbon fixed by the plant. A recent study revealed that the arbuscular mycorrhiza *Gigaspora margarita* benefits from harboring the β-proteobacteria endosymbiont *Candidatus* Glomeribacter gigasporarum. The bacteria increase the sporulation success of the fungus, in exchange for nutrients and a safe environment (Salvioli et al. 2010). A *G. margarita* line devoid of *Ca.* G. gigasporarum formed half the spores than the one harboring its endosymbiont.

Burkholderia gladioli pathovar agaricicola is an important pathogen in the mushroom industry and represents a good example of parasitism. Cavity disease results from this bacterium feeding on white button mushrooms. To be able to feed on the fungus, *B. gladioli* uses a type II secretion system (T2SS). Bacterial mutants for a T2SS component are unable to degrade *Agaricus bitorquis* tissue*,* display reduced number of flagella and have compromised protease and chitinase activities (Chowdhury & Heinemann 2006).

Listing mutualistic and parasitic interactions is way easier than finding commensalism and amensalism examples. True commensalism or amensalism symbioses are hard to prove. To achieve this, a neutral effect is required throughout the development of both organisms, under all environmental conditions. We could consider the following interaction as an example of commensalism. The bacterium *Burkholderia terrae* BS001 is able to migrate through hyphae of several fungal species. While the bacterium gets transportation, a benefit for the fungus is not always obvious (Nazir et al. 2014).

Symbiotic interactions are dynamic as they can change over time due to population density, third party influence, nutritional and environmental conditions (Holland & Bronstein 2008). There are even experiments to turn pathogens into mutualists. A remarkable study started by mutagenizing the pathogen *Ralstonia solanacearum* harboring the symbiotic plasmid of the nitrogen-fixing mutualistic *Cupriavidus taiwanensis*. Mutants able to infect Mimosa nodules were isolated after one round of selection. Legume nodule invasion and nitrogen fixation are considered characteristic traits of plant mutualistic bacteria. The mutation of two genes was found relevant for the pathogen to mutualism shift, virulence regulator *hrpG* and a type III secretion system component. Nonsense mutants for these genes presented different mutualistic capabilities; lack of the type III secretion system resulted in early infections; while the *hrpG* mutant was able to stimulate nodulation and even invade the cytoplasm of nodule cells (Marchetti et al. 2010).

Symbiotic interactions can also be classified by the degree of closeness for the species involved: co-ocurrence, ectosymbiosis and endosymbiosis. Co-ocurrent species simply share a given environment. A co-occurrence analysis of an intercontinental soil sample collection revealed that bacteria belonging to the genus *Burkholderia* frequently co-occur with *Alternaria alternata* and *Fusarium solani* fungi (Stopnisek et al. 2015). In this study, authors argue that bacteria could use fungal hyphae to be dispersed in soils devoid of water films. They found that genes involved in motility where undetected in co-cultures of *Burkholderia glathei* and *Fusarium solani*, relative to the bacterium cultured alone.

In ectosymbiosis, a closer association, symbionts live on the body surface of the host. This is the case for ectoparasitic mites living attached to *Drosophila* hosts. *Macrocheles subbadius* ectoparasitic mites feed on haemolymph of *Drosophila nigrospiracula* (Polak 1996). Endosymbiosis is the phenomenon in which one organism lives inside the body or the cells of another organism, regardless of the benefit or harm to the host (Wernegreen 2012). Endosymbiosis represents the most intimate of species associations.

Based on the codependence and evolutionary age of interaction, animal endosymbionts are categorized as “primary” or “secondary”. Primary or obligate endosymbionts are required by the host and usually reside in specialized host organs such as bacteriomes. Secondary or facultative endosymbionts are not needed for host survival and they don’t reside in specialized organs (Dale & Moran 2006).

The endosymbiosis theory for the origin of chloroplasts was originally proposed by Russian biologist Konstantin Sergejewiz Merezhkovsky in 1905. He argued that chromatophores (chloroplasts) are plant symbionts, based on observations that plastids proliferate through self-division and are not formed *de novo*. He thus speculated that the first chloroplast would have migrated into a colorless organism. He also mentioned examples of probable free-living chromatophores and even some examples of recent invasion of chromatophores, such as *Paulinella chromatophora* (Martin & Kowallik 1999). Ivan Wallin brought the endosymbiosis theory to mitochondria, proposing that this organelle is of bacterial origin. His claims were based on the successful isolation of mitochondria, followed by a comparison with free-living bacteria (Wallin, 1922). The endosymbiosis theory became popular in the first two decades of the twentieth century and was then abandoned (Martin & Kowallik 1999).

The endosymbiosis theory re-gained popularity with Lynn Margulis’s 1967 paper "On the origin of mitosing cells". In this work, Margulis proposed that mitochondria, chloroplasts and flagella basal bodies originated from ancient endosymbiosis events. She described a specific order of appearance of these cellular components, based on their phylogenetic distribution, geological and fossil record. Margulis listed general properties that an endosymbiont must have:

1. Symbionts must have had their own DNA, replication, transcription and translation machinery.
2. A mechanism must exist to ensure symbiont heritability after cell division.
3. The capabilities conferred to the host by the symbiont behave as a unit: they are not fragmented capacities.
4. If the symbiont is lost all capabilities conferred are lost, once lost these capabilities cannot be restored by nuclear genes.
5. Non-Mendelian genetics should be found in organisms in which mitochondria and chloroplast are inherited uniparentally.
6. Free-living relatives may be found among extant organisms.

In 1978 Schwartz and Dayhoff presented the first sequence-based evidence favoring the endosymbiosis theory. They used sequences corresponding to ferredoxins, 5S RNA and c-type cytochrome. With individual and composite phylogenies they were able to construct a model for the evolution of bacteria, eukaryotes, mitochondria and chloroplasts. In their model, chloroplasts cluster with free-living blue-green algae (cyanobacteria), on the other hand, mitochondria cluster with the free-living photosynthetic and aerobic bacterium *Rhodopseudomonas* (Schwartz & Dayhoff 1978). These results are in accordance with the endosymbiotic theory for the origin of mitochondria and chloroplasts. Nowadays, it is well accepted that mitochondria and chloroplasts are the result of endosymbiosis events. But the endosymbiont origin of flagella proposed by Margulis is still controversial.

The eukaryote cell is probably the best example that endosymbiosis has been a powerful source of evolutionary innovation. There's no reason to think that the processes that resulted in the formation of mitochondria and chloroplasts are not still active. In this regard, it's worth mentioning that endosymbiotic bacteria have been found throughout the eukaryote domain.

The first evidence for bacterial endosymbionts in fungi (endofungal bacteria) was published in 1970. Barbara Mosse described the development of *Endogone* spores using microscopy. The *Endogone* genus belongs to the mucoromycotina subdivision. Mosse found bacterial-like objects inside *Endogone* spores. These objects had an outer/inner membrane and ribosomes. She was even able to capture the very moment of division of these bacterial-like objects (Mosse 1970). Since then, a plethora of endofungal bacteria have been reported involving phylogenetically diverse bacteria and fungi (Hoffman & Arnold 2010; Salvioli et al. 2010).

Mycoviruses

Mycoviruses are viruses that infect fungi, and were first discovered in fungi in 1962. To be considered mycoviruses, candidate particles need to be able to infect healthy fungi, if this condition is not met, they are considered virus-like particles (Bozarth 1972).

Mycoviruses have been found in all major groups of fungi such as Ascomycota, Basidiomycota, Zygomycota and Chytridiomycota. According to Bozarth, double stranded RNA mycoviruses have been found in Mucorales such as *Rhizopus* and *Mucor*. Mucorales is the largest studied order of Zygomycota and is of particular interest to our study. Random sampling revealed that 10 to 15% of fungal species contain mycoviruses (Bozarth 1972).

Some genera are rich in mycoviruses; for example, five out of six screened *Penicillium* species where shown to harbor mycoviruses in random sampling. Other genera seem to be reluctant to mycovirus infection; none out of 10 isolates of *Verticilium* harbored mycoviruses (Bozarth 1972).

Classical detection methods for mycoviruses include the recognition of disease symptoms, physical methods such as partial purification or electron microscopy (Bozarth 1972). More recent detection techniques include PCR based methods, hybridization of probes or RNA sequencing.

Mycovirus particles cannot invade fungi by themselves. Mycelia fragments of infected *Helminthosporium victoriae* mushroom cultured with healthy mushrooms resulted in mycovirus transmission, but filtrates of the infected tissue did not (Bozarth 1972). Later, genome sequencing of mycoviruses revealed that they lack genes for ‘cell-to-cell movement’ or for external infections.

Mycoviruses reproduce by two different mechanisms:

1. Host sporulation
2. Hyphal anastomosis, plasmogamy, cytoplasmic exchange

When mycoviruses are present in fungal spores, they are able to reproduce in the newly formed colony. On the other hand, anastomosis, a process that involves hyphal fusion, whereby mycoviruses may move from infected to non-infected fungi (Nuss 2005).

A study in *Aspergillus* populations found that loss of mycoviruses is infrequent. (van Diepeningen et al. 2006). This work suggests that mycovirus infection is a stable process, although more studies are needed to generalize this observation.

The majority of mycovirus infections are asymptomatic, although advantageous and deleterious effects have been reported. Some of the negative effects in fungal host are:

- Reduced growth rate
- Lack of sporulation
- Attenuation of virulence
- Decreased germination of spores

Lysis of infected fungal cells is rare, in contrast to phages, which invade and destroy their bacterial hosts (Bozarth 1972).

Beneficial interactions include the killer phenotype in *Saccharomyces cerevisisae* and *Ustilago maydis*. Killer yeasts destroy cells of the same species with secreted toxins, while killer cells are immune (Schmitt & Breinig 2006). Two dsRNA virus molecules, named L-A and M, confer the killer yeast phenotype in *S. cerevisiae*. It is worth mentioning that the RNAi components, which are usually used to silence transposons and foreign DNA, are lost in *S. cerevisiae.* Interestingly, the number of killer yeast is diminished when the RNAi machinery is restored in *S. cerevisiae*. In these strains, endogenous L-A and M dsRNAs are processed into small interfering RNAs and then are lost in most cells (Drinnenberg et al. 2011).

Another example of a beneficial interaction between mycoviruses and fungi was reported in 2007. High thermal tolerance was discovered in a three-part system conformed by a mycovirus, the endophytic fungus *Curvularia protuberate* and the grass *Dichanthelium lanuginosum* (Márquez Luis M., Redman Regina S., Rodriguez Russell J. 2007). Heat-stress tolerance was only achieved by *D. lanuginosum* when harboring *C. protuberate,* and the endophyte fungus contained the mycovirus.

Mycovirus infected fungi also typically contain defective RNAs and satellite RNAs. These sub viral components depend on virus replication machinery for their maintenance (Hillman & Cai 2013).

Cryphonectria parasitica hypovirus 1 (CHV1) is the best-studied mycovirus. This is due to CHV1 success in biocontrol of chestnut blight and its use as a model for hypovirulence in fungi (Nuss 2005). CHV1 is transmitted via anastomosis between infected and healthy *Cryphonectria parasitica* fungi. Sometimes incompatibility reactions occur in anastomosis, defining vegetative compatibility groups (VCGs). VCGs limit CHV1 host range and presumably mycoviruses's host range in general.

The majority of mycoviruses have a double stranded RNA (dsRNA) genome and isomeric particles, but approximately 30% have positive single stranded RNA (+ssRNA) genomes. By 2011 a total of 90 mycovirus species and 10 viral families were described. Viruses of families Partitiviridae, Totiviridae and Narnaviridae are the most common representatives the 'mycovirus sphere'.

The Narnaviridae family of Mycovirus

The Narnaviridae family is composed of two genera, *Narnavirus* and *Mitovirus*. Both genus members consist of a single molecule of positive strand RNA (+ssRNA) that encodes a single protein, an RNA-dependent RNA polymerase (RdRp). Instead of coding for a capsid or envelope protein, RdRp molecules are bound to the RNA genome. Similarities and differences between *Mitovirus* and *Narnavirus* are shown in Table 1.

Table 1. Similarities and differences between Mitovirus and Narnavirus

Genus	Mitovirus	Narnavirus
Composition	+ssRNA	+ssRNA
Code for	RdRp	RdRp
Capsid	Non-enveloped	Non-enveloped
Location in host	Mitochondria	Cytoplasm
GC content	30%	60%

Mitovirus

Mitoviruses are among the most common viruses of fungi. They have been found in Ascomycete genera such as *Botrytis*, *Cryphonectria*, *Gremmeniella*, *Ophiostoma*, *Sclerotinia*, *Thielaviopsis* and *Tuber*. Mitoviruses have been found infecting basidiomycotan genera such as *Helicobasidium* and *Rhizoctonia*. Additionally, they have been found in the arbuscular mycorhizae genus *Glomus*

(Hillman & Cai 2013). Mitoviruses reside in mitochondria and are sometimes associated with mitochondrial deformations and with mitochondrial recombination between infected and non-infected fungi. It is worth mentioning that Mitoviruses have a GC content of 30%, similar to the fungal mitochondrial DNA genomes. In *Ophiostoma novo-ulmi,* mitovirus presence is associated with a disease phenotypes such as reduced growth and aberrant colony form (Brasier 1983).

Narnavirus

Narnavirus reports are far less common than those of Mitoviruses. Narnaviruses were discovered in *Saccharomyces cerevisiae* where two different varieties were found: ScNV-20S and ScNV-23S. They were named after the sedimentation coefficient of their RNA particles in sucrose gradients (Hillman & Cai 2013). Narnaviruses are found in their host cytoplasm and have a GC content around 60% (Hillman & Cai 2013).

A 2002 study surveyed industrial and natural isolates of *Saccharomyces cerevisiae* and *Saccharomyces diastaticus* in search of Narnaviruses. The authors found that 46 of 160 isolates were infected with these viruses (López et al. 2002). Narnaviruses have also been found in the oomycete *Phytophtora infestans* (Hillman & Cai 2013).

The presence of Narnaviruses has not been associated with a clear phenotype in their fungal host (Hillman & Cai 2013). Nevertheless in *Saccharomyces,* viral particles increase when these yeasts are grown in stressful conditions such as heat shock and nitrogen starvation (López et al. 2002). Some authors suggest that Narnavirus presence helps yeast cope with these stresses.

Eukaryotic chromatin manipulation by bacteria

Bacterial pathogens reproduce at the expense of Eukaryotic hosts. To achieve reproductive success bacterial pathogens have developed a wide diversity of strategies and molecular mechanisms. For example, varying antigens recognized by the immune system may attain host immune evasion. Another strategy is host manipulation, some bacteria employ secretion mechanisms to transfer proteins that mimic host structure or functions (Dean 2011). Some recent studies reveal that chromatin is a possible target to achieve host manipulation (Li et al. 2013; Alvarez-Venegas 2014).

Chromatin

Chromatin is a macromolecular complex found in eukaryotic nuclei. It is composed of DNA, protein and RNA that on a larger scale form chromosomes. The main roles of chromatin are:

1. To compact DNA in order to fit within the nucleus
2. To support DNA during mitosis

3. To protect DNA from damage
4. To control gene expression and DNA replication.

Two broad types of chromatin exist. The first type, named euchromatin is less condensed and can be transcribed. The second form, heterochromatin, is highly condensed and is generally not transcribed. The chromatin can achieve such level of compactation during mitosis than chromosomes are visible under the microscope as the DNA is dispensed between dividing cells.

Chromatin structure and regulation

Histones are primary components of the chromatin. These alkaline proteins compact the DNA into structural units called nucleosomes. There are five major or canonical histone families H2A, H2B, H3, H4 and H1. Histones H2A, H2B, H3 and H4 are core histones, while histones H1 and H5 are linker histones. Nucleosomes are composed of two H2A-H2B dimers, a H3-H4 tetramer and approximately 146 bp of DNA wrapped around the histone core. Linker histone H1 binds the nucleosome locking the DNA into place and allowing the formation of higher-order chromatin structures (Berger 2007).

There are three types of chromatin modifications that alter gene expression:

1. DNA methylation
2. Histone post-translational modification
3. Exchange of histone variants

In mammals, DNA methylation is generally associated with transcriptional repression. DNA methylation occurs globally in CG dinucleotides or CNG trinucleotides. In vertebrate genomes the CpG dinucleotide is found less often that it would be expected by chance.

In fungi, much less is known about the distribution and the regulatory role of DNA methylation. In *Neursopora crassa* and *Ascobolus immersus,* DNA methylation displays a mosaic distribution in the genome. Mosaic methylation comprises regions of heavily methylated DNA interspersed with regions with no methylation (Suzuki & Bird 2008). In *N. crassa* 1.5% of the cytosines are methylated, DNA methylation is absent in yeast species such as *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* (Capuano et al. 2014).

Heterochromatin formation in *N. crassa* occurs as follows. This fungus has a system named repeated-induced point mutation (RIP) that mutates and methylates repetitive sequences. These point mutations reduce the GC content of repetitive sequences. AT rich repetitive sequences are recognized by a histone methyl transferase that performs histone 3 lysine 9 tri-methylation (H3K9me3). Interestingly H3K9me3 modifications are recognized by heterochormatin protein 1, which recruits Dim-2, a DNA methyl transferase. With this recruitment Dim-2 is able to methylate DNA cytosines around H3K9me3 histone modifications.

Therefore, *N. crassa* DNA methylation is linked to histone H3K9me3 modifications (Du et al. 2015).

Histone modification language

Core histone structures are mostly globular except for their amino-terminal tails, which are non-structured and can be post-translationally modified. Possible histone modifications include methylation, acetylation, ubiquitination, phosphorylation and SUMOylation (Table 2, taken and modified from (Berger 2007)). Methylation and acetylation are small modifications while ubiquitination and SUMOylation involve large moieties, two-thirds of the size of a histone protein. Methylation can occur more than once with mono-, di- or tri-methylation on the same residue (typically lysine), and each modification level can have different biological outcomes. All histone modifications are reversible; acetylations are removed by histone deacetylases, Ser/Thr phosphatases remove phosphate groups, ubiquitin proteases remove ubiquitin from H2B, and two classes of lysine demethylases remove methyl groups from histone tails: the LSD1/BHC110 class and the jumonji class (Berger 2007).

Table 2. Histone modifications, residues and role in transcription (Berger 2007)

Histone modification	Histone residue	Transcriptional role
Acetylated lysine	H3 (9, 14, 18, 56), H4 (5, 8, 13, 16), H2A, H2B	Activation
Phospholylated serine/threonine	H3 (3, 10, 28), H2A, H2B	Activation
Methylated arginine	H3 (17, 23), H4 (3)	Activation
Methylated lysine	H3 (4, 36, 79) H3 (9, 27), H4 (20)	Activation Repression
Ubiquitinylated lysine	H2B (123) H2A (119)	Activation Repression
Sumoylated lysine	H2B (6/7), H2A (126)	Repression

There are two models in which histone modifications influence gene expression:

1. Chromatin packing is altered directly to open or close the DNA accessibility and by this way, regulating the access of DNA-binding proteins such as transcription factors.
2. The post-translational modifications alter the nucleosome surface and promote the recruitment of chromatin-binding proteins.

The interaction of all these players results in a very complex chromatin language and dynamics that regulates gene expression.

SET domain and histone methylation

The SET domain was named after three proteins in which it is present: **S**uppressor of variegation 3-9 [Su(var)3-9], **E**nhancer of zeste [E(z)], **T**rithorax [Trx]. This domain is commonly present in histone lysine methyltransferase proteins. SET proteins utilize the S-adenosyl-L-methionine cofactor to accomplish substrate methylations.

A summary of representative SET domain containing methyltransferases is shown in Table 3, modified from (Shilatifard 2008).

Table 3. Lysine histone methyltransferses (KMT), domain architecture, associated modifications and functions. Domain architecture was determined with Pfam,

Protein	Domain architecture	Modification	Function
Su(Var)3–9/Clr4 (KMT1)	Chromo, pre-SET, SET	H3K9	Transcriptional repression
Set1/COMPASS (KMT2)	RNA recognition motif, SET assoc, COMPASS, SET	H3K4	Activation
ySet2 (KMT3)	SET, SRI	H3K36	Elongation form of Pol II
yDot1 (KMT4)	DOT1	H3K79	Activation
spSet9 (KMT5)	SET	H4K20	DNA-damage response
EZH2 (KMT6)	SET	H3K27	Polycomb silencing
SET7/9 (KMT7)	MORN repeats, SET	H3K4	Not determined
RIZ1 (KMT8)	C2H2-type zinc finger	H3K9	Transcriptional repression

It is worth mentioning that some SET proteins only methylate non-histone substrates. As an example, a SET domain-containing protein in many plants methylates a lysine in the Rubisco large subunit, prior to large and small subunit joining (Herz et al. 2013). Other examples of non-histone substrates for SET proteins include the tumor suppressor protein p53, androgen receptor and estrogen receptor α (Herz et al. 2013).

SET domain containing proteins are present in both eukaryotes and prokaryotes. All eukaryotic genomes sequenced so far harbor proteins with the SET domain. In eukaryotes, SET proteins tend to include additional domains and to be part of complexes involved in chromatin remodeling. For example, EZH2 is part of the polycomb complex involved in chromatin compaction and consequently gene silencing. Another case is the Set1 part of the COMPASS complex that is associated with chromatin relaxation and gene expression. COMPASS performs H3K4 tri- and dimethylation on promoters and bodies of actively transcribed genes (Herz et al. 2013).

In mammals Heterochromatin protein 1 (HP1) proteins are part of H3K9 histone methyl transferase complexes such as SETDB1 and SUV39H1. These complexes are associated with H3K9 di and tri-methylation repressive marks (Herz et al. 2013). HP1 proteins are characterized by the presence of Chromo (PF00385) and Chromo shadow domains (PF01393). According to the Pfam database HP1 proteins are found in Opisthokonts, which includes, animals and fungi (http://pfam.xfam.org/family/PF01393#tabview=tab7).

Histone variants

Histone variants increase the complexity of chromatin. The presence of these variants can alter nucleosome stability. Histone variants display differential mRNA characteristics and expression timings. H3.2, H3.3 and CENP-A are histone variants of H3, while H2A.X, H2A.Z and macroH2A are histone variants of H2A (Biterge & Schneider 2014).

H2A.Z is one of the most conserved histone variants; it is present in all multicellular organisms. In most species in which is found, H2A.Z has a sequence similarity of 65% to the canonical H2A. Canonical histones lack introns and are not polyadenylated while H2A variants do have introns and are polyadenylated (Biterge & Schneider 2014). Canonical mRNAs need to be accumulated rapidly in the S phase to match DNA replication (Marzluff et al. 2008). The position of H2A.Z-containing nucleosomes around transcription start sites can affect the expression of downstream genes. H2A.Z is associated with an open chromatin conformation and transcriptional activity. The acetylation of this histone variant can define telomere heterochromatin boundaries (Dehé & Géli 2006). H2A.Z containing nucleosomes display an acidic patch on their surface, provoking subtle destabilization of interactions between H2A.Z-H2B and with the H3-H4 tetramer, and altering linker H1 binding. Deletion of H2A.Z is lethal in *Drosophila*, *Tetrahymena* and mouse (Biterge & Schneider 2014).

Host chromatin manipulation by pathogens and symbionts

Approximately 20% of the sequenced bacterial genomes to date have at least one SET protein, but their functional roles remain largely underexplored. Likely these proteins may have more than one biological function in bacteria, such as regulating bacterial growth as well as manipulating host transcription machinery (Alvarez-Venegas 2014; Escoll et al. 2015).

Characterized bacterial SET-containing proteins include *Chlamydia trachomatis*, *Chlamydia pneumonia*, *Legionella pneumophila*, *Bacillus anthracis*, *Burkholderia thailandensis* and the Archaea *Methanosarcina mazei*. In *Legionella pneumophila* and *Burkholderia thailandensis,* SET proteins interact with host HP1α and HP1γ proteins. This interaction promotes host rDNA expression. In vitro, *B. thailandensis* SET mono and di-methylates H3K4, a histone mark associated with transcriptional activation. The authors proposed that host cells with enforced higher rDNA transcription could provide a better niche for bacterial replication, or infecting bacteria could use host ribosomal activity for its own adventage (Li et al. 2013).

Fungal immunity and symbiont perception

The immune system is a set of many biological processes and structures within an organism that confer protection against disease. To work properly, the immune system must be able to distinguish pathogens from healthy self-tissue.

In vertebrate species the immune system is divided into adaptive immune system and innate immune system. The adaptive immune system consists of highly specialized and systemic cells that eliminate pathogens. Adaptive immunity is capable of recognizing new pathogen molecular patterns (antigens). Adaptive immunity also creates immunological memory, which leads to a better response after a first encounter with a given pathogen.

The innate immune system is non-specific and targets pathogens based on general microbe associated molecular patterns (MAMPs). Appropriate MAMPs have to be conserved among targeted microbes and absent in the host. Fungal MAMPs include chitin and beta glucans, which are components of the fungal cell wall. Bacterial detection relies on molecules such as peptidoglycan, lipopolysaccharides and flagellin (Newman et al. 2007). There are receptors that interact with MAMPs by using certain protein domains such as leucine rich repeats, lysine motifs, etc. These pattern recognition receptors are found on the surface of cells responsible for inflammation initiation in animals (Akira et al. 2006).

The innate immune system confers immediate protection against infections. This system is an ancient defense line; it is found in vertebrates, insects, fungi, plants and primitive multicellular organisms (Nürnberger et al. 2004).

The fungal immune system is poorly studied relative to animals and plants. Leucine rich repeats and lysine motif domain-containing proteins are present in fungal genomes (Buist et al. 2008); these proteins represent good candidates for bacterial perception.

Leucine rich repeats

Leucine rich repeats (LRR) fold into curved solenoidal structures and evolve specific bindings for several biological molecules such as proteins, lipids and carbohydrates (Bryant et al. 2010).

LRR domains are present in ~250 human proteins. Among these are TLR (Toll-like receptors) and NOD (nucleotide-binding oligomerization domain) proteins. All these proteins recognize specific pathogen components and trigger a downstream innate immune response. In these proteins LRR are used as ligand recognition domains. TLR are membrane receptors with the LRR domain facing the extracellular region and the intracellular Toll IL-1 receptor domain. On the other hand, NOD proteins are receptors found in the cytoplasm. In mammals, NOD1 binds to lipopolysaccharide while NOD2 binds to peptidoglycan (Inohara & Nunez 2003).

LRR membrane receptors are also present in plants and share some similarities with those present in animals. They both have an extracellular LRR domain and a single alpha helix transmembrane domain. However, in plants the intracellular effector is often a kinase. The plant LRR receptor domain architecture is also present in Oomycetes (Soanes & Talbot 2010).

A bioinformatic search revealed that LRR receptors are nearly absent from fungal genomes in stark contrast to animals and plants (Soanes & Talbot 2010). This same study revealed a fungal specific adenylyl cyclase that has a LRR domain. Adenylyl cyclases turn ATP to cyclic adenosine monophosphate (cAMP) a second messenger that regulates morphological transitions in fungi. The LRR harboring adenylyl cyclase has a widespread distribution in fungi. In *Candida albicans* this adenylyl cyclase senses peptidoglycans with its LRR domain and drives a morphological transition; from non-pathogen yeast to pathogenic hyphal growth (Xu et al. 2008). This discovery shows the potential influence of bacteria in fungal physiology.

Lysin motif

Lysin motif (LysM) displays a wide distribution, being present in more than 4000 proteins from bacterial and eukaryote genomes. LysM domains typically bind to *N*-acetyl glucosamine containing molecules such as bacterial peptidoglycan (Buist et al. 2008).

The majority of LysM-containing proteins are bacterial hydrolases used for cell wall remodeling. A great number of these bacterial LysM proteins have a secretion signal peptide at their amino terminus.

Plants use LysM-kinase receptor proteins to sense bacteria. These proteins have a similar architecture to plant LRR membrane receptors. They have an intracellular kinase domain, a transmembrane alpha helix, but a LysM domain replaces the LRR domain. In plants, LysM harboring proteins are relevant for sensing both pathogen and mutualistic bacteria (Gust et al. 2012). A LysM protein was found to be relevant for symbiosis between *Medicago truncatula* legume and nitrogen fixing *Mesorhizobium loti* (Kawaharada et al. 2015). Soanes and collaborators found no evidence of plant-like LysM-kinase receptors in fungi (Soanes & Talbot 2010).

Toxin-antitoxin systems

Toxin-antitoxin (TA) systems are small genetic modules consisting of two linked components, a stable toxin and a labile antitoxin that nullifies the effects of its partner. Toxin-antitoxin systems were first discovered as a plasmid inheritance safety measure in the 80's (Ogura & Hiraga 1983). As long as the TA harboring plasmid is kept by the bacterial host, both toxin and antitoxin are produced and the toxin effects are negated. However, when the plasmid is lost, the remaining toxin and antitoxin in the cytoplasm are inherited; eventually the labile antitoxin decays and the remaining stable toxins kill the bacterium. Consequently, at the population level, the plasmid prevalence is increased. This phenomenon is referred as post-segregational killing or addiction (Figure 2).

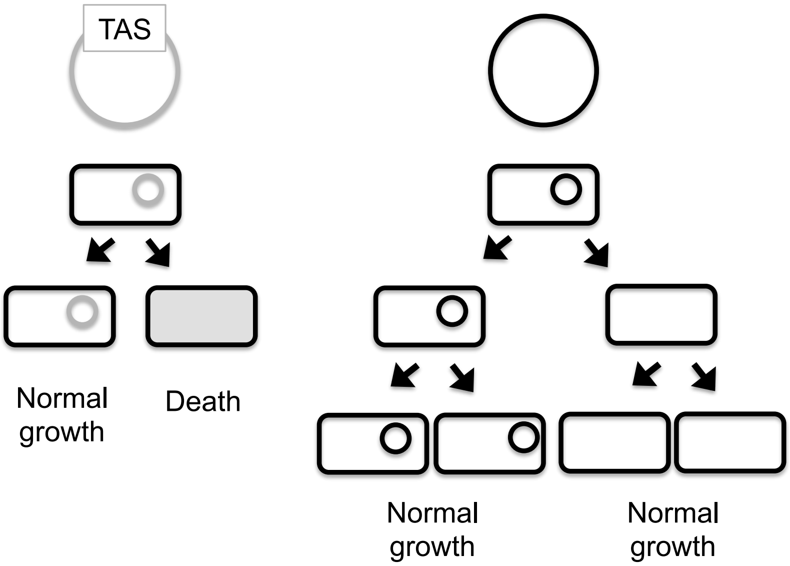


Figure 2. TA role in plasmid inheritance, TA harboring and not harboring plasmids are shown in grey and black respectively

TA are classified according to the antitoxin composition:

- Type I: Antitoxin is an antisense RNA of the protein, that prevents toxic protein from being translated
- Type II: Antitoxin is a protein
- Type III: Antitoxin is a RNA that inhibits the toxic effects of the protein

Type II TA are highly represented among bacterial genomes and tend to move through horizontal gene transfer. Type I TA show a more constrained distribution and seem to evolve by gene duplications (van Melderen 2010). Type III class is far less represented in bacterial genomes than Type II and Type I.

Type III TA was found in a cryptic plasmid of the phytopathogen bacterium *Erwinia carotovora.* This TA was named ToxIN and was proven to provide protection against bacteriophages. Antitoxin ToxI consists of 36 nt tandem repeats that are transcribed and interfere with ToxN activity. The ToxI locus in *E. carotovora* has five tandem repeats, although one is sufficient to confer toxin resistance. Finally, authors searched Type III TA in other bacterial genomes. ToxN protein sequence similarity and the identification of contiguous tandem repeats suggested that 13 other genomes have Type III TA (Fineran et al. 2009)

Mechanisms for toxin effects are shown in Table 4, information taken from (van Melderen 2010).

Table 4. Toxins, targets, activities and cellular processes affected by type 1 and 2 toxins

Toxin family	Target	Activity	Cellular process
CcdB	DNA gyrase	Generates DS breaks	Replication
RelE	Translating ribosome	Induces mRNAs cleveage	Translation
MazF	RNAs	Endoribonuclease	Translation
ParE	DNA gyrase	Generates DS breaks	Replication
Doc	Translating ribosome	Induces mRNAs cleveage	Translation
VapC	RNAs	Endoribonuclease	Translation
ζ	Unknown	Phosphotransferase	Unknown
HipA	EF-Tu	Protein kinase	Translation
HigB	Translating ribosome	Induces mRNAs cleveage	Translation
HicA	RNAs	Induces mRNAs cleveage	Translation

It was originally believed that TA were absent in obligate host associated bacteria. A study analyzing 126 completely sequenced prokaryotic genomes revealed that TA systems are absent in obligate host associated bacteria such as *Rickettsia prowazekii*, *Buchnera aphidicola*, *Wigglesworthia brevipalpis*, *Borrelia burgdorferi* and *Treponema pallidum*, etc (Pandey & Gerdes 2005). *Rickettsia felis* was the first exception to this trend; this obligate host-associated bacterium has at least 13 predicted type II TA. Later, TA proteins were found in other members of the order Rickettsiales.

The widespread distribution of type II TA and their presence in chromosomes raised questions about a functional effect to the host or if they are predominantly selfish elements. Several functions have been proposed for chromosomal encoded TA such as survival under stress conditions, guarding against DNA loss and protection against invading DNA (van Melderen 2010). TA have also been linked to pathogenesis. Elimination of VapBC homologues in *Haemphilus influenzae* results in a reduced virulence in tissue and animal models for ostitis media. In *Saphylococcus aureus,* MazF ribonuclease recognizes specific pentad sequences that are overrepresented in the mRNA of virulence genes (Wen et al. 2014). A role in biofilm formation has been proposed as some TA mutants display a reduced biofilm formation phenotype. However, a direct link between TA and biofilm formation is still controversial (Wen et al. 2014). To the best of our knowledge there are no reports of TA implicated in bacterial-fungal interactions.

Non-Ribosomal Peptide Synthetases

Non-ribosomal peptide synthetases (NRPS) are gene clusters that, in stark contrast to ribosomes, produce peptides independently of messenger RNA. They are capable of incorporating non-standard amino acids, as well as performing cyclizations and modifications of their products. Each NRPS can produce one type of peptide. Non-ribosomal peptides are often dimers, trimers or polymers of identical joined amino acids, cyclized or even branched. They are a subset of secondary metabolites or natural products, and can exhibit a broad range of biological activities such as antibiotics, immunosuppressants, siderophores, toxins, nitrogen storage polymers and phytotoxins.

The highest diversity of natural products comes from Actinomycetes, with a typical actinomycte genome containing ~20 natural product gene clusters (Traxler & Kolter 2015). NRPS are common in bacteria and fungi, but are much less represented in other higher eukaryotes. Genes devoted to production of a certain non-ribosomal peptide are usually organized in one operon in bacteria and in gene clusters in eukaryotes. NRPS are organized as modules; each module adds an amino acid residue to the growing product. Each module has several domains with specific functions; these domains are separated by approximately 15 amino acids.

There are around a dozen domains that can be found in NRPS. However, three domains are required to be present in each NRPS module:

1. Adenlytation, activates the amino acid to be incorpotared with ATP
2. Thiolation and peptide carrier protein
3. Condensation, this domain forms the amide bond

A thio-esterase domain is present once in each NRPS, this domain terminates the synthesis reaction by releasing the final product. All other modules such as formylation, cyclization, oxidation, reduction and epimerization are optional; these domains contribute to product diversity by altering its chemical composition and/or structure.

NRPS share similarities with polyketide synthases. Polyketide synthases (PKS) are multi-domain enzymes that produce polyketides, a large class of secondary metabolites. PKS products are based on malonyl blocks just like NRPS products are based on amino acids. There are even cases of hybrid NRPS-polyketide synthases; an example of this is the Rhizoxin biosynthetic cluster, present in the endofungal bacteria *B. rhizoxinica* of *Rhizopus microsporus* (Partida-Martinez & Hertweck 2007)

A comparative genomics analysis described the distribution of NRPS in fungi. NRPS are abundant in Euascomycetes, while they are scarce in Chytridiomycota, Zygomycota, Schizosaccharomycota and Hemiascomycota (Bushley & Turgeon 2010). The analyzed Zygomycota genomes included *Phycomyces blakeesleanus* and *Rhizopus oryzae*. These genomes only harbor a single NRPS that corresponds to a α-aminoadipate reductase.

The function of NRPS metabolites in the producing organism has been overlooked, mainly because the interest in natural products is biased towards health and industry applications. However, NRPS products can play a role in microbial interactions. A known NRPS based bacterial-bacterial interaction involves *Bacillus* and *Streptomyces*. Some *Bacillus* species produce surfactin, a NRPS cyclic lipopeptide product. Surfactin is able to inhibit *Streptomyces* hyphae production. However, it is not clear if *Bacillus* and Streptomyces co-ocurr in the same environments, so this interaction is considered an “off-target” effect (Traxler & Kolter 2015).

Non-ribosomal peptide synthetases can also influence bacterial-fungal interactions. The bacterium *Pseudomonas tolaasii* is the causal agent of blotch brown disease in several cultivated mushroom species. Blotch brown disease symptoms include lesions in the basidiocarp and dark brown stains. *P. tolaasii* synthesizes the toxin tolaasin, which is produced by a NRPS. Bacterial mutants for the NRPS were unable to produce tolaasin and didn’t caused the characteristic symptoms in *Agaricus* mushrooms (Scherlach et al. 2013).

The *Rhizopus microsporus – Burkholderia rhizoxinica* symbiosis

The mucoralean fungus *Rhizopus microsporus* is the causal agent of the rice seedling blight disease. Rhizoxin is a toxin necessary and sufficient to produce the disease symptoms. This toxin is a macrolide natural product that inhibits cell cycle progression and has been shown to have anti-tumor activity (Sublines et al. 1986). Rhizoxin inhibits tubulin assembly by obstructing the α- β-tubulin interface (Schmitt et al. 2008). For a long time it was believed that Rhizoxin was produced by *R. microsporus*, until in 2005, Partida-Martínez proved that the toxin was in fact produced by bacterial endosymbionts. The sequencing of endosymbiont 16S revealed that these bacteria belong to the *Burkholderia* genus and form a new species named *Burkholderia rhizoxinica* (Partida-Martinez & Hertweck 2005). The bacterium produces the toxin that confers pathogenicity to the fungus, in exchange, the fungus gives the bacterium a place to live. *R. microsporus* and potentially many other mucorales are resistant to Rhizoxin. This resistance is given by a mutation in β-tubulin. The presence of asparagine in the 100 β-tubulin position results in Rhizoxin-sensitive fungi, while the presence of other amino acids such as isoleucine, valine, serine or alanine confers resistance (Schmitt et al. 2008). This host resistance could have been important for establishment of the symbiosis.

B. rhizoxinica can be cultured in laboratory conditions (Scherlach et al. 2006), contrary to many other endosymbionts, such as the chromatophore of *Paulinella chromatophora*, *Candidatus* Glomeribacter gigasporarum (*Gigaspora margarita*), *Buchnera aphidicola* (Aphididae family), among others. As *B. rhizoxinica* can be cultured, it can be genetically manipulated and candidate symbiosis-relevant genes can be tested. Some successful examples of *R. microsorporus*-*B. rhizoxinica* symbiosis molecular mechanisms are mentioned below. The genome of *B. rhizoxinica* was sequenced and reported in 2011. A comparison with free-living relatives revealed that it is the smallest *Burkholderia* genome available, with 3.7 Mb relative to 5.8 – 8 Mb. *B. rhizoxinica* has less transcriptional regulator genes (5% of the proteome, compared to an average of 9%) and has accumulated transposons (Lackner, Moebius, Partida-Martinez, et al. 2011). *B. rhizoxinica* has a remarkable secondary metabolism potential. Besides the Rhizoxin cluster, the endosymbiont genome harbors 14 non-ribosomal peptide synthases (NRPS). A total of 9% of the genome length is composed of PKS and NRPS clusters, although NRPS distribution hasn't been analyzed in the *Burkholderia* genus.

The most common form of reproduction in *Rhizopus* is by the production of non-sexual spores in sporangia. These spores are product of mitosis and are named sporangiospores. Sexual reproduction results in the formation of a dark thick-walled zygospore, the classical defining structure in zygomycetes. A sporangium emerges from the zygospore, but resulting spore products are genetically different from either parent due to meiosis. *Rhizopus* is haploid for most of its lifecycle, the zygospore is its only diploid stage. It is also heterothallic, meaning that the sexual form of reproduction is only possible when two mycelia of opposite mating type reach each other and fuse (Figure 3, Blakeslee, 1904).

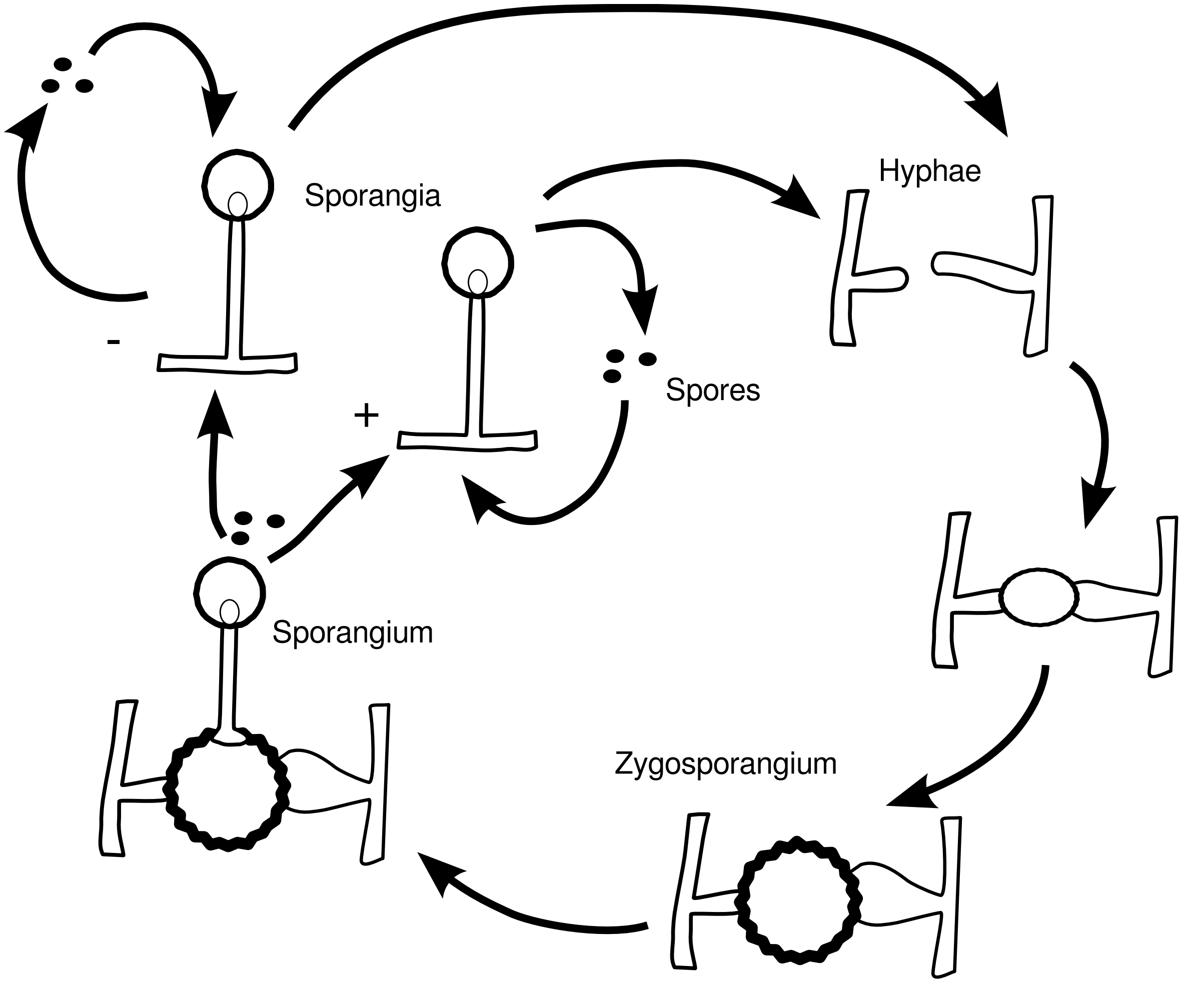


Figure 3. Rhizopus lifecycle. Plus (+) and minus (-) symbols represent different mating types

A host strain "cured" of endosymbionts can be generated treating *R. microsporus* with antibiotics (ciprofloxacin). However, the "cured" fungal strain losses its capacity to form asexual sporangiospores, which is the most common mode of reproduction for *R. microsporus* (Partida-Martinez, Monajembashi, et al. 2007). The "cured" fungal strain is still able to grow as a sterile mycelium that can be propagated via plating, albeit frequently showing a stunted behavior. Sporulation can be restored by co-culturing the "cured" host strain with the isolated endosymbiont. The same study showed that endosymbionts are vertically transmitted inside fungal spores (Partida-Martinez, Monajembashi, et al. 2007). Host-dependence and vertical transmission are hallmarks of primary endosymbiosis (Dale & Moran 2006). Interestingly, *B. rhizoxinica* is not contained in a specialized host organ (Partida-Martinez, Groth, et al. 2007), another hallmark of primary endosymbiosis in animals.

Notably, there are strains of *Rhizopus microsporus* that don't produce Rhizoxin, don't harbor *B. rhizoxinica* and don't rely on endosymbionts to complete their asexual cycle. Partida-Martínez characterized 29 *R. microsporus* strains, of which 8 harbor *B. rhizoxinica* (Partida-Martínez PhD thesis 2007), this study suggests that host strains represent a minority among the *R. microsporus* clade. The facts that less *R. microsporus* strains rely on endosymbionts and that a related species *R. oryzae* RG is also independent (Partida-Martínez PhD thesis 2007), suggest that the ancestral *Rhizopus* was independent of endosymbionts. Surprisingly, endosymbionts are interchangeable, this means that all 8 endosymbionts are able to rescue all 8 *R. microsporus* "cured" strains (Partida-Martínez PhD thesis 2007).

Dolatabadi and collaborators argue that all *R. microsporus* strains comprise a single species. They analyzed 48 *R. microsporus* strains from the CBS-KNAW Fungal Biodiversity Centre using molecular phylogenies, MALDI-ToF profiles, physiological, and mating experiments (Dolatabadi et al. 2013). Their study included six host and four true non-host *R. microsporus* strains. Interestingly, the four host strains behaved as outgroups to the main clade, which included 42 of the 48 analyzed isolates. This behavior was consistent in two out of three molecular phylogenies constructed using the rDNA internal transcribed spacer (ITS), actin and partial translation enlongation factor (TEF). The use of more molecular markers could help determining the phylogenetic distribution of the host trait. Currently, three host *Rhizopus* and three non-host *Rhizopus* genome sequences are available. This opens the opportunity for phylogenomic analysis of the *Rhizopus* genus and comparative genomics in search for the molecular basis of the host trait.

On the other hand, the possibility to culture and genetically modify *B. rhizoxinica* has already led to the discovery of some bacterial molecular mechanisms relevant in the *R. microsporus* - *B. rhizoxinica* symbiosis, which are described below.

The outer membrane lipopolysaccharyde (LPS) of Gram-negative bacteria is a key determinant for symbiotic interactions with eukaryotic hosts (Bryant et al. 2010). The O antigen of *B. rhizoxinica*'s LPS is important for the maintenance of the symbiosis (Leone et al. 2010). This was demonstrated by mutating the O antigen ligase, and comparing sporulation re-establishment to the wild type strain. The O antigen is composed of Galactofuranose, an unprecedented structure in *Burkholderia,* but common in the fungal kingdom (Morelle et al. 2005). With all this in mind, the authors proposed that *B. rhizoxinica* mimics the fungal host with its LPS (Leone et al. 2010).

Type III secretion systems (T3SS) are huge molecular complexes that span both membranes of some Gram-negative bacteria. They are typically found in bacteria capable of invading eukaryotic hosts, both in pathogenic and mutualistic relationships (Dale & Moran 2006). T3SS are used to secrete proteins (effectors) to the host and affect its behavior in a variety of ways (Dale & Moran 2006). Type II secretion systems (T2SS), also known as the general secretion pathway, are frequently used to secrete extracellular lytic enzymes and toxins (Korotkov et al.

2012). Both type III and II secretion system mutants fail to restore sporulation in *Rhizopus microsporus* (Lackner, Moebius & Hertweck 2011; Moebius et al. 2014). Also, components of both secretion systems are up-regulated in co-cultivation with the fungal host. Notably, T2SS is down-regulated once sporulation is reestablished (Moebius et al. 2014). Effector proteins for the T2SS were found via secretome proteomics. Secreted proteins include a chitosanase, a chitin-binding protein and a chitinase. The chitinase mutant was unable to restore sporulation just as a T2SS mutant. Snapshots of active invasion of *R. microsporus* hyphae were obtained via cryo-electron microscopy. It was thus concluded that the T2SS is used to enter the host in a local diffusion-like process involving the T2SS and the chitinase (Moebius et al. 2014).

In 2013 Li and collaborators reported a SET domain bacterial effector secreted via a T3SS in *Burkholderia thailandensis* (BtSET). BtSET was shown to mono and di-methylate the lysine 4 of histone H3 *in vitro*, was localized in the host nucleolus, binded to ribosomal DNA promoters and activated their transcription. Additionally BtSET contributed to intracellular replication (Li et al. 2013). It is not clear if BtSET has an impact on the expression of other host genes apart from pre-ribosomal genes. Interestingly, *Burkholderia rhizoxinica* possesses a T3SS and a SET domain containing protein (BrSET) that has been shown to methylate histones *in vitro* (Baruch, Brieba-Castro & Partida-Martínez, unpublished results). Thus, it is possible that BrSET may play a role in the bacterial-fungal symbiosis, and it may be an effector protein of the T3SS.

Stephen Mondo and Teresa Pawlowska were pioneers in the search for symbiosis relevant genes in *R. microsporus*. To achieve this goal, they compared RNA-Seq data from two *Rhizopus* host strains with and without their endosymbionts. They found a Ras2 protein that was up-regulated in the presence of *B. rhizoxinica*. Some other up-regulated genes include White collar light response regulator, CMGC/CLK, osmolarity sensing Sho1, pheromone response regulator Prr2p and Mucin. Down-regulated genes included genes involved in cell wall regulation, energy production, cytoskeleton and DNA regulation (Mondo & Pawlowska, unpublished results). Unfortunately, no statistical model for differential gene expression was used; fold change was the only criterion to define differentially expressed genes. With this criterion, lowly expressed genes are more likely to be considered differentially expressed just by chance. This makes the whole analysis susceptible to a high degree of false positives. We are currently collaborating with Teresa Pawlowska's lab and decided to revisit their analysis.

Objective

Identify fungal and bacterial candidate genes involved in the symbiosis between *Rhizopus microsporus - Burkholderia rhizoxinica*.

Specific objectives

1. Identify differentially expressed genes in *Rhizopus microsporus* in presence and in absence of *Burkholderia rhizoxinica* to propose candidate genes.
2. Identify symbiosis candidate genes in *Rhizopus microsporus* by comparing genomes of host and non-host strains.
3. Identify symbiosis candidate genes in *Burkholderia rhizoxinica* by comparing its genome with those of free-living relatives.

Justification

Differential expression analysis

The analysis made by Mondo (Mondo & Pawlowska, unpublished results) can be improved by using a statistical framework. In their analysis, differentially expressed (DE) genes were defined with a fold-change cut-off approach and a *de novo* assembly of transcript reads. We propose to use the recently sequenced genome of *R. microsporus* and a statistical model for differential expression analysis implemented in the edgeR (Robinson et al. 2009) package, within the R environment for statistical computing. Statistical models consider data variance to build an expected expression distribution. Contrary, the fold-change cut-off method sets the same threshold for all genes regardless of their expression average and variance. Lowly expressed genes are prone to have high fold-changes by chance more often than highly expressed genes. Setting a fold-change cut-off to define DE genes may require further filtering for lowly expressed genes. Additionally, we can correct the p-values according to the number of analyzed genes, to estimate the False Discovery Rate within our results.

Comparative genomics

We aim to understand the molecular basis of *R. microsporus* and *B. rhizoxinica* symbiosis. With this objective it sounds natural to compare genomes of symbionts with those genomes of non-symbiont relatives. To our knowledge nobody has made genome comparisons of host and non-host *Rhizopus* strains.

The paper of *B. rhizoxinica* genome made some comparisons with those of free-living realtives, but there is room for improvement (Lackner, Moebius, Partida-Martinez, et al. 2011). A quantitative genome scale comparison with a background statistic model could aid these analyses. Here we include more genomes that those considered in the original genome paper. Additionally we include *Candidatus* Glomeribacter gigasporarum, the fungal endosymbiont of the arbuscular mycorrhizal *Gigaspora margarita*. The inclusion of this genome adds a new layer to our comparison, the endofungal lifestyle of *Burkholderia* relatives.

Chapter II: Methods

In this chapter, I describe the methods and tools used in this project. Table 5 lists the different software tools used through this project.

Table 5. Software and versions used

Software	Version	Reference
HMMER	3.1b1	(Mistry et al. 2013)
R	3.1.3	(R Development Core Team 2008)
Bioconductor	3.0	(Gentleman et al. 2004)
Perl	v5.12.3	http://www.perl.org/
fastQC	0.10.1	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
samtools	0.1.19	(Li et al. 2009)
Trinity	v2.0.6	(Grabherr et al. 2011)
BLAST+	2.2.28	(Altschul et al. 1997)
Bowtie2	2.2.3	(Langmead & Salzberg 2012)
Muscle	v3.8.31	(Edgar 2004)
Glocks	0.91b	(Castresana 2000)
MrBayes	v3.2.5	(Ronquist et al. 2012)
MCL	12-135	(Enright et al. 2002)

The R programming language was a key tool in our analyses. R packages used are mentioned in Table 6.

Table 6. R packages and versions used in this project

Package	Version	Reference
GO.db	3.0.0	http://bioconductor.org/packages/release/data/annotation/html/GO.db.html
edgeR	3.8.6	(Robinson et al. 2009)
goseq	1.18.0	(Young et al. 2010)
RColorBrewer	1.1-2	https://cran.r-project.org/web/packages/RColorBrewer/index.html
ape	3.2	(Paradis et al. 2004)
limma	3.22.7	(Ritchie et al. 2015)
rtracklayer	1.26.3	(Lawrence et al. 2009)
seqinr	3.1-3	(Charif & Lobry 2007)

Annotations

Pfam

All analyzed genomes in this project were annotated with the Pfam database (Finn et al. 2014) using HMMER (Eddy 1998). Pfam is a database of protein families generated using profile hidden markov models (Finn et al. 2014). With profile hidden markov models (HMM) it is possible to find homologous protein domains even over long evolutionary distances.

HMMER is a software suite capable of performing a number of profile HMM operations (Eddy 1998). With HMMER it is possible to build a HMM profile from a multiple sequence alignment, create a HMM database flat file, compress and index a HMM database file, search for a given profile in a sequence database and search sequences in a HMM profile database. This last functionality is implemented by the program *hmmscan*, and it can be used together with the Pfam-A database to annotate protein sequences. Pfam-A is a curated version of the Pfam database. *hmmscan* offers a trusted cutoff threshold, indicated by the --cut_tc flag. Trusted cutoffs are set manually for each HMM profile in Pfam-A and are generally considered to be the score of the lowest-scoring known true positive that is above all known false positive protein sequences.

Example lines for annotation of proteins for organism X:

```
hmmpress Pfam-A.hmm
hmmfetch --index Pfam-A.hmm
hmmscan --tblout X.tblout --cut_tc Pfam-A.hmm X.aa.fasta
```

Previous commands assume that both Pfam database and input protein file are in the same directory. Database, input and output file directories may be added.

Protein domain identification was useful for transcriptome and comparative genomics analyses of both fungi and bacterial endosymbiont.

GO

Gene Ontology annotations (GO terms) are hierarchical descriptions with a controlled vocabulary for gene products. They have multiple parental and embedded son categories, being broader and more specific respectively. For example, nucleus is a broader term than nucleolus. There are three ontologies to describe gene products: molecular function, biological process and cellular component. Gene Ontology annotations are meant to be species independent.

GO annotations for *Rhizopus microsporus* ATCC52813 were downloaded from the Joint Genome Institute Fungi portal.
http://genome.jgi-psf.org/programs/fungi/index.jsf

GO terms are quite useful for functional enrichments analyses. To use them properly two steps are necessary. The first step is to obtain the complete hierarchies for every gene. This step is needed because usually only the most specific GO terms are given, to avoid redundancy. This is sufficient, since a gene assigned with a specific GO term (e.g. nucleolus) will necessarily also include all the more general direct parental terms (e.g. nucleus). We used the GO.db R package to reconstruct whole ontologies for each fungal gene to ensure they were complete. Ontologies are filled from the most specific term available to the broader ones for each gene. Complete GO hierarchies can be very large and redundant; this situation could result in a severe multiple testing problem. For this reason, the second step is to reduce the number of GO terms, preferably in an automated and non biased way. To perform enrichment analyses we eliminated those categories that have a low gene representation and those that are very broad.

We used an R script to reduce the number of GO terms. First it loads complete GO reconstructions for all genes. Then the script goes down for a given number of levels (we used 6 levels) starting from the root. With each descended level the script searches for children terms, discards those terms that don’t pass a minimum gene representation threshold (3 in our case) and splits categories that are larger than a maximum size (300 in our case). Additionally, if those genes present in a split category have a 80% overlap with its parental category we give priority to keep the broader category. We used this approach to diminish the multiple testing problems in a non-biased manner.

GO terms were relevant in the interpretation of *R. microsporus* differential expression analysis (See chapter III).

RAST

The RAST server offers annotations based on subsystems for bacterial and Achaea genomes (Aziz et al. 2008). Subsystems are sets of abstract functional roles manually curated by experts. RAST utilizes its own collection of protein families, named FIGfams. Proteins assigned to a functional role are then used to construct a subsystem-based protein family (FIGfam). Each FIGfam is associated with a two-tiered accurate decision procedure to determine family membership for new protein sequences (Meyer et al. 2009).

We annotated all the bacterial genomes of interest for this study, using the RAST server. Proposed functions and subsystem categories found by RAST were useful to compare *Burkholderia* genomes and interpret expression of *B. rhizoxinica* inside its host.

AntiSMASH

Antibiotics and Secondary Metabolite SHell (AntiSMASH) is well suited for genome mining of biosynthetic gene products. In bacteria, genes responsible for production of secondary metabolites tend to be grouped together in biosynthetic gene clusters.

All bacterial genomes analyzed in this study were submitted and annotated by the AntiSMASH server. It uses a hidden Markov model to predict biosynthetic gene clusters regions based on the frequencies of observed Pfam domains inside and outside a set of known biosynthetic gene clusters (Weber et al. 2015).

AntiSMASH predictions of clusters for natural product biosynthesis aided *Burkholderia* comparative genomics and analysis of expression of *B. rhizoxinica* inside its fungal host.

RNA-Seq analysis

Experimental design, extraction & sequencing

Bukholderia rhizoxinica is essential for *Rhizopus microsporus* asexual sporulation (Partida-Martinez, Monajembashi, et al. 2007). Recently, it was shown that *B. rhizoxinica* is also relevant for its host to reproduce sexually (Mondo & Pawlowska, unpublished results).

R. microsporus expression differences due to *B. rhizoxinica* may be relevant to explain the symbiont-host dependence. In order to find transcriptional differences, RNA-Seq was chosen as an explorative approach. Six different conditions were selected for the experiment (Figure 4).

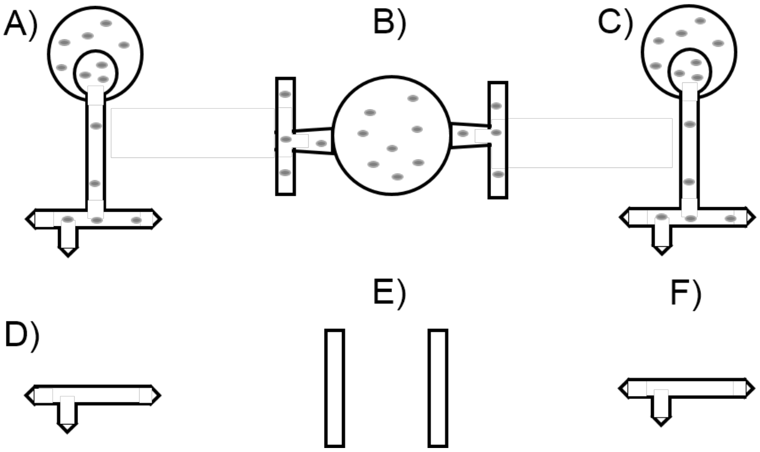


Figure 4. Experimental design and physiological state of RNA-Seq samples. A) ATCC52813 with endosymbionts, asexual sporulation via sporangia. B) Mating of ATCC52813 with ATCC52814, formation of zygospores, both strains harbor endosymbionts. C) ATCC52814 with endosymbionts, asexual sporulation via sporangia. D) ATCC52813 without its endosymbionts growing as sterile mycelium. E) Co-culture of ATCC52813 and ATCC52814 without their endosymbionts, both strains grow as sterile mycelia. F) ATCC52814 without its endosymbionts growing as sterile mycelium.

Two different host strains, *R. microsporus* var microsporus ATCC52813 and *R. microsporus* var microsporus ATCC52814, were grown alone with their cognate endosymbionts. In these conditions, *Rhizopus* strains reproduced asexually via sporangiospore formation. As a contrast setting, the same fungal strains were grown without their endosymbionts; these "cured" strains are unable of reproduce asexually and grow as sterile mycelia. Additionally, both strains were co-cultured with and without their bacterial symbionts. When co-cultured and in presence of their endosymbionts, these compatible mating type strains reproduce sexually forming zygospores. In contrast, when "cured" compatible strains are grown together they are unable to initiate the zygospore formation program and remain in the mycelium physiological state (Mondo & Pawlowska, unpublished results).

Stephen Jay Mondo performed all experimental procedures at Cornell University, NY, USA. All fungal strains were grown on half-strength potato dextrose agar (PDA) containing 2 g L^{-1} potato extract, 10 gL^{-1} dextrose, and 15 g L^{-1} agar. Strains were incubated at 30°C for six days. In this time point, opposite mate strains were undergoing reproduction in the condition containing both fungi and their endosymbionts. Each biological replicate consists of five plates that were pooled before RNA extraction. Equivalent sections of mycelia were taken. Total RNA was extracted with the ToTALLY RNA kit (Ambion ®) recovering fungal and bacterial transcripts. All conditions were treated with two Ribo-Zero rRNA Removal Kits (Epicentre, Madison, WI): Human/Mouse/Rat to remove fungal rRNA and Gram-Negative Bacteria for endosymbiont rRNA. After RNA removal, total RNA sequencing libraries were constructed and sequenced by the Cornell Sequencing Facility using the TruSeq RNA Sample Preparation Kit (Illumina). Samples were sequenced using the Illumina Hi-Seq 100 bp paired end platform.

With this data we sought to find *R. microsporus* differentially expressed genes due to presence of its endosymbiont and to explore *B. rhizoxinica* expression inside its host.

We defined short names for each library to be shown in plots (Table 7). These names will be used in the following chapters.

Table 7. RNA-Seq libraries naming conventions. b denotes bacterial endosymbiont presence. c denotes bacterial endosymbiont absence, or cured host. B4 and B7 stand for different B. rhizoxinica strains.

Library name	ATCC52813	ATCC52814	Bacterial endosymbiont
A13b	+	-	+ (B4)
A13c	+	-	-
A14b	-	+	+ (B7)
A14c	-	+	-
A13bA14b	+	+	+ (B4 & B7)
A13cA14c	+	+	-

Quality control

Sequence quality analysis was done with fastQC, including calculating the GC content distribution.

Mapping

We wanted to know how many reads corresponded to rRNA as special kits were used to remove them. We couldn't locate *R. microsporus* rRNAs in the genome using RNAmmer, an rRNA annotation tool (Lagesen et al. 2007), so, we created an artificial chromosome containing rRNAs from *R. microsporus* reported in the Silva database (Quast et al. 2013). The artificial rRNA chromosome was added to the reference genome prior to mapping the reads.

We mapped the reads directly to a mixed database of the genomes of *R. microsporus* ATCC52813 and *B. rhizoxinica* HKI 454, B1 isolate (Lackner, Moebius, Partida-Martinez, et al. 2011). By doing so, the mapping quality of a read is simultaneously evaluated against both genomes. This strategy facilitates distinguishing bacterial from fungal reads. All read mapping procedures were done with Bowtie2 (Langmead & Salzberg 2012) using paired-end reads. We used the parameters listed in Table 8.

Table 8. List and explanation of parameters used in Bowtie2

Parameter	Explanation
--local	Local alignment; ends might be soft clipped
-p 8	Number of alignment threads to launch (default 1)
-x index_file	Path and name of Bowtie2 index file
-1 paired_left_reads.fq	Path and name of paired end file 1
-2 paired_right_reads.fq	Path and name of paired end file 2
	Look for multiple alignments, report best, with MAPQ (default)

We ordered resulting alignment file with samtools sort, a necessary prior step for gene counting.

Gene counts

To count the reads mapping to each gene we used HTSeq-count tool (Anders et al. 2014). This program uses an alignment file (SAM/BAM) and a General Feature Format (GFF) file to count the reads matching gene coordinates. Resulting gene counts were used for differential expression analyses. Parameters used are shown in Table 9.

Table 9. List and explanation of parameters used in HTSeq-count for both B. rhizoxinica and R. microsporus

Parameter	Explanation
--order=name	Sorting of the alignment file. Paired-end data are sorted by name, this means that all pairs are contiguous in SAM file (choices: 'pos' or 'name')
--format=sam	Input alignment file is in SAM format (choices: 'sam' or 'bam')
--stranded=no	Data is not strand-specific (choices: 'yes', 'no' or 'reverse')
--minaqual=10	Skip all reads with alignment quality less than 10 (default: 10)
-m union	Use a conservative mode that considers a read as ambiguous when it maps to mode than one gene (choices: union, intersection-strict, intersection-nonempty; default: union) For further explanation http://www-huber.embl.de/users/anders/HTSeq/doc/count.html

Differential Expression Analysis

Differential expression (DE) analysis was done using the edgeR package in R (Robinson et al. 2009). We filtered low expression genes because differences between them have little statistical support. A gene was considered for DE analysis if at least two samples had five or more counts per million mapped reads. About 16% of the *R. microsporus* genes didn't meet this criterion and weren't considered for DE analysis.

Multiple contrasts are possible, but we focused on the following:

- Anamorphic growth: Isolated sporulating libraries with endosymbiont (Figure 4, A and C) vs isolated mycelium libraries without endosymbiont (Figure 4, D and F).
- Teleomorphic growth: Mating strains libraries (Figure 4, B) with endosymbionts vs co-cultured mycelia strains libraries without endosymbionts (Figure 4, E).

We used the method of trimmed mean of M-values (Robinson & Oshlack 2010) to calculate normalization factors to deal with different library sizes. Normalization factors were calculated with *calcNormFactors* function. Then the common, trended and tagwise dispersions were estimated using the functions *estimateCommonDisp*, *estimateTrendedDisp* and *estimateTagwiseDisp*. Differential expression was determined using the "Generalized linear model likelihood ratio" test (McCarthy et al. 2012) with the function *glmLRT*. We considered a false discovery rate (FDR) < 0.05 to define differentially expressed genes.

To perform functional enrichment analysis in DE genes we used GOseq R package (Young et al. 2010), using Pfam domains (Finn et al. 2014) and Gene Ontology annotations for functional categories.

B. rhizoxinica expression inside its host

B. rhizoxinica genes that didn't display at least five counts per million in two or more libraries were not considered as being expressed inside *R. microsporus*.

Tracing the origin of a high-GC peak

We detected an abnormal high-GC peak of reads in ATCC52814 libraries (see chapter III). To determine if these high-GC reads mapped to the fungus, bacteria or remained unmapped we used an id added to the headers of all fungal and bacterial sequences. These ids facilitated the identification of bacterial and fungal transcripts as mapping alignment files include the name of the reference sequence that a read maps to. We joined the mapped organism information with the GC content of each read to determine if the high GC peak reads mapped to the fungus or bacterium (See Chapter III). We processed SAM files with samtools (Li et al. 2009) and perl scripts to extract organism mapping information and to calculate the GC content of each read.

We designed a strategy to track the possible origin of unmapped reads (Figure 5). To achieve this, 1) we extracted unmapped reads from alignment files using samtools and custom perl scripts. 2) These unmapped reads were assembled using Trinity (Grabherr et al. 2011) performing independent assemblies for each library (Table 10).

Table 10. List of parameters used in transcriptome assembly by Trinity

Parameter	Explanation
--seqType fq	Type of reads (choices: fa, fq)
--max_memory 50G	Suggested max memory to use by Trinity where limiting can be enabled. (jellyfish, sorting, etc) provided in Gb of RAM
--left left.fq	Left reads, one or more file names (separated by commas, no spaces)
--right right.fq	Right reads, one or more file names (separated by commas, no spaces)
--CPU 8	number of CPUs to use (default 2)

3) The assembled transcripts were then compared to NCBI non-redundant database using blastx (Table 11). This database has taxonomy information for every single sequence it contains. We added taxonomy information to the transcripts in order to determine their origin.

Table 11. List of parameters used in transcripts comparison by blastx

Parameter	Explanation
-db nr	Database path and name
-num_threads 8	Number of processor to use
-query contigs.fasta	Input file in fasta format
-evalue 0.1	Threshold e-value to report alignments
-outfmt 6	Tabular output format without header
-out out_file.tab	Output file name

4) To consider expression, we mapped back orphan reads to the assembled transcripts using Bowtie2 using the same parameters as (Table 8). We then inherited the taxonomy information to all reads depending on which transcript they mapped to. Finally we calculated the GC content of each read and considered their taxonomical information (See Chapter III).

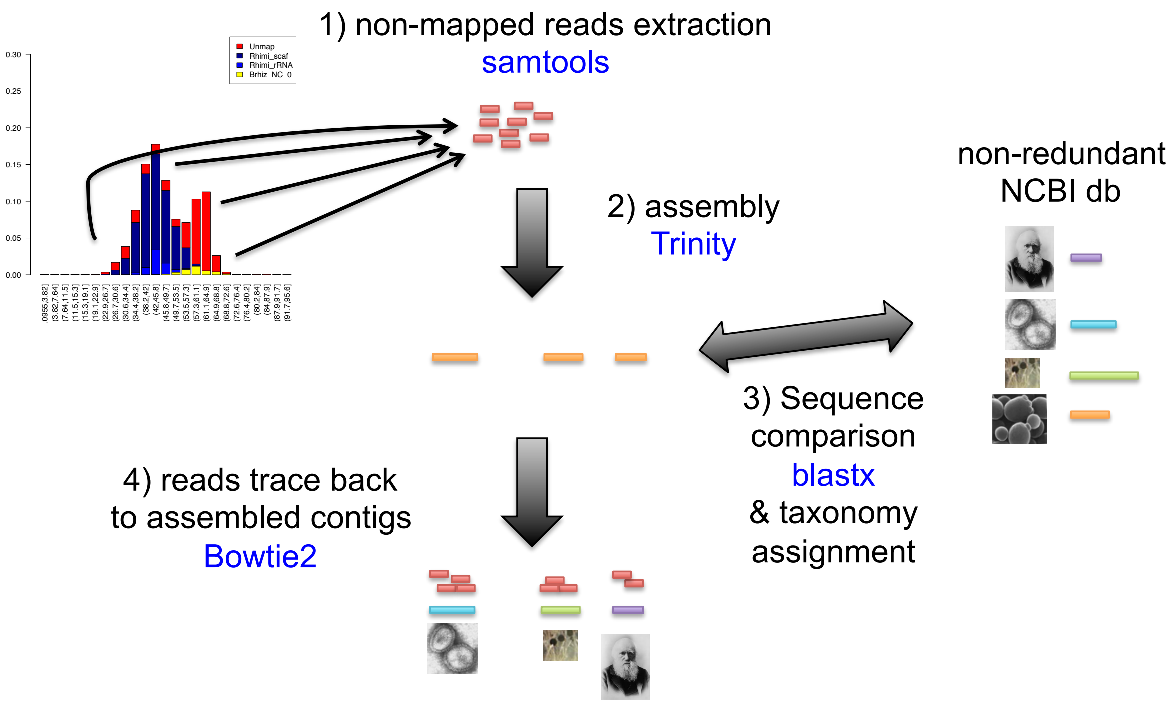


Figure 5. A strategy to trace origin of unmapped reads. Software used in each step is shown (blue) with a brief description (black). Unmapped reads (red), assembled contigs (orange), and sequences with associated taxonomy information display different colors

Fungal comparative genomics

Our fungal comparative genomics analysis involves two different layers. The first one involves Mucorales genomes with an emphasis on host and non-host *Rhizopus* strains. The second layer involves a broad comparison of Mucorales and Dykaria (Ascomycota and Basidiomycota) representative genomes.

Mucorales genomes analyzed are shown in Table 12.

Table 12. Mucorales genomes used for comparative genomics. JGI Joint Genome Institute, HKI Hans Knoll Institute. CCTCC China Center for Type Culture Collection. BI, Broad Institute

Strain	Symbiont	Genome size (Mb)	Prot genes	Sequenced by	N50 Kb	# Scaffolds	Key
Rhizopus microsporus var microsporus ATCC52813	B4	26	10,905	JGI	111	131	Rm13
Rhizopus microsporus var microsporus ATCC52814	B7	25	11,502	JGI	105	560	Rm14
Rhizopus microsporus (*Rhizopus chinensis* Rh-2) ATCC62417	B1	48	18,869	HKI	198	1,386	Rm17
Rhizopus microsporus var chinensis ATCC11559	-	48	19,563	HKI	53	1,554	Rm59
*Rhizpus microsporu*s var chinensis CCTCC M201021	-	45	17,676	CCTCC	30	3,281	Rm21
Rhizopus oryzae 99-880	-	46	17,467	BI	310	81	Ro
Mucor circinelloides CBS 277.49 V2.0	-	36	11,719	JGI	4,318	80	Mc
Phycomyces blakesleeanus NRRL1555 V2.0	-	53	16,528	JGI	1,515	26	Pb

Defining protein families in Mucorales

We built a protein database containing the predicted proteins of Mucorales genomes. We used blastp to make an all versus all protein comparison (Altschul et al. 1997). We asked for a particular output format that specified the compared sequences length, alignment length and number of gaps in the alignment. With this, it was possible to calculate mutual sequence coverage for each alignment (Table 13).

Table 13. Parameters used in blastp sequence comparison

Parameter	Explanation
-db Mucorales	Mucorales database path and name
-num_threads 8	Number of processor to use
-query Mucorales.fasta	Input file in fasta format
-evalue 0.001	Threshold e-value to report alignments
-outfmt '6 qacc sseqid pident evalue bitscore length qlen slen gaps'	Tabular output format without header with specified fields
-out out_file.tab	Output file name
qacc	Query sequence name (accession name)
sseqid	Subject sequence id
pident	Percentage of identity
bitscore	Bitscore of alignment
length	Length of alignment
qlen	Query sequence length
slen	Subject sequence length
gaps	Number of gaps present in alignment

We used a mutual coverage of 70%, and a minimum identity of 50% as thresholds. A network was defined by connecting all pairs of proteins satisfying these criteria. We ran MCL (Enright et al. 2002) to define clusters in this network, using an inflation parameter of 1.4. Low inflation leads to coarser clusters, high inflation leads to fine-grained clusters. The resulting MCL clusters were considered protein families.

Descriptions of general Mucorales genome properties were made in R, aided by the seqinr package.

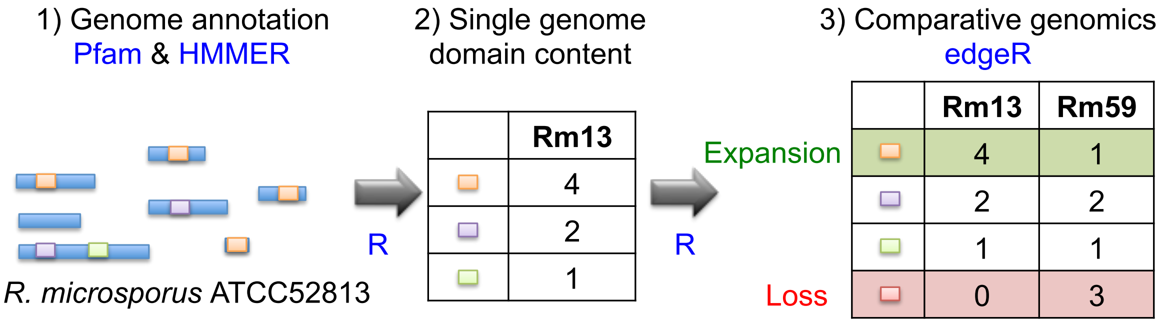
Mucorales phylogeny reconstruction

We built a phylogeny of 8 Mucorales selecting protein families present in all genomes with a single member. A total of 46 families meet this criterion. We aligned each protein family separately with muscle (Edgar 2004) and then concatenated the alignment with perl scripts. We built a phylogeny with MrBayes (Ronquist et al. 2012). We used a Poisson amino acid substitution model. We ran two independent Monte Carlo Markov chains for 1,130,000 generations, sampling every 1,000 generations. All node posterior probabilities have 100% support and a standard deviation of 0. We rooted the resulting phylogeny manually with Phycomyces as outgroup. The phylogeny was drawn using FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Comparison of protein domains in fungi

We compared protein domain content in fungi. We analyzed the Mucorales listed in Table 12 and included the following Ascomycota representatives: *Neurospora crassa*, *Aspergillus nidulans*, *Trichoderma atroviride* and *Saccharomyces cerevisisae,* as well as Basidiomycota representatives: *Cryptococcus neoformans* JEC21 and *Ustilago maydis.* We selected these Dykarial genomes based on the abundance of molecular studies and their relevance as fungal model organisms.

We came up with a strategy to compare the protein domain contents in fungal genomes; our approach is independent of the organisms studied (Figure 6).



	Rm13
🟧	4
🟪	2
🟩	1

	Rm13	Rm59
🟧	4	1
🟪	2	2
🟩	1	1
🟥	0	3

Figure 6. Comparative genomics strategy for functional categories, Pfam domains are shown as an example. Software used in each step is shown (blue) with a brief description (black). Protein sequences (blue), domains are represented as different color rectangles.

1) The first step involves the annotation of all genomes with Pfam as previously mentioned.
2) Then we built a table with the number of proteins associated with each domain for every analyzed genome.
3) We merged this information in a single table were columns represent fungal genomes and rows represent every Pfam domain. In this table, every cell corresponds to the number of proteins that have that particular domain in a given genome. All table processing was done in R. To perform comparative analysis of fungal protein domain contents we used the edgeR R package (Robinson et al. 2009).

The edgeR package was originally designed to deal with digital expression data such as serial analysis of gene expression (SAGE), or more recently, next-generation RNA sequencing (RNA-Seq) (McCarthy et al. 2012). Their developers argue that its software could have broader applications; it is well suited to manage discrete counts that follow a negative binomial distribution. There are certain analogies between differential expression and comparative genomics; these are listed in Table 14.

Table 14. Analogies between RNA-Seq and comparative genomics analyses

RNA-Seq	Comparative genomics analog
Genes	Protein domains
Over-expression	Expansion
Down-regulation	Reduction
RNA-Seq library size	Genome size

The package edgeR has tools that are useful to our comparative genomics analysis. First, it is common that sequencing libraries differ in sizes, to deal with this edgeR uses normalization factors to correct for different library sizes (with the function *calcNormFactors*). Differing genome sizes are a common scenario in comparative genomics analysis. By using this package we gain a statistical framework and our results are supported by significance values. Finally, our analysis implies many individual tests. edgeR considers the number of test performed and corrects reported p-values with the *decideTestDGE* function. We used the BH method for multiple testing correction of p-values (Benjamini & Hochberg 1995), which returns false discovery rate values (FDR).

We filtered for lowly represented protein domains. These were defined as those having less than 10 counts per million in at least 5 genomes.

The value of cpms for each Pfam category relies on the size of the genome. In this case, the number of proteins that have any Pfam annotation for a given genome represents the size of that genome. In a category where all genomes have one protein, the cpm values range from 29 cpm in *R. microsporus* ATCC62417 to 113 cpm in *Saccharomyces cerevisiae.* Therefore the filter of 10 cpms will consider all non-zero categories. The true filter then turns to be the distribution of that category, as it needs to be represented in at least 5 out of 14 fungal genomes.

With this filter we reduced de number of Pfam domains to test from 4,895 to 3,610.

We focused our efforts in two comparisons:

- Host strains vs non-host *Rhizopus* strains: ATCC52813 + ATCC52814 + ATCC62417 vs ATCC11559 + CCTCC M201021.
- Mucorales vs Dykaria: *Rhizopus* + *Mucor* + *Phycomyces* vs *C. neoformans* + *U. maydis* + *N. crassa* + *A. nidulans* + *T. atroviridae* + *S. cerevisisae*.

All edgeR steps are similar to a differential expression analysis, with the exception of data dispersion estimation. We used the general dispersion estimation (*estimateGLMCommonDisp* function) instead of using more sophisticated trended and tagwise estimations.

Bacterial comparative genomics

All finished *Burkholderia* genomes were downloaded from the NCBI ftp site (Table 15). By analyzing finished genomes we guarantee that observed differences are not caused by assembly differences. We allowed a couple of exceptions and included two draft genomes of particular interest: *Candidatus* Glomeribacter gigasporarum and *Burkholderia terrae* BS001. *Ca.* G. gigasporarum is a fungal endosymbiont of the arbuscular mycorrhizae *Gigaspora margarita* (Salvioli et al. 2010), on the other hand, *B. terrae* BS001 is known to establish mutualistic relationships with several fungal species (Nazir et al. 2014). Caution should be taken when interpreting results related to these two genomes.

Table 15. Burkholderia genomes analyzed. In status column F and D stand for Finished and Draft respectively. BCC stands for Burkholderia cepacia complex.

Name	Short name	Lifestyle	Status
Burkholderia pseudomallei 1026b	pseu_1026b	Human pathogen	F
Burkholderia pseudomallei 1106a	pseu_1106a	Human pathogen	F
Burkholderia pseudomallei 1710b	pseu_1710b	Human pathogen	F
Burkholderia pseudomallei 668	pseu_668	Human pathogen	F
Burkholderia pseudomallei BPC006	pseu_BPC006	Human pathogen	F
Burkholderia pseudomallei K96243	pseu_K96243	Human pathogen	F
Burkholderia pseudomallei MSHR305	pseu_MSHR305	Human pathogen	F
Burkholderia pseudomallei MSHR346	pseu_MSHR346	Human pathogen	F
Burkholderia pseudomallei NCTC 13179	pseu_13179	Human pathogen	F
Burkholderia cenocepacia AU 1054	ceno_1054	Human pathogen	F
Burkholderia cenocepacia HI2424	ceno_HI2424	Human pathogen	F
Burkholderia cenocepacia J2315	ceno_J2315	Human pathogen	F
Burkholderia cenocepacia MC0 3	ceno_3	Human pathogen	F
Burkholderia cepacia GG4	cepa_GG4	Human pathogen	F
Burkholderia ambifaria AMMD	ambi_AMMD	Human pathogen, BCC	F
Burkholderia ambifaria MC40 6	ambi_6	Human pathogen, BCC	F
Burkholderia thailandensis E264	thai_E264	Seldom pathogen, soil	F
Burkholderia thailandensis MSMB121	thai_MSMB121	Seldom pathogen, soil	F
Burkholderia lata 383	lata_383	BCC, soil	F
Burkholderia gladioli BSR3	glad_BSR3	Human pathogen, plant pathogen, plant, fungi symbiont	F
Burkholderia glumae BGR1	glum_BGR1	Plant pathogen	F
Burkholderia KJ006	KJ006	Endophyte, antifungal activity	F
Burkholderia multivorans ATCC 17616	mult_17616	Human pathogen	F
Burkholderia vietnamiensis G4	viet_G4	Human pathogen	F
Burkholderia rhizoxinica HKI 454	rhiz_454	Fungal endosymbiont	F
Burkholderia CCGE1001	CCGE1001		F
Burkholderia CCGE1002	CCGE1002	Plant endophyte, Legume nodule	F
Burkholderia CCGE1003	CCGE1003		F
Burkholderia phymatum STM815	phym_STM815	Plant endophyte, Nitrogen fixation	F
Burkholderia phytofirmans PsJN	phyt_PsJN	Plant endophyte, Onion roots	F
Burkholderia phenoliruptrix BR3459a	phen_BR3459a	Plant endophyte, Nitrogen fixation, legume nodule	F

Burkholderia RPE64	RPE64	Insect endosymbiont	F
Burkholderia YI23	YI23	Soil, bioremediation	F
Burkholderia xenovorans LB400	xeno_LB400	Bioremediation, PCB degrader, contaminated soil	F
Burkholderia mallei ATCC 23344	mall_23344	Human pathogen	F
Burkholderia mallei NCTC 10229	mall_10229	Human pathogen	F
Burkholderia mallei NCTC 10247	mall_10247	Human pathogen	F
Burkholderia mallei SAVP1	mall_SAVP1	Human pathogen	F
Burkholderia terrae BS001	terr_BS001	Fungal symbiont	D
Candidatus Glomeribacter gigasporarum	CaG_gig	Fungal endosymbiont	D
Pandoraea RB 44	Pan_RB		F
Pandoraea pnomenusa 3kgm	Pan_pnomen	Human pathogen	F
Ralstonia eutropha JMP134	Ral_eu		F
Ralstonia solanacearum CFBP2957	Ral_sol	Plant pathogen	F

Genomes were annotated with Pfam, RAST and antiSMASH as previously described. We used these annotation tools for all genomes with the aim of making data more comparable.

Defining protein families in *Burkholderia*

We built families in bacterial genomes based on sequence similarity. First we built a protein database containing the predicted proteins of all selected bacterial genomes. We used blastp to make an all versus all protein comparison (Altschul et al. 1997); with same parameters as in Mucorales protein families (Table 13). We asked for a particular output format that specified the compared sequences length, alignment length and number of gaps in the alignment. With this, it was possible to calculate mutual sequence coverage for each alignment. We used a mutual coverage of 70%, a minimum identity of 50% as thresholds. A network was defined by connecting all pairs of proteins satisfying these criteria. We ran MCL (Enright et al. 2002) to find clusters in this network, we used an inflation parameter of 1.4. The resulting clusters were considered protein families.

We calculated the GC content of all bacterial genomes with the aid of seqinr R package.

Burkholderia core genome & phylogeny

The core genome of chosen bacteria was found with BBH-star, a software suite developed by Nelly Selem Mojica in the laboratory of Dr. Francisco Barona. Orthologous proteins are found by blastp, then groups are formed using the best bidirectional criteria. Proteins are considered part of the core genome if a complete graph of best bidirectional hit relationship is generated among all group members.

The core genome consists of 541 genes; these were used to build a phylogeny of the *Burkholderia* genus. Independent alignments were made using Muscle (Edgar 2004) for every group of orthologous proteins with default parameters. Then we eliminated non-informative alignment sites with Gblocks (Castresana 2000).

Table 16. Parameters used in Gblocks

Parameter	Explanation
-b4=5	Minimum length of a block (default 10)
-b5=n	Allowed gap positions (none=n, with half=h, all=a)
-b3=5	maximum number of contiguous non conserved positions (default 8)

The phylogeny was constructed using MrBayes (Ronquist et al. 2012). Two Monte Carlo Markov chains were run for 10,000 generations. We used a Poisson distribution model for amino acid substitutions.

All analyzed genomes belong to the Burkholderiales order. *Pandoraea* and *Ralstonia* genomes were chosen as out-groups to the *Burkholderia* genus. *Pandoraea* belongs to the Burkholderiaceae family, while *Ralstonia* belongs to the Ralstoniaceae family. We manually rooted the tree with *Ralstonia* genomes being the out-group, since *Ralstonia* doesn't belongs to the Burkholderiaceae family but to Ralstoniaceae.

Functional categories comparison

To perform comparative analysis of Pfam domains, RAST subsystems and predicted functions we used the edgeR R package (Robinson et al. 2009) in a similar way to fungal comparative genomics. We filtered lowly represented functional categories are shown in Table 17.

Table 17. Low represented functional categories. Cpm stands for counts per million

Category	Low representation filter
Pfam	Less than 100 cpm in at least 30 bacterial genomes
Subsytems	Less than 100 cpm in at least 30 bacterial genomes
Proposed functions	Less than 100 cpm in at least 22 bacterial genomes

In subsystems, a count of 1 corresponds to 375 cpm in *B. rhizoxinica,* 817 cpm in *Ca.* G. gigasporarum and a mean of 193 cpm for free-living relatives. Therefore we asked for subsystems to be represented by at least one protein in at least 30 of 44 bacterial genomes. With this filter we reduced de number of subsystems to test from 813 to 511.

Using Pfam domains, a count of 1 corresponds to 191 cpm in *B. rhizoxinica,* 817 cpm in *Ca.* G. gigasporarum and a median of 90 cpm for free-living relatives. Therefore we asked for Pfams to be represented by at least two proteins in at least 30 of 44 bacterial genomes. With this filter we reduced de number of Pfam domains to test from 4,533 to 1,121.

Using RAST proposed functions, a count of 1 corresponds to 459 cpm in *B. rhizoxinica,* 1074 cpm in *Ca.* G. gigasporarum, and a mean of 218 cpm for free-living relatives. Therefore we asked for Pfams to be represented by at least two proteins in at least 22 of 44 bacterial genomes. With this filter we reduced de number of Pfam domains to test from 11,557 to 2,918.

Our filters can be considered very lax, as most of them allow scenarios of at least one protein per category. The scrict component of our filters is comprised by the widespread occurrence of the category as 22 or 30 members of the Burkholderaceae most have the category to be considered.

We compared functional categories contents of:

- *B. rhizoxinica* vs all free-living *Burkholderia*.
- endofungal bacteria (*B. rhizoxinica* HKI 454 + *Ca*. G. gigasporarum) vs all free-living *Burkholderia*.

We used the “Analysis of Phylogenetics and Evolution” (ape) R package (Paradis et al. 2004) to import our *Burkholderia* core genome phylogeny and to add information of Pfam, predicted functions and subsystem counts to the species. More than 14,000 images were generated this way; all these images are available in the digital supplementary material (Supporting_material/2.Fungal_comparative_genomics/Mucorales_core_family_proteins). With functional category counts in a phylogenetic context it is easier to infer expansions and reductions.

Search for *Burkholderia* toxin-antitoxin proteins in Mucorales

We identified all Pfam domains associated with toxin-antitoxin systems present in *Burkholderia* genomes (Table 18).

Table 18. Toxin-antitoxin associated Pfam domains. T and A stand for toxin and antitoxin respectively. U stands for unkown

Pfam id	Description	TA
PF04221.7	RelB antitoxin	A
PF02604.14	Antitoxin Phd_YefM, type II toxin-antitoxin system	A
PF12910.2	Antitoxin of toxin-antitoxin stability system N-terminal	A
PF06769.8	Plasmid encoded toxin Txe	T
PF02452.12	PemK-like protein	U
PF01850.16	PIN domain	T
PF05509.6	TraY domain	U
PF05015.8	Plasmid maintenance system killer protein	T
PF02661.13	Fic/DOC family	U
PF13470.1	PIN domain	T
PF04014.13	Antidote-toxin recognition MazE	A
PF05016.9	Plasmid stabilisation system protein	U

We then looked for these domains in all previously annotated Mucorales proteins. These searches were performed in R.

Chapter III: RNA-Seq analysis Results

Narnaviruses in *Rhizopus*

A strange peak of reads with high-GC content

The GC-content distribution of the reads is a routine quality control analysis. The GC-content of a transcriptome of a single organism should be found near the mean GC-content of that organism. This distribution can be used to detect potential contaminations, which may be represented by peaks in the distribution with an unexpected GC. In a transcriptome of two organisms with dissimilar GC contents, these differences should be reflected as separate peaks in the distribution if enough reads are sampled from both organisms.

R. microsporus and *B. rhizoxinica* differ significantly in their average GC content, with 37% and 59% respectively. We wanted to know if this difference could be reflected in the GC content distribution for all reads. An example result is shown in Figure 7. The distribution for ATCC52813 libraries with the bacterium (A13b) shows a slight positive skew, this skew is not apparent in cured libraries (A13c) (Figure 7).

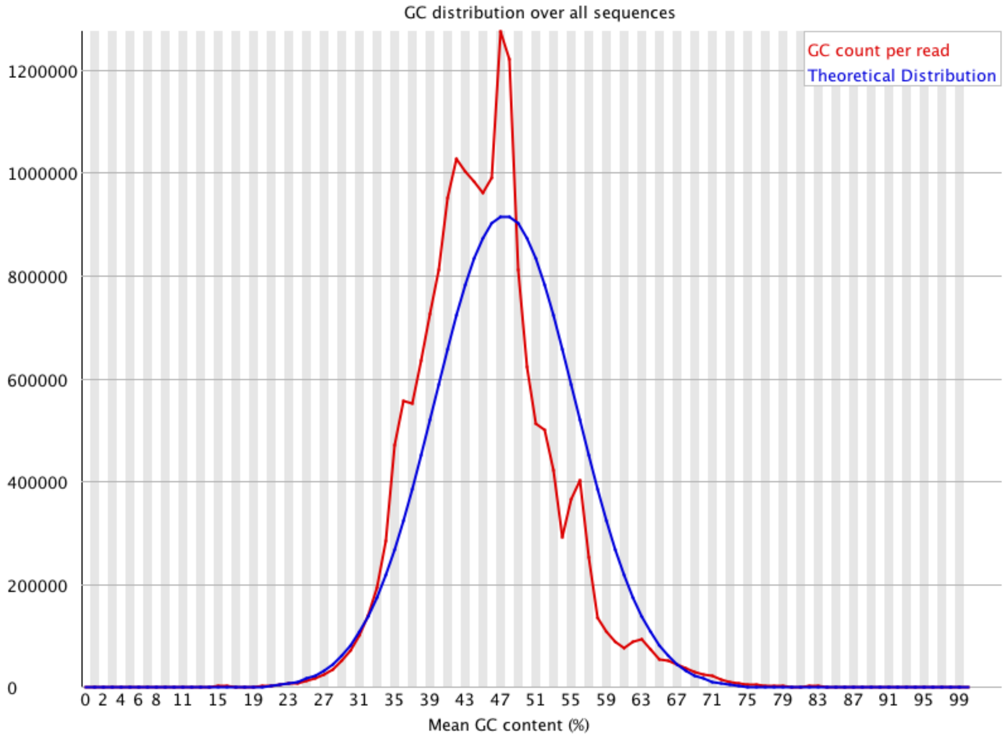


Figure 7. Per sequence GC content distribution. ATCC52813 with B. rhizoxinica is shown. R. microsporus and B. rhizoxinica GC content are 37% and 59%, respectively. GC count per read (red). Theoretical normal distribution (blue) centered on observed mean data.

We detected an abnormal GC peak at around 60% in all A14 libraries (*Figure 8*). Interestingly the peaks don't disappear in A14 cured samples (A14c), this made us suspect that they did not correspond to *B. rhizoxinica* transcripts.

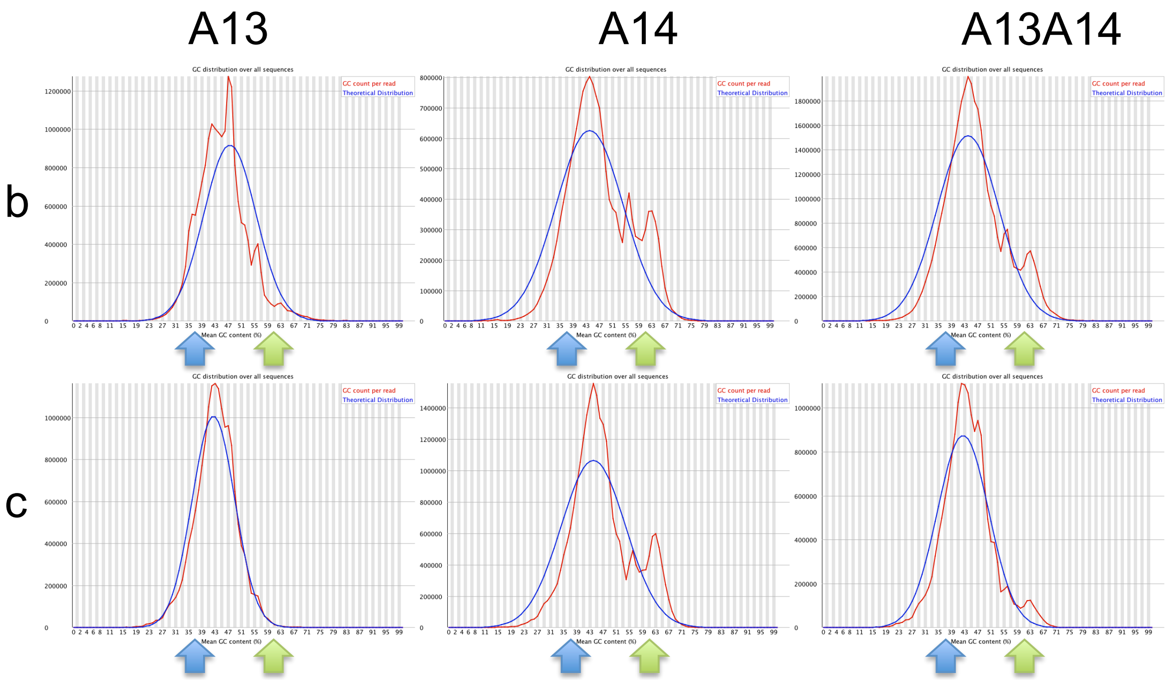


Figure 8. High GC reads in ATCC52814 libraries. GC content distribution for all libraries. R. microsporus GC content is denoted by a blue arrow. B. rhizoxinica GC content is denoted by a green arrow. Libraries information is shown in a matrix-like arrangement.

To rule out the possibility that the high GC reads mapped to *B. rhizoxinica*, we determined the GC content of each read and whether it mapped to *R. microsporus*, *B. rhizoxinica* or none of them. An example result of this analysis is shown in Figure 9. Mapping information of A14b library is binned according to the reads' GC content. This approach revealed that the high GC reads predominantly remain unmapped. Endosymbiont transcripts match their expected GC content.

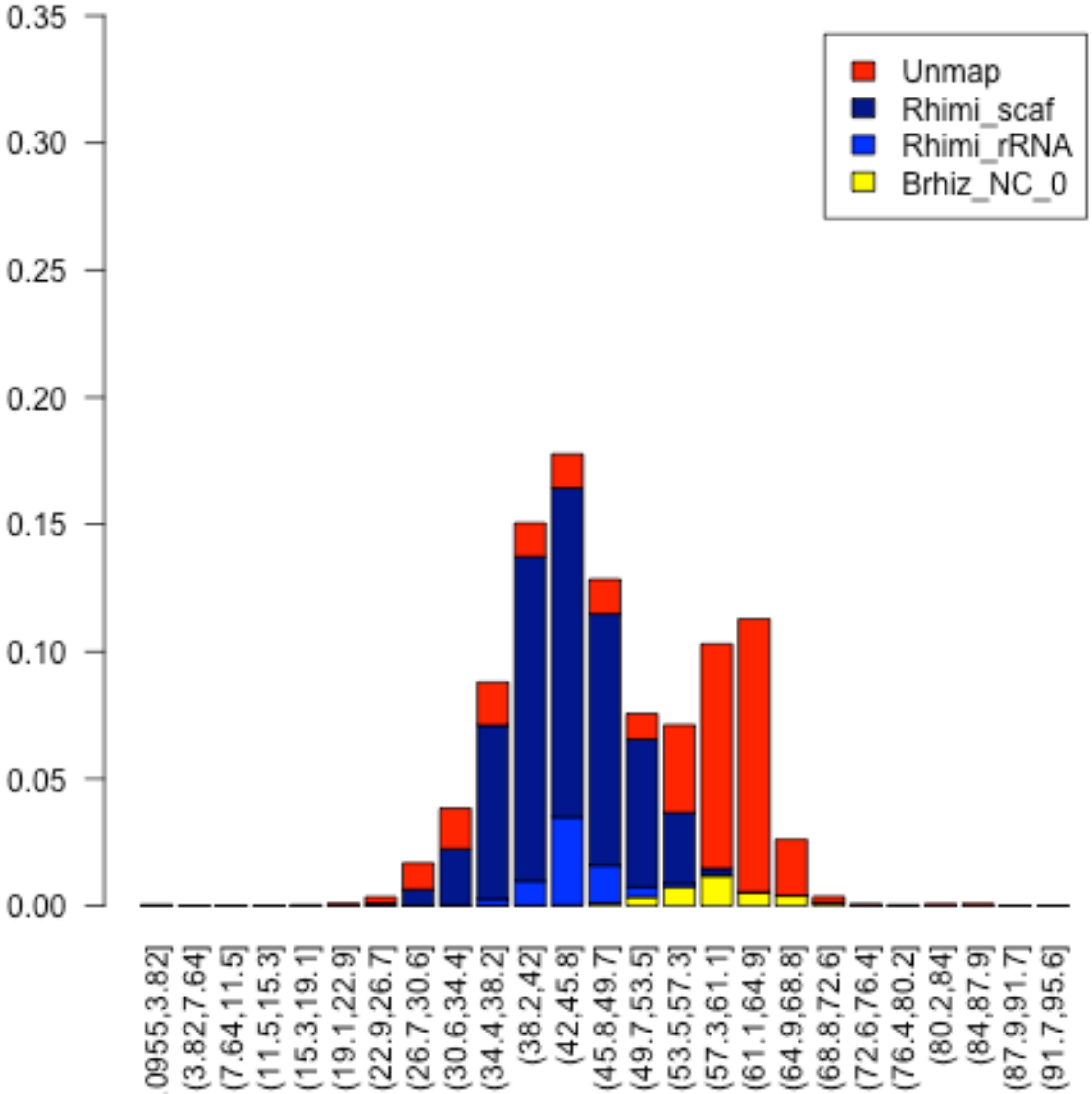


Figure 9. GC content of all reads with mapping information for ATCC52814 with B. rhizoxinica. Reads mapping to: B. rhizoxinica (yellow), R. microsporus (dark blue), R. microsporus rRNA (light blue) and non-mapping reads (red).

The same information for all libraries is shown in Figure 10. The reads of the abnormal ATCC52814 high GC peak don’t map to the host or the endosymbiont genome. Biological replicates behave the same way.

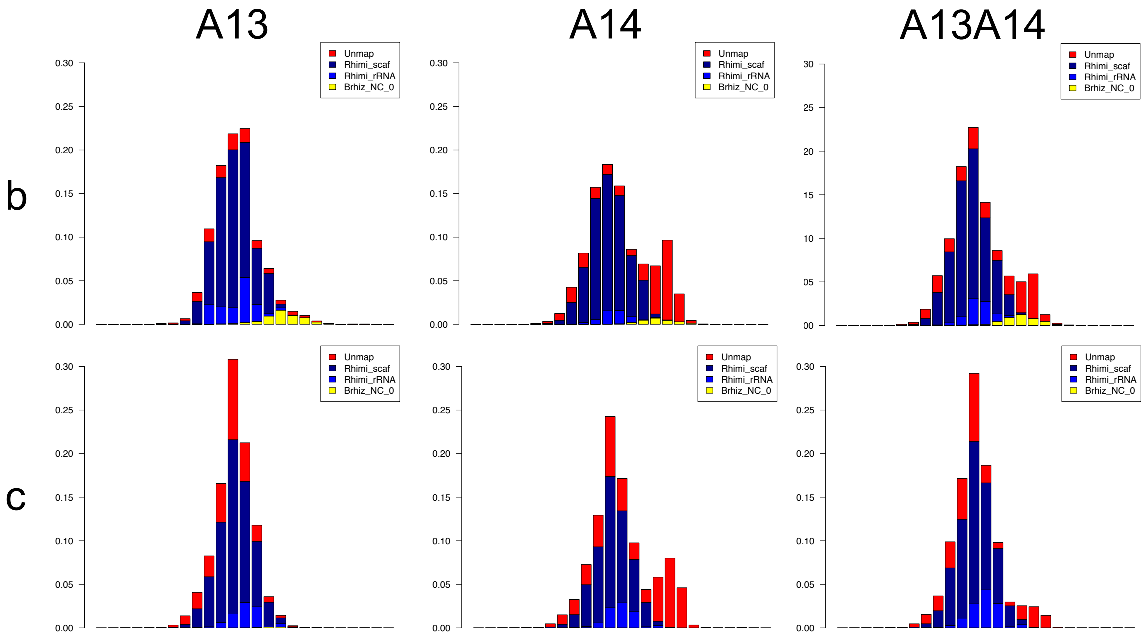


Figure 10. High GC reads of ATCC52814 don’t map to R. microsporus or B. rhizoxinica. Per sequence GC content with mapping information. Reads mapping to: B. rhizoxinica (yellow), R. microsporus (dark blue), R. microsporus rRNA (light blue) and non-mapping reads (red).

Tracing peak origin and discovery of *Narnavirus* sequences

We were curious about the origin of these high GC reads. We assembled the unmapped reads and added taxonomy information based on sequence similarity to the nr NCBI database (for detailed procedure see Methods section). To consider the expression levels, we mapped the reads back onto the newly assembled contigs. The result of this analysis is shown in Figure 11.

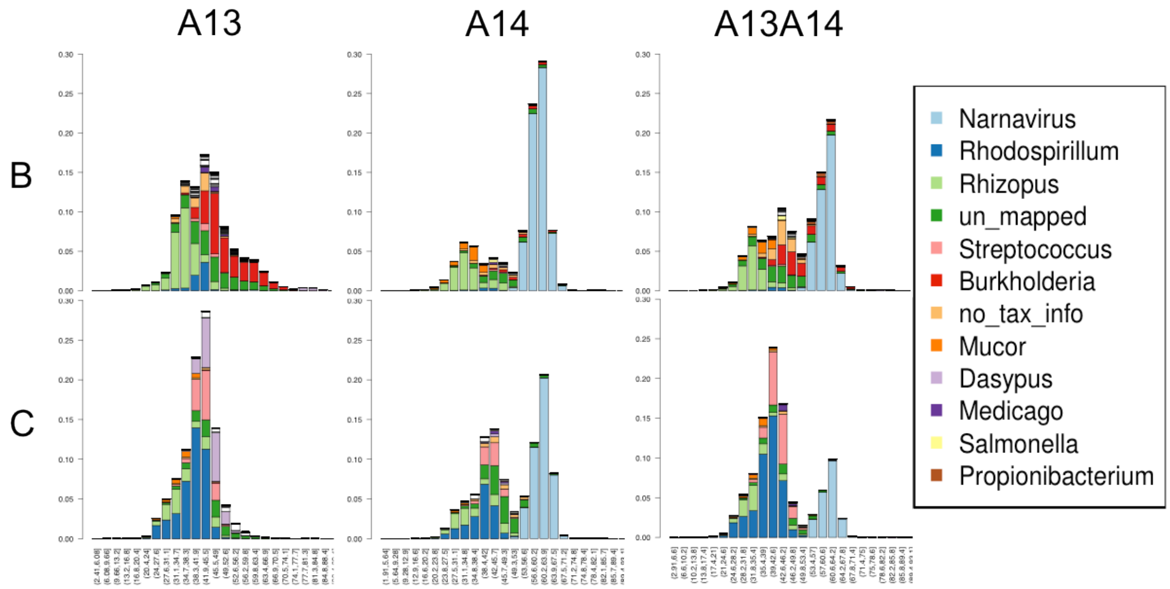


Figure 11. The high GC peak probably originates from narnaviruses. A13, ATCC52813. A14 ATCC52814, A13A14, ATCC52813 co-cultered with ATCC52814. B, with B. rhizoxinica. C, cured host strain(s).

We found some *Rhizopus* sequences in all cases and *Burkholderia* sequences only in the non-cured libraries (Figure 11). These reads mapped to the assembled transcriptome, yet they didn't map to the reference genome. This could be due to reads aligning to splice-junctions or genes that are not present in the genomes used as reference.

Sequences similar to *Narnavirus* genus coincide with the abnormal high GC peak (Figure 11). Narnavirus reads are found consistently in all libraries involving ATCC52814 strain, but they are absent from ATC52813 isolated libraries. These observations fit a scenario in which ATCC52814 is infected with *Narnavirus* but ATCC52813 is not. Biological replicates behave in a similar way (data not shown).

Narnavirus signal is present in all replicates of all ATCC52814 libraries, which indicates that it is consistently present in the strain used in this experiment. An important point to determine is whether ATCC52814 viral infection is transient or is maintained through several generations.

General libraries description

The percentage of reads mapping to *B. rhizoxinica*, *R. microsporus* or *R. microsporus* rRNAs are shown in Figure 12. As expected, samples harboring the endosymbiont show many more reads mapping to *B. rhizoxinica* than cured samples. On average 3.8% of the reads map to *B. rhizoxinica* in wild type host libraries, in contrast 0.16% of the reads map to the endosymbiont in cured libraries. We manually explored the "endosymbiont" reads from the cured libraries, finding that most of these reads map to rDNA regions. Upon further verification, they most likely come from mitochondria (data not shown).

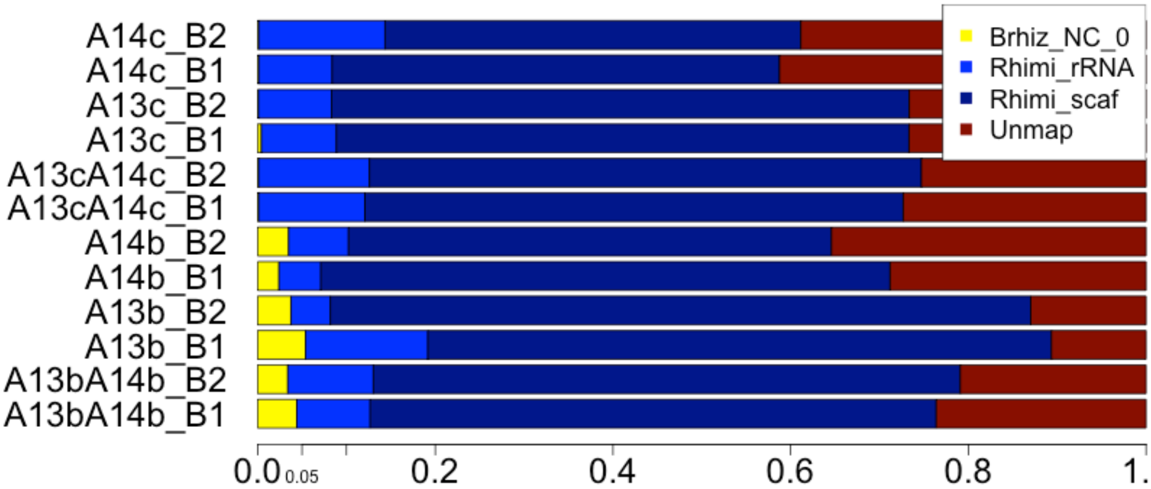


Figure 12. Percentage of reads mapping to: B. rhizoxinica (yellow), R. microsporus (dark blue), R. microsporus rRNA (light blue) and non-mapping reads (red). Both biological replicates are shown (B1 and B2).

Ribosomal RNA (rRNA) constitutes the majority of RNA molecules in any cell; it typically constitutes 95% of total cellular RNA (Westermann et al. 2012). Sequencing total RNA samples would result in 95% of non-informative rRNA molecules. In order to avoid this redundancy, special kits have been designed to remove as much rRNA as possible. In our case, rRNA removal kits were used when preparing RNA for sequencing (Mondo and Pawlowska, unpublished results). According to our mapping results, approximately 9.25% of the reads map to *Rhizopus* rRNAs. We consider this a successful rRNA removal, given the expected initial rRNA quantities.

As an overview of the behavior of the number of the reads mapping to each fungal gene, we made a principal component analysis (PCA) plot (Figure 13). The PCA analysis correctly separates all samples according to their biological properties. Two diagonal lines separate mating (top left corner of the plot), asexually sporulating (intermediate) and cured libraries (bottom and right). All biological replicates cluster together. Therefore, principal component analysis separates libraries according to developmental stage of the fungus.

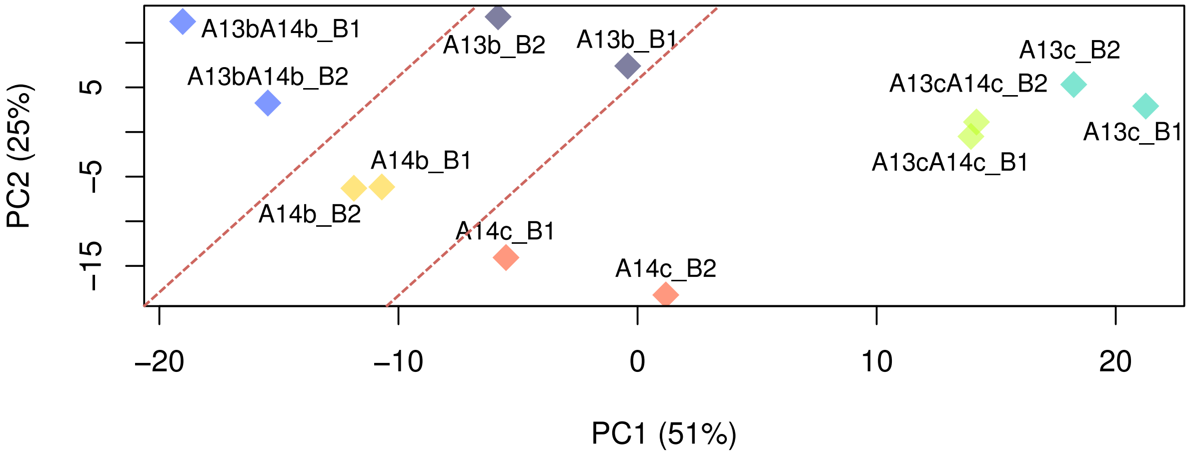


Figure 13. Principal component analysis for gene counts of R. microsporus. Biological replicates are denoted as B1 or B2 and are colored the same.

We conclude that the RNA-Seq results have good quality based on the previous analyses. Cured fungal libraries show little if any reads mapping to the endosymbiont, biological replicates behave similarly and data separates according to physiological state. We then proceeded to perform differential expression of *R. microsporus* genes.

Housekeeping genes proposal of *Rhizopus microsporus* and *Burkholderia rhizoxinica*

We looked for genes with low variance across all libraries with the aim of using them as reference genes in relative quantification experiments such as qRT-PCR. We performed this analysis for both *Rhizopus microsporus* and *Burkholderia rhizoxinca*. These genes would represent an important resource for further studies.

Proposed reference housekeeping genes for *Rhizopus microsporus* are shown in Table 19. We included genes with different expression levels and a low variance across libraries.

Table 19. Proposed housekeeping genes for Rhizopus microsporus ATCC52813 and ATCC52814

No.	Protein id	Expression level	Annotation
1	239509	Low	mRNA splicing factor SYF2
2	251630	Medium	Mediator of RNA polymerase II transcription subunit 1
3	68175	Medium	DNA polymerase family B
4	8316	Medium High	Ribosomal protein L13e
5	287373	High	Acetyl-coenzyme A transporter 1
6	244183	High	Ribosomal S17

We next compared the expression of our proposed housekeeping genes with those used by Dolatabati. These reported genes include actin (ACT) and translation enlongation 1-alpha (Dolatabadi et al. 2013). We generally observe that these two genes display high expression levels. To perform relative expression quantifications it is preferable to interrogate a desired gene with a gene reference that has similar expression levels. We propose six genes with different expression levels that may serve as reference for *R. microsporus* qRT-PCR experiments Figure 14.

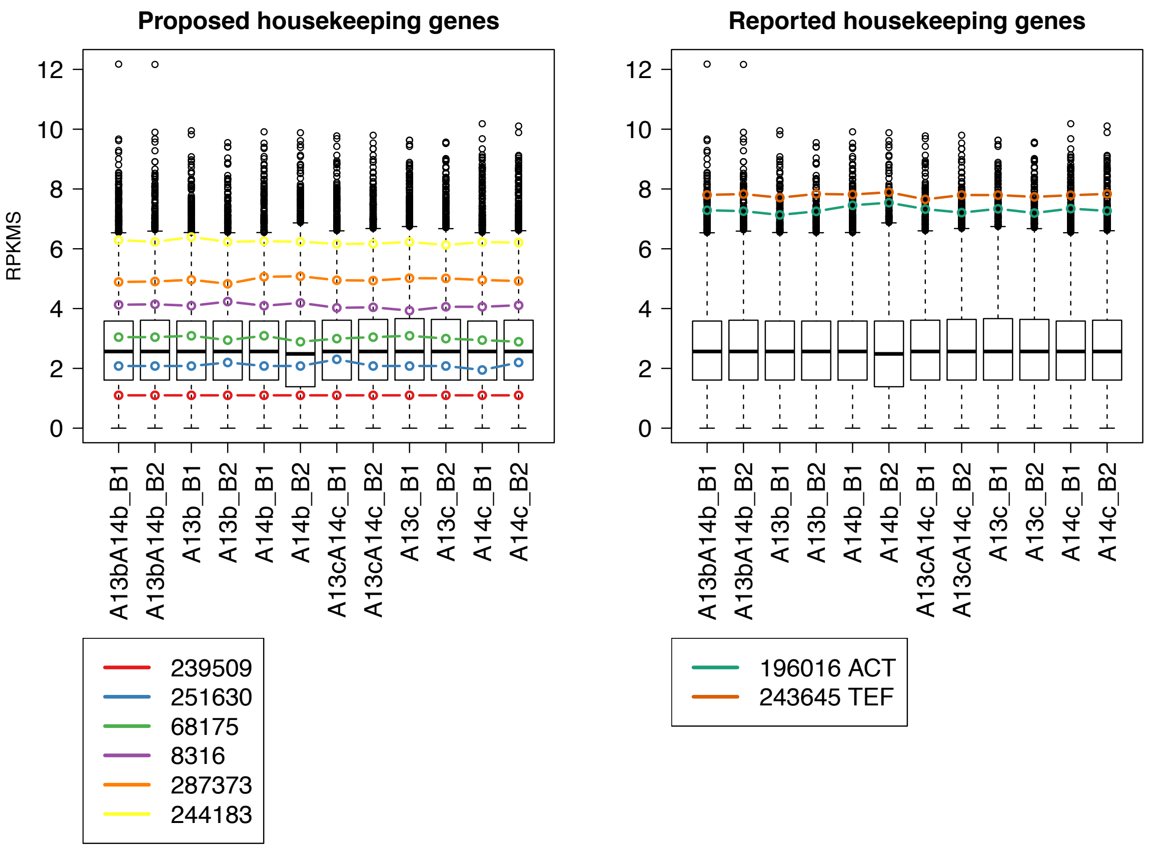


Figure 14. Expression of proposed and reported housekeeping genes for Rhizopus microsporus. Protein Ids for proposed and reported genes. ACT, actin. TEF, Translation enlongation factor 1-alpha

We then performed the same procedure and looked for genes with low expression variance and different expression levels in *Burkholderia rhizoxinica*. The proposed seven housekeeping genes are shown in Table 20. Chosen genes expression levels range from low to very high.

Table 20. Proposed housekeeping genes for Burkholderia rhizoxinica B4 and B7

No.	Protein id	Expression level	Annotation
1	fig|32008.60.peg.1051	Low	UDP-N-acetylmuramate--alanine ligase (EC 6.3.2.8)
2	fig|32008.60.peg.1895	Medium	Chorismate synthase (EC 4.2.3.5)
3	fig|32008.60.peg.1098	Medium	Biotin carboxyl carrier protein of acetyl-CoA carboxylase
4	fig|32008.60.peg.2835	Medium High	DNA polymerase III alpha subunit (EC 2.7.7.7)
5	fig|32008.60.peg.3072	Medium High	RecA protein
6	fig|32008.60.peg.2328	High	Glucose-6-phosphate isomerase (EC 5.3.1.9)
7	fig|32008.60.peg.3501	Very High	DNA topoisomerase III, Burkholderia type (EC 5.99.1.2)

We then compared the expression of our proposed bacterial housekeeping genes with those used by Lackner et al. (Lackner et al. 2009). These reported bacterial housekeeping genes are listed in Table 21 and the expression levels comparison is shown in Figure 15.

Table 21. Reported housekeeping genes for the Burkholderia genus

gene	Gene ids	Annotation
lipA	RBRH_02213 YP_004030340.1 fig|32008.60.peg.3312	Leipoate synthase
lipA	RBRH_03512 YP_004022683.1 fig|32008.60.peg.715	Leipoate synthase
gmhD	RBRH_01723 YP_004028383.1 fig|32008.60.peg.1428	ADP-L-glycero-D-manno-heptose-6-epimerase (EC 5.1.3.20)
lepA	RBRH_02448 YP_004028462.1 fig|32008.60.peg.1509	Translation elongation factor LepA
gltB	RBRH_02313 YP_004028754.1 fig|32008.60.peg.1776	Glutamate synthase [NADPH] large chain (EC 1.4.1.13)
ace	RBRH_03709 YP_004029038.1 fig|32008.60.peg.2048	Acetoacetyl-CoA reductase (EC 1.1.1.36)
ndh	RBRH_02530 YP_004029817.1 fig|32008.60.peg.2779	NADH-ubiquinone oxidoreductase chain G (EC 1.6.5.3)

We observe that reported housekeeping genes mostly display an above the median expression level. Our seven proposed *Burkholderia* housekeeping genes intentionally have different expression levels Figure 15.

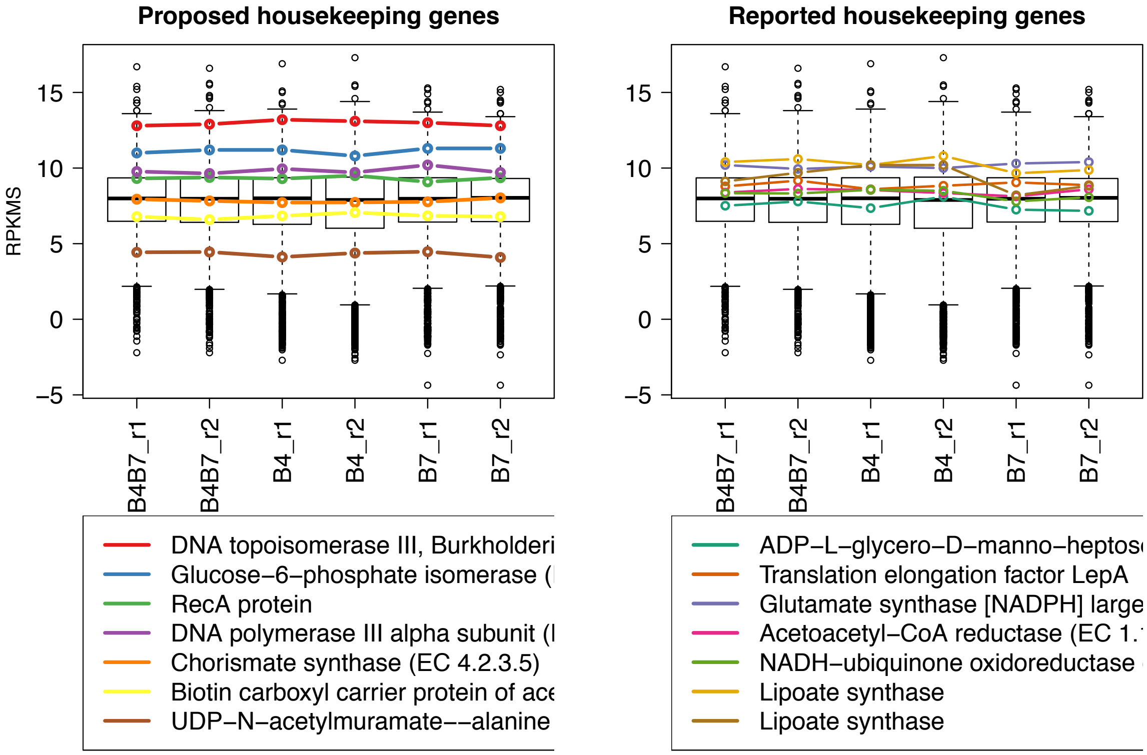


Figure 15. Boxplot proposed and typical housekeeping genes for Burkholderia

Differential expression analysis

The goal of this section was to identify differentially expressed (DE) genes in *R. microsporus* due to the presence or absence of *Burkholderia rhizoxinica.* These genes could be relevant for the interaction with the bacterial endosymbiot.

When interpreting a DE result it is important to note that it can't reveal the underlying mechanism. For example, if a gene shows more counts in condition A vs B, it doesn’t necessarily means that it is activated in A. An alternative explanation is that it could be repressed in B. To simplify the description of these results we shall use the convention of referring to up-regulated genes as those with more counts in presence of the bacterium and we shall call down-regulated genes those with more counts in absence of the bacterium. This convention will be followed for the rest of this text.

Several comparisons are possible, resulting in different sets of DE genes (Table 22).

Table 22. Possible contrasts comparing conditions with to without endosymbiont. Numbers and percentages of differentially expressed (DE) genes. Using a FDR of 0.05 and a logFC of 0 as thresholds.

Contrast	DE genes No. and (%)	Up-regulated No. and (%)	Down-regulated No. and (%)	Libraries involved
A13	2752 (25%)	1209 (11%)	1543 (14%)	A13b vs A13c
A14	2633 (24%)	1189 (10%)	1442 (13%)	A14b vs A14c
Asexual sporulation	3647 (33%)	1721 (15%)	1926 (17%)	A13b and A14b vs A13c and A14c
Mating	3831 (34%)	1874 (16%)	1957 (17%)	A13bA14b vs A13cA14c
All bacterial effect	4720 (43%)	2375 (21%)	2345 (21%)	A13b and A14b and A13bA14b vs A13c and A14c and A13cA14c
Interaction	1467 (13%)	-	-	(A13b vs A14b) vs (A13c vs A14c)

Given so many possible comparisons, it is preferable to focus in just a few for a deeper exploration. The asexual sporulation comparison is particularly interesting for a couple reasons:

1. Asexual sporulation is the most common reproductive mode of *R. microsporus*.
2. In plates where mating occurs, asexual reproduction also happens. Zygospore formation concentrates in hyphae meeting points, and asexual sporangiophores develop elsewhere in the plate. By choosing the asexual sporulation comparison we make sure that DE are devoted to a single developmental program: sporangiophore formation and maturation.

The rest of this text shall focus mainly on the asexual sporulation contrast, unless something else is specified.

Functional enrichment analysis

DE analysis revealed a large fraction of the host genes being differentially expressed. The more drastic case is the comparison of all bacteria harboring libraries versus the “cured” ones, in this contrast 43% of *Rhizopus microsporus* genes are differentially expressed. Given that we have thousands of DE genes we used a functional enrichment analysis to get an overview of the biological processes being affected. We used Gene Ontology and Pfam domains as functional categories to be tested (see Methods).

Selected enriched functional categories for up-regulated and down-regulated genes are shown in Table 23 and in Table 24 respectively. We added comments for each category; some of them required a deeper exploration of available annotations. Complete tables are available in digital supplementary material (Supporting_material/1.Rhizopus_microsporus_differential_expression/functional_enrichment_analysis).

Up-regulated

Table 23. Selected functional categories enriched in up-regulated DE genes

	pvalue	FDR	Fun Cat	Comments
pyridoxal phosphate binding	7.5E-4	0.053	MF	Pyridoxal phosphate is a coenzyme frequently found in amino transferase enzymes, such as those involved in amino acid synthesis
DNA replication	1.6E-3	0.18	BP	Including two B family DNA polymerases and members of the mini chromosome maintenance complex, which have a role in the formation and elongation of the replication fork
iron ion binding	2E-3	0.092	MF	Includes two catalases, which are involved in protecting the cell against oxidative stress
steroid dehydrogenase activity	3.3E-3	0.11	MF	Enzymes responsible for the oxidation of sterols, components of fungal membranes
guanyl nucleotide binding	5.3E-3	0.15	MF	Including two tubulins and eight septins
carbohydrate derivative biosynthetic process	9.4E-3	0.35	BP	glycolipid and protein mannosyl-transferases, mannoproteins are part of the zygomycete cell wall
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	0.013	0.23	MF	10 of 22 have the cytochrome P450 domain. These heme proteins are the terminal oxidase enzymes in electron transfer chains
ribonucleoside-diphosphate reductase complex	0.015	0.26	CC	Catalyzes deoxyribonucleotides from ribonucleotides, deoxyribonucleotides are then used to synthesize DNA.
aromatic amino acid biosynthesis	0.016	0.35	BP	
Permease family	4E-05	0.024	Pfam	Members of this family have ten predicted transmembrane helices. For transport of xanthine, uracil and vitamin C. Many

				members of this family are uncharacterized and may transport other substrates.
Septin	4.6E-04	0.086	Pfam	Compartamentalization proteins. Involved in septum formation in fungi
AIG1 family	7.9E-04	0.086	Pfam	Arabidopsis AIG1 protein appears to be involved in plant resistance to bacteria
short chain dehydrogenase	8.5E-4	0.086	Pfam	Found in fatty acid biosynthesis proteins and in polyketide synthases
MCM2/3/5 family	1.1E-3	0.086	Pfam	Minichromosome maintenance complex has a role in the initiation and elongation phases of eukaryotic DNA replication, specifically the formation and elongation of the replication fork
Proteasome subunit	1.4E-3	0.095	Pfam	The majority of the proteasome complex components are up-regulated
KR domain	2.9E-3	0.17	Pfam	Found in fatty acid biosynthesis proteins and in polyketide synthases
ADP-ribosylation factor family	7.1E-3	0.34	Pfam	regulators of vesicular traffic and actin remodelling
Ras superfamily	0.044	0.73	Pfam	Small GTPases. Involved in cell proliferation, cytoskeletal dynamics and membrane trafficking
SET domain	0.048	0.73	Pfam	Histone methyl transferase domain

Down-regulated

Genes with a negative fold change are those that have more counts in absence of the bacterium. The results of the GO enrichment analysis for these genes are shown in Table 24.

Table 24. Selected functional categories enriched in Down-regulated DE genes

Category	P-value	FDR	Fun Cat	Comments
nucleic acid binding transcription factor activity	3.4E-14	4.7E-12	MF	Of the 100 down-regulated TFs, 62 have the sequence-specific DNA binding annotation and 37 bind to zinc ions
chromosome	1.8E-4	4.3E-3	CC	Including proteins with the CHROMO (CHRromatin Organisation Modifier) domain, core histones, H1 linker histone and a subunit of the origin recognition complex
cytoskeletal protein binding	5.8E-4	0.016	MF	Including proteins with the domain architecture of formins. Formins are involved in actin polymerization. In fungi, they are located in the hyphae tip
carbohydrate metabolic process	2.9E-3	0.088	BP	Hydrolases acting on glycosidic bonds and also polysaccharide deacetylases
lipase activity	6.7E-3	0.1	MF	5 of the 14 class III lipases. Class III lipases hydrolyze ester links in triacylglycerol releasing fatty acids. We also find 3 of the 6 D phospholipases, these enzymes release the head chains of phospholipids, leaving phosphatidic acid as a product. We also find a lysophospholipase, responsible for releasing fatty acids from lysophospholipids
cell communication	6.9E-3	0.1	BP	We find 8 of the 19 Ras guanine nucleotide exchange factors (GEF); these proteins activate Ras domain subfamily proteins. We also find 9 of the 30 RhoGEF and 10 of the 32 Rho GTPase activating proteins (GAP). GAP proteins

				deactivate its cognate G protein. RhoGEF and RhoGAP are specific regulators for Rho subfamily G proteins.
RNA polymerase complex	0.023	0.26	CC	Transcription initiation factor components such as the TFIID or TATA binding protein (TBP), TAF4 and TAF7. TFIID is part of the RNA polymerase II pre-initiation complex
Helix-loop-helix DNA-binding domain	1.6E-07	1E-4	Pfam	Present in some transcription factors
Chitin synthase	3.1E-06	6.3E-4	Pfam	Essential in synthesis of chitin, an important fungal cell wall compound. 9 out of 25 chitin synthases have a myosin domain
LysM domain	2.3E-4	0.022	Pfam	LysM containing domains bind to peptidoglycan and in plants these proteins are used to sense bacteria
Ricin-type beta-trefoil lectin domain	3.2E-4	0.022	Pfam	Lectins are carbohydrate-binding proteins. They bind sugar moieties with high specificity. In fungi, binding of lectins to non-self glycans results in toxicity towards predators.
Basic region leucine zipper	3.8E-4	0.023	Pfam	Domain commonly present in transcription factors
HSF-type DNA-binding	5.6E-4	0.027	Pfam	Domain commonly present in transcription factors
C2H2-type zinc finger	6.2E-4	0.027	Pfam	Domain commonly present in transcription factors
Polysaccharide deacetylase	0.0015	0.041	Pfam	Some of these proteins are annotated as chitin deacetylase. Chitin deacelyases are involved in chitosan biosynthesis
Core histone H2A/H2B/H3/H4	0.05	0.54	Pfam	Involved in DNA packing. Histone tails are modified to alter DNA compactation

Nuclear DE proteins: TF, Histones & SET

We found that 41% of the expressed transcription factors (TF) are down-regulated in presence of *B. rhizoxinica* (Figure 16). We also found an enrichment of many nuclear proteins in the list of down-regulated genes; this raises the possibility of TFs being enriched just because nuclear proteins are enriched.

To determine if TF are enriched because most of them are nuclear proteins we compared the fold change distribution of:

1. Transcription factors that are predicted to be located in the nucleus
2. Nuclear proteins that are not transcription factors
3. All other proteins that have a predicted cellular component location different than nucleus and a molecular function other than TF.

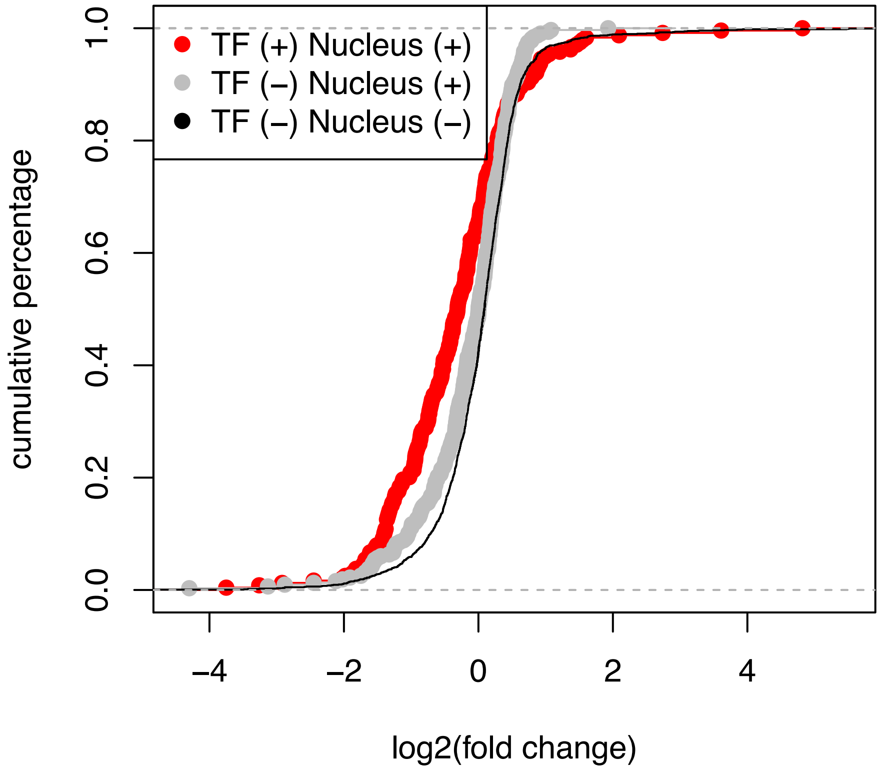


Figure 16. Fold change distribution for nuclear transcription factors (TF) (red) and non-transcription factor nuclear proteins (grey) and remaining proteins with a molecular function and a compartment prediction (black)

The distribution of non-TF and non-nuclear genes represents a reference to compare the expression patterns of nuclear proteins and transcription factors. A bias to down-regulation is detectable in non-TF nuclear proteins relative to our reference distribution (Wilcoxon test, non-paired data, p-value = 0.0006). Down-regulation of nuclear proteins represents more read counts in absence of bacteria. This result suggests the presence of more nuclei in plain mycelia relative to sporulating mycelia in *R. microsporus*. Nuclear transcription factors show an even more biased distribution towards down-regulation (Wilcoxon test, non-paired data, p-value = 1.32e-14). In summary, TFs are enriched towards down-regulation independently of nuclear proteins, which are also enriched.

The enrichment of TF holds both for asexual and sexual development. However the number of TFs being down-regulated is greater for asexual vs sexual development being 100 and 84 respectively. Given this difference we wanted to know if all types of TF had a similar deregulation in both development programs (Figure 17).

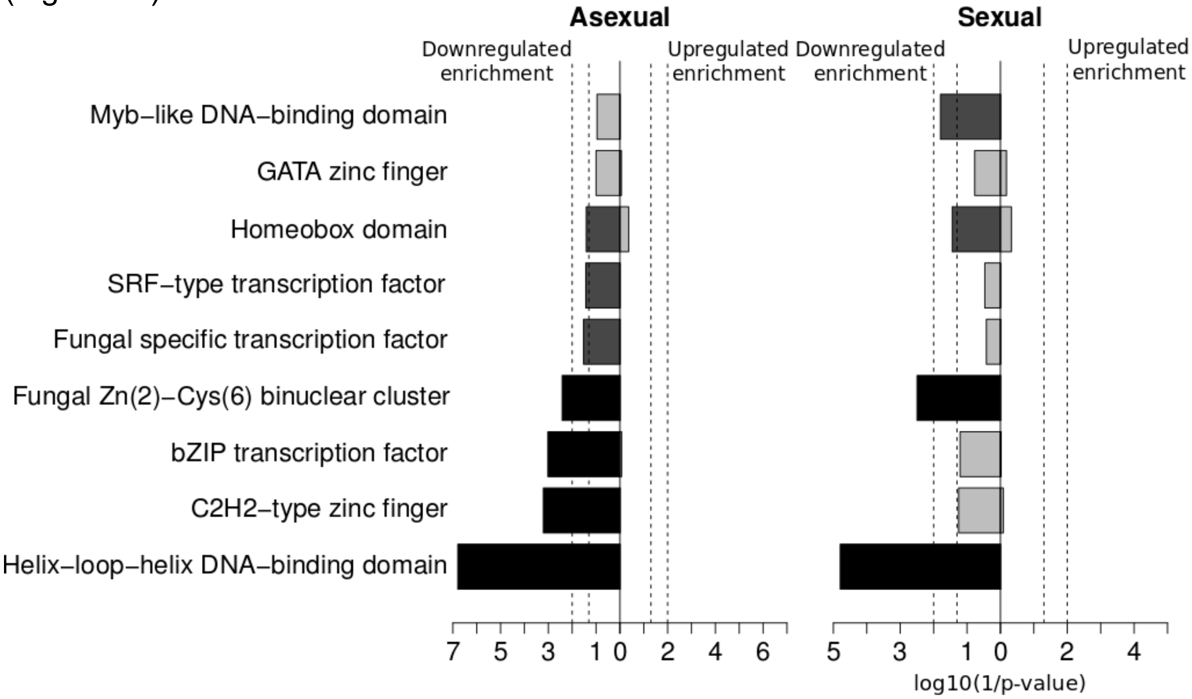


Figure 17. Similarities and differences in transcription factor families in asexual and sexual development of Rhizopus microsporus. The first and second dotted lines correspond to 0.05 and 0.01 p-value thresholds respectively. Bars passing 0.05 and 0.01 p-value thresholds are colored dark grey and black respectively.

In Figure 17, the bars represent enrichments of these domains in down-regulated and up-regulated genes in asexual (left) and sexual (right) development, the larger the bar the more significant the enrichment of that TF class. P-values are very small numbers ranging asymptotically from 1 to 0. To represent the enrichment as bars we transformed enrichment p-values dividing 1 by the p-value and calculating its logarithm (base 10). Several classes of DNA binding domains fall in the TF category: bZIP, fungal specific TF, fungal Zn(2)-Cys(6) binuclear cluster domain, GATA zinc finger, Homeobox and SRF-type TF. We considered enrichment information for all these classes.

The enrichment of TFs is always biased to down-regulation; however, there are similarities and differences among families in asexual and sexual development. Helix-loop-helix DNAbinding domain, fungal Zn(2)-Cys(6) binuclear cluster, homeobox domain and GATA zinc finger show a similar behavior regardless of the reproductive mode. On the other hand, more Myb TFs are repressed in sexual development. Additionally, SRF type, fungal specific, bZIP and C2H2-type zinc finger TFs are repressed in asexual sporulation (Figure 17). This result suggests that different types of TF may be used in the two reproductive modes of *R. microsporus*.

We further expored the number of common and uncommon TFs during mating and asexual development (Figure 18). Although the majority (57%) of down-regulated TFs are shared in both reproductive ways, 28% and 13% are exclusive of asexual and sexual programs respectively.

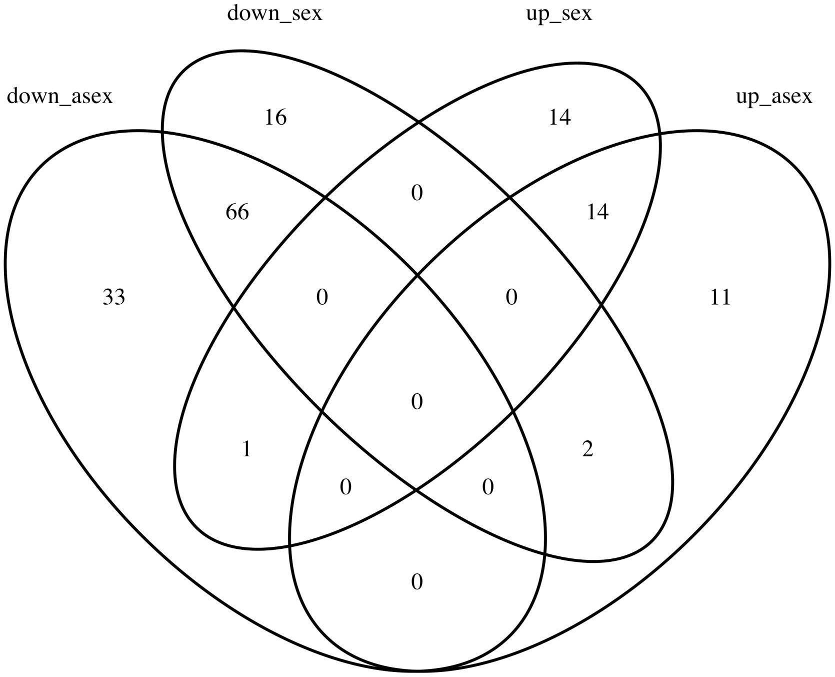


Figure 18. Shared and unshared transcription factors (TF) in sexual and asexual development. Down-asex represents down-regulated TFs in asexual development, down-sex represents down-regulated TFs in mating, up-asex represents up-regulated TFs in asexual development, and finally up-sex represents up-regulated TFs in mating.

Some of the down-regulated nuclear genes are annotated as chromosomal proteins. Among these we find a subunit of the origin recognition complex. This complex binds to origins of replication in eukaryotes and assists during the formation of the pre-replication complex. We also find six CHROMO domain-containing proteins (CHRomatin Organization MOdifier). CHROMO domains are present in proteins such as the chromatin remodelers polycomb, and trithorax, Heterochromatin protein 1 and SET domain containing protein SU(VAR)3-9 (Eissenberg 2001). This suggests that host chromatin remodelers may be down-regulated in presence of *B. rhizoxinica*.

We also find core histones H3, H4, H2A, H2B (Table 24) and linker histone H1 among nuclear down-regulated genes (Figure 19).

A manual exploration revealed that there are two histone cluster loci in ATCC52813. The first one is located in scaffold 7 and has histones H2B, H2A, H4, H3 (Figure 19, red). The second one is found in scaffold 8 with histones H4, H3 and two copies of H2A and H2B (Figure 19, blue). Together they harbor 10 of 14 histones present in ATCC52813. The enrichment of histones in down-regulated genes could be due to a co-regulation of these clusters. A smaller cluster composed of H3 and H4 (Figure 19, green) is up-regulated during sexual development.

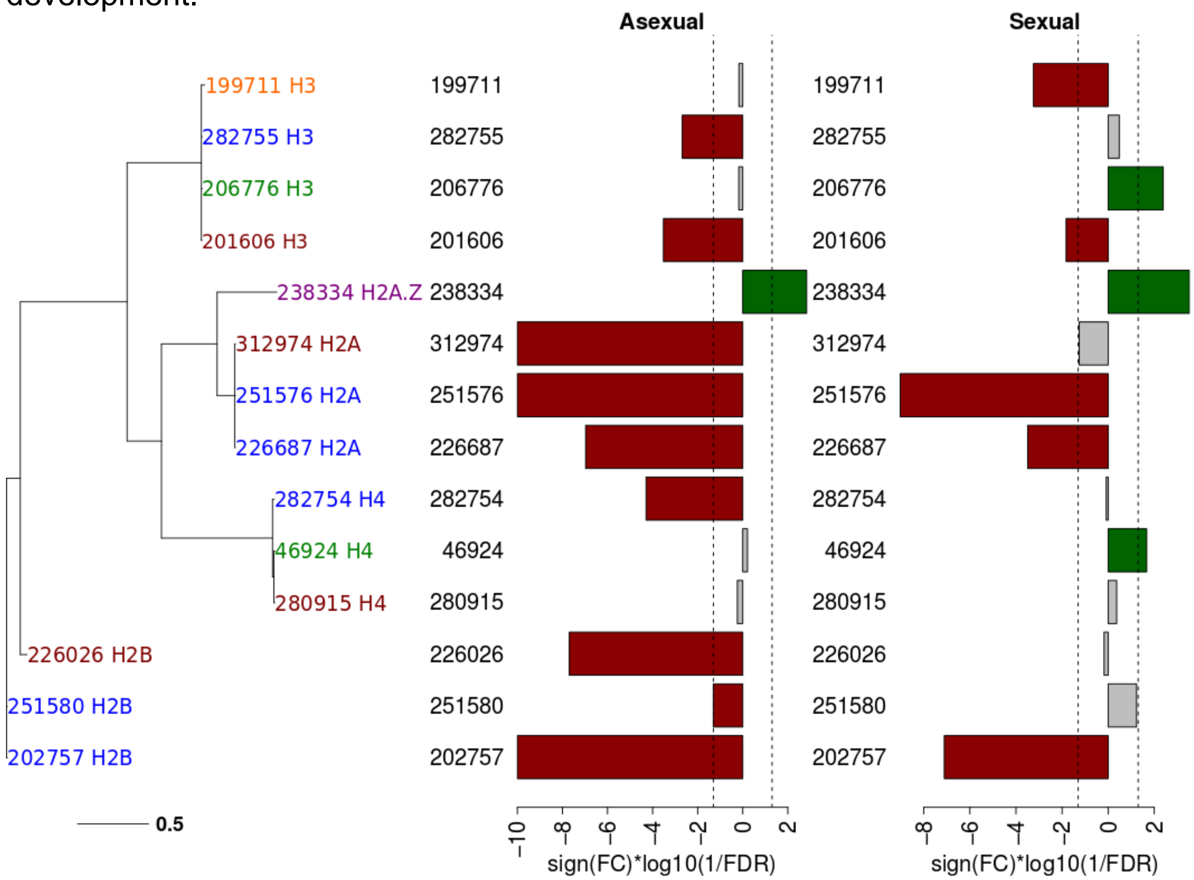


Figure 19. Histone bayesian phylogeny and differential expression (DE) among asexual and sexual development. DE data is truncated to 10 and -10 for up-regulation and down-regulation respectively. Positively and negatively fold changing genes are shown in dark green and dark red respectively. Vertical lines represent thresholds for up and down-regulation with an FDR of 0.05.

Only one histone is up-regulated in both sexual and asexual contrasts. This H2A (protein id 238334) histone was named H2A.Z based on the following findings:

1. Canonical H2A and the H2A.Z variants have a similarity of 65% (Biterge & Schneider 2014). H2A 238334 has only 65% identity and 79% similarity to the other H2A histones, the large branch of 238334 reflects these differences (Figure 19)

2. Canonical histones don’t have introns while H2A variants do, H2A 238334 has three introns (Biterge & Schneider 2014)

SET proteins are enriched in up-regulated genes contrary to general enrichment of nuclear proteins in down-regulated genes. This was discovered using a Pfam enrichment analysis (Table 23). Interestingly, SET enrichment holds for asexual (p-value 0.048) but not for sexual development (p-value 0.85). Apparently asexual development displays a higher diversity of *R. microsporus* SET proteins (Figure 20).

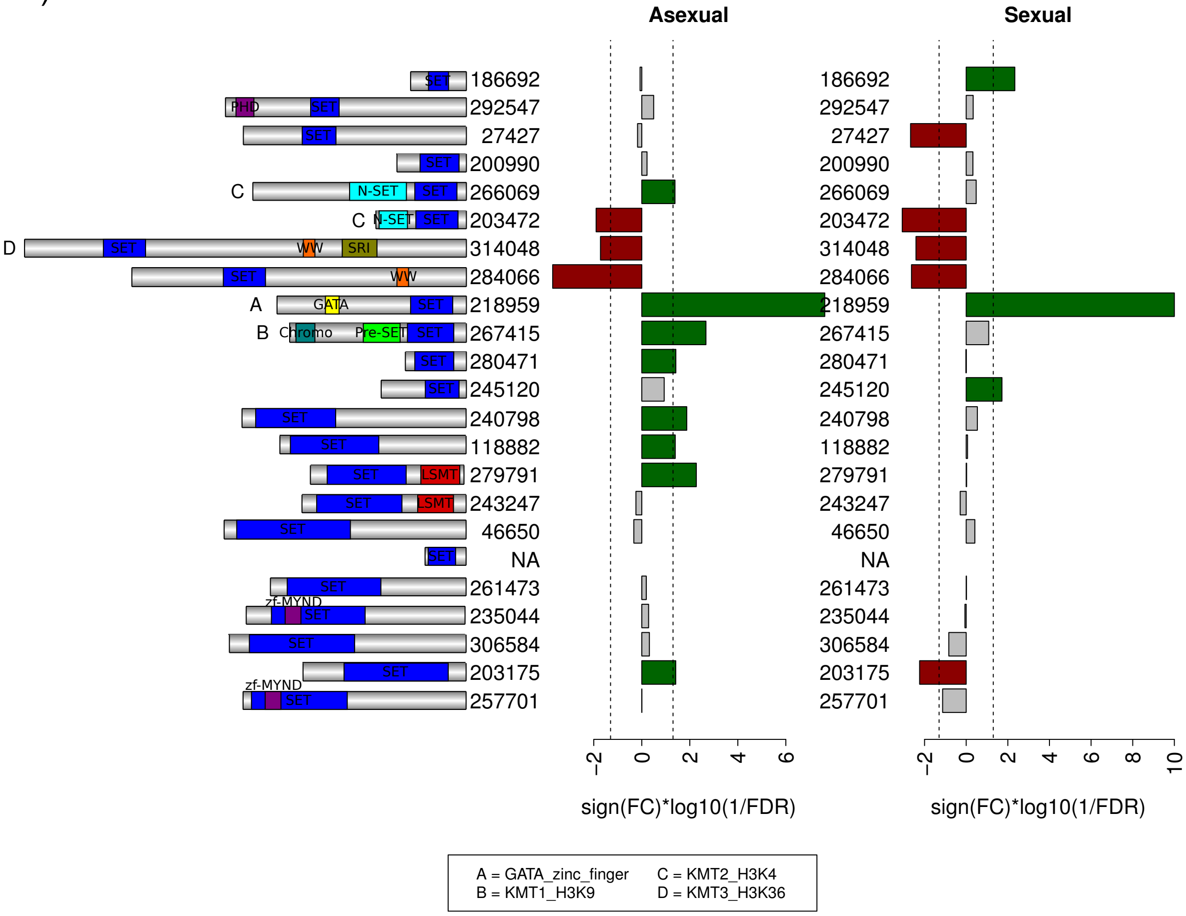


Figure 20. SET domain-containing protein architectures and differential expression (DE) among asexual and sexual development. DE data is truncated to 10 for up-regulation. Positively and negatively fold changing genes are shown in dark green and dark red respectively. Vertical lines represent thresholds for up and down-regulation with an FDR of 0.05.

We assigned probable functions for SET proteins based on sequence similarity and shared domain architecture (Table 3) with characterized homologs. The protein with Id 267415 is a H3K9 methyl-transferase and is up-regulated in asexual but not sexual development. We couldn’t predict a substrate for the majority of *R. microsporus* SET proteins. A H3K36 methyl-transferase (protein Id 314048) is down-regulated in both sexual and asexual development. The protein 203472 is a H3K4 methyl-transferase that is consistently down-regulated in both reproductive modes, while another H3K4 methyl-transferase, 266069 is up-regulated in asexual sporulation (Figure 20). Generally H3K4me3 and H3K36me1 are associated with active transcription, while H3K9me1 is correlated with transcriptional repression.

We found Rubisco methyl transferases in *R. microsporus*. These proteins are characterized by having the SET domain and Rubisco (large subunit methyltransferase) LSMT substrate-binding domain (PF09273).

A protein with a GATA zinc finger and a SET domain (218959) is up-regulated in both reproductive modes and has a very small p-value. We found no proteins having the same architecture in the Pfam domain database or in any of the selected Dykarial genome. But when we looked for them in our annotated genomes, we found that GATA-SET proteins are present in all *Rhizopus microsporus* members including both host and non-host strains (Table 25). Interestingly, all host strains have only one GATA-SET protein while the two non-host strains have two. This observation raises the possibility that endosymbiont dependence relies on the absence of one of these GATA-SET proteins.

Table 25. GATA zinc finger - SET is a new domain architecture exclusive of R. microsporus species

	Dykaria						Mucorales		Non-host *Rhizopus*			Host strains		
	An	Ta	Nc	Sc	Um	Cn	Pb	Mc	Ro	Rm 21	Rm 59	Rm 17	Rm 14	Rm 13
GATA +SET	0	0	0	0	0	0	0	0	0	2	2	1	1	1
N-SET +SET +SET assoc	1	1	1	1	1	0	0	0	0	0	0	0	1	1
N-SET +SET	0	0	0	0	0	1	0	0	2	2	2	1	1	1
SET +chromo +Pre-SET	0	0	0	0	0	0	0	0	1	2	0	1	1	1
SET +SRI +WW	0	0	0	0	0	0	0	0	2	1	2	1	1	1

In summary, GATA-SET architecture represents a new domain architecture that is exclusive of the *R. microsporus* clade. Host strains lack an additional copy that non-host strains have, and the only representative in Rm13 and Rm14 are up-regulated in presence of *B. rhizoxinica*. All of these observations make 215989 GATA-SET an excellent gene candidate for further explorations.

It is interesting that host SET proteins are being up-regulated in presence of *B. rhizoxinica* because this endosymbiont codes for its own SET protein (BrSET). BrSET has been shown to methylate histones *in vitro* (Baruch, Brieba-Castro & Partida-Martínez, unpublished results). BrSET has sequence similarity to

Burkholderia thailandensis SET protein (80% identity), Therefore we expect them to have similar substrate specificity and activity. BtSET performs mono- and di-methylations in H3K4 (Li et al. 2013).

According to our analysis, BrSET is expressed inside its host along with a type III secretion system (T3SS) (Figure 21). We propose that BrSET is secreted to *R. microsporus* cytoplasm through T3SS. This proposal is based on the finding that *Burkholderia thailandesis* secretes a SET protein (BtSET) in invading hosts via a T3SS (Li et al. 2013).

R. microsporus ATCC52813 has two HP1 proteins which satisfy the criterion of having Chromo and Chromo shadow domains. One of these proteins is down-regulated in presence of *B. rhizoxinica* and the other one is not differentially expressed. BrSET could interact with these proteins in a similar way as BtSET does in cell lines.

BtSET could potentially up-regulate some *R. microsporus* genes through H3K4me and H3K4me2 modifications.

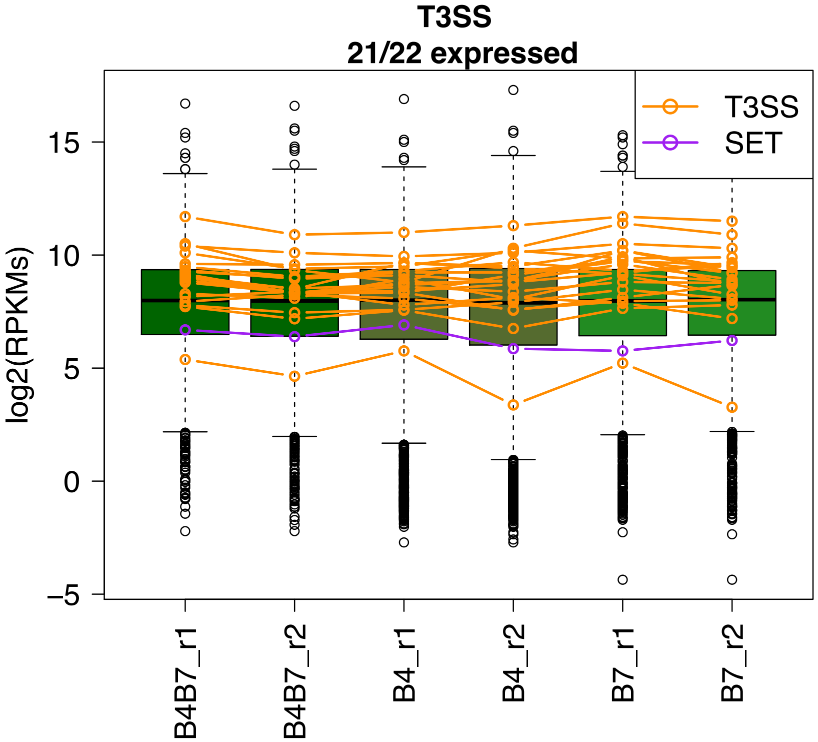


Figure 21. B. rhizoxinica SET protein (purple) and a type III secretion system (orange) are expressed inside R. microsporus

We envision that both BrSET and *R. microsporus* SET proteins are present simultaneously in host nucleus (Figure 22). Less TF and Nuclei genes are found in presence of the endosymbiont.

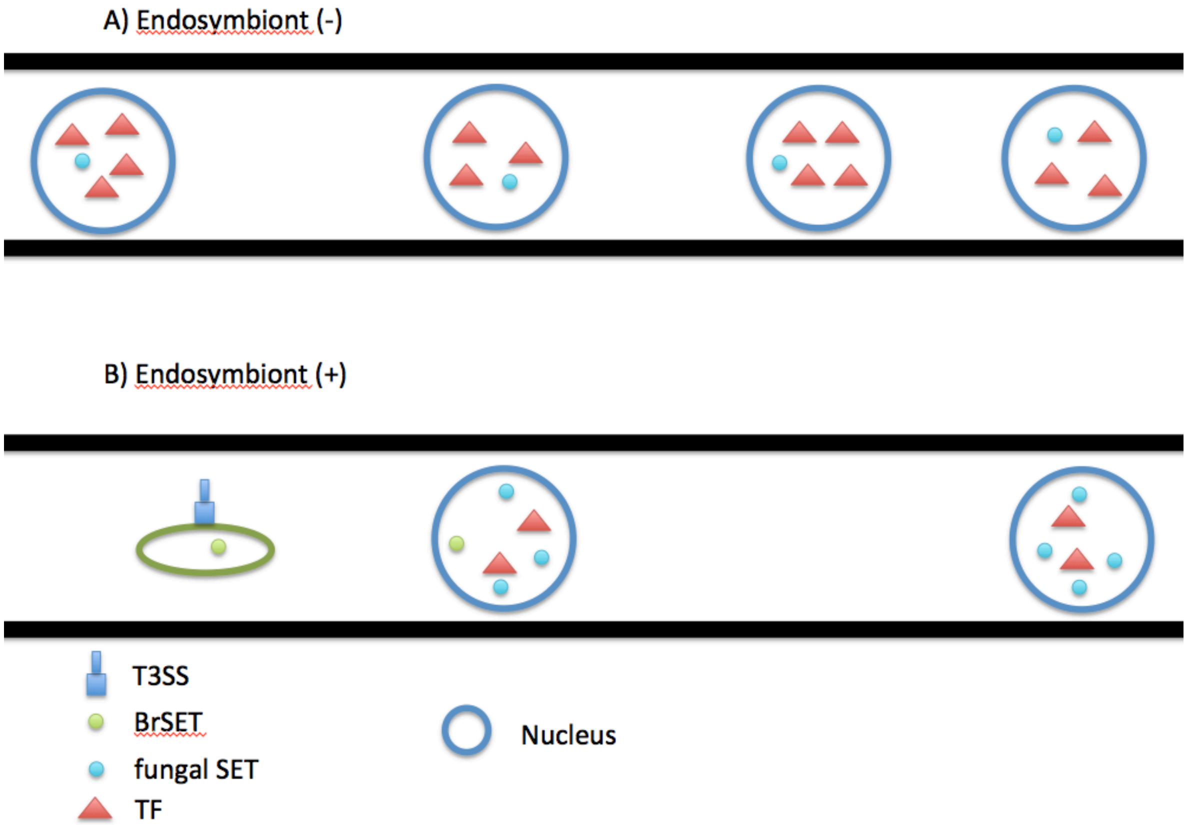


Figure 22. Model of R. microsporus nuclei in presence and absence of B. rhizoxinica

Chapter IV: Results of fungal comparative genomics & expression

Mucorales phylogeny

To get insights into the evolutionary relationships of the *Rhizopus* genus, we built a Mucoralean phylogeny. We selected a total of 46 protein families that are present in all Mucoralean genomes with only a single member based on sequence similarity (see Methods). Among these 46 proteins we find the L27 ribosomal protein, translation initiation factor eIF2A, a subunit of the origin recognition complex for DNA replication, a Myb-like transcription factor, chorismate synthase, among others. Annotations for these proteins are available in the digital supplementary material (Supporting_material/2.Fungal_comparative_genomics/Mucorales_core_family_proteins).

We built a phylogeny of Mucorales with bayesian inference (Figure 23). *Phycomyces* is the most distant Mucoralean genome analyzed, for this reason we manually rooted our phylogeny with *Phycomyces* as the outgroup.

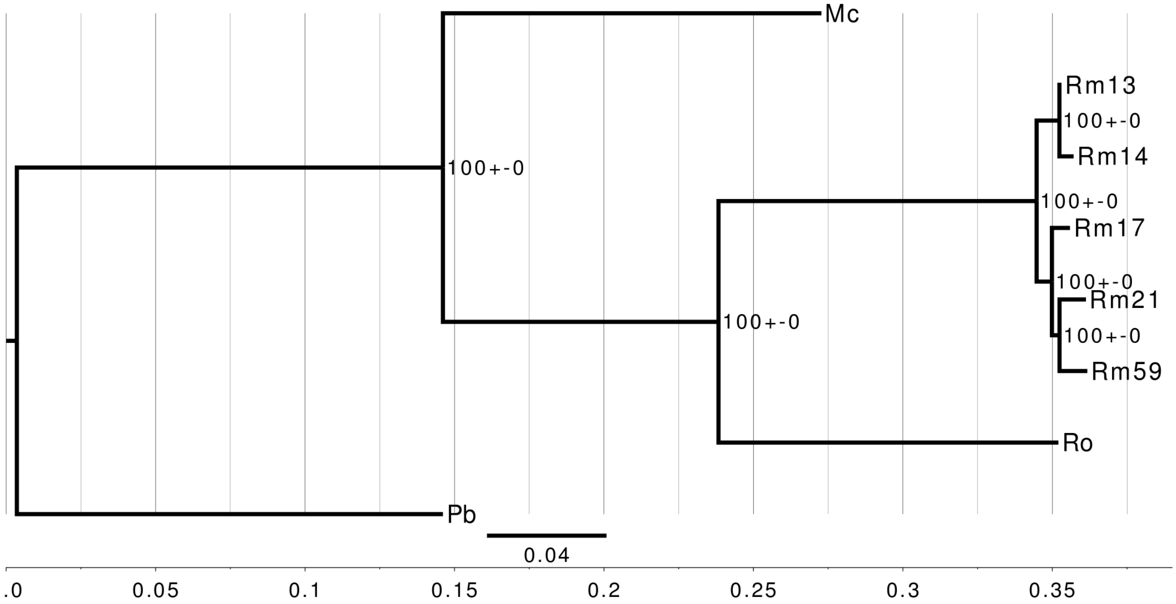


Figure 23. Mucoralean bayesian phylogeny, made with 46 conserved single member protein families. Mc stands for Mucor circinelloides, Rm stands for R. microsporus, Ro stands for Rhizopus oryzae and Pb stands for Phycomyces blakeesleanus

According to our phylogeny *R. oryzae* is a clear outgroup to all *Rhizopus microsporus*. This is consistent with *R. oryzae* being a different species. In the *R. microsporus* clade we have two groups, the first one coincides with var microsporus containing Rm13 and Rm14, the second one with var chinensis, containing Rm17, Rm59 and Rm21. The chinensis clade has larger genomes than microsporus (Table 12). According to our phylogeny Rm21 and Rm59 are closer to each other than Rm17 (Figure 23).

The host trait is present in the var. microsporus and var. chinensis clades. This result suggests that host capabilities are paraphyletic rather than monophyletic. This means that the host trait is not found in a single clade, but is dispersed in var. microsporus and var. chinensis clades. Another possible scenario is that the ancestor of all Rhizopus microsporus was symbiotic but then the clade that gave rise to Rm21 and Rm59 lost the symbiotic trait.

Rhizopus genome description

We wanted to compare the genomes of *Rhizopus* strains that naturally harbor endosymbiotic bacteria to those that don’t, in order to discover differences that correlated with this behavior. For this purpose we selected the genomes shown in Table 12. We included available host *Rhizopus* genomes, these are *Rhizopus microsporus* var microsporus ATCC52813, *R. microsporus* var microsporus ATCC52814 and *R. microsporus* (*Rhizopus chinensis* Rh-2) ATCC62417 (Horn et al. 2015). Non-host *Rhizopus* genomes include *R. microsporus* var chinensis ATCC11559 (Horn et al. 2015), *R. microsporus* var chinensis CCTCC M201021 and *R. oryzae* 99-880.

R. microsporus ATCC11559 and *R. oryzae* 99-880 are validated as non-host (Partida-Martinez, PhD dissertation, 2007). Since all known host strains produce the Rhizoxin toxin (Partida-Martinez & Hertweck 2005), they are considered highly toxigenic, making them unsuitable for traditional food fermentation. We consider that *R. microsporus* var chinensis CCTCC M201021 is a non-host strain because it was isolated from Daqu, a traditional fermentation starter (Wang et al. 2013). Rhizoxin is produced in fermentation conditions for sufu or tempe at concentrations that can be highly toxic for humans (Rohm et al. 2010). If CCTCC M201021 were a host strain it would likely produce Rhizoxin, resulting in toxic beverages.

All the *Rhizopus* genomes that we have access to are drafts and are thus assembled to scaffolds and not chromosomes. They also differ substantially in assembly quality, as judged by their N_{50} values and number of scaffolds (Table 12) N_{50} is used as a measure of assembly quality; N_{50} value sets a lower limit for scaffold sizes that contain half of the total bases of the assembly. A greater N_{50} value represents a better assembly as more bases are found in bigger scaffolds. The best *Rhizopus* assembly is *R. oryzae* with just 81 scaffolds and a N_{50} of 310 Kb. On the other hand, the worst assembly is represented by *R. microsporus* var chinensis CCTCC M201021 with a N_{50} of 30 Kb and over 3,000 scaffolds. It is worth noting that *Phycomyces* and *Mucor* genomes have better assemblies than any *Rhizopus* genome, with N_{50} values of 1,515 and 4,318 Kb respectively.

Genome size varies almost two-fold among currently sequenced *Rhizopus* genomes. We have two main categories regarding assembly size: reduced genomes with a size around 25 Mb and bigger genomes with sizes around 48 Mb. The two genomes from the Joint Genome Institute have a genome size similar to 25 Mb, all other *Rhizopus* genomes sizes are around 48 Mb regardless of where they were sequenced. Differences in genome size are also reflected in protein numbers (Figure 24). In this barplot, we also show the number of protein families according to sequence similarity (see Methods).

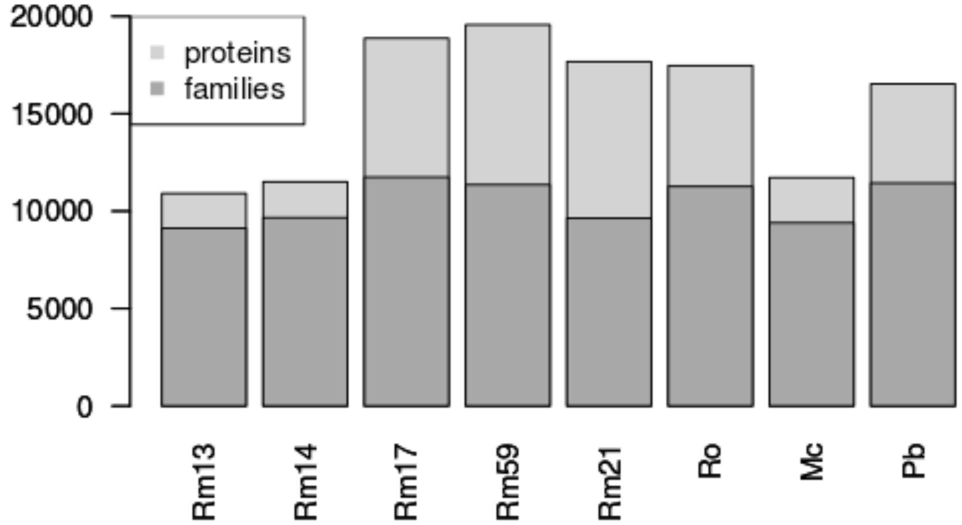


Figure 24. Number of proteins and families in Mucoralean genomes. Mucoralean abbreviations are mentioned in Table 12

The number of proteins varies much more than the number of families in Mucoralean genomes. The core genome represents the genes shared by a group of organisms. The difference between the number of proteins and families is related to the number of paralogs in each genome. The number of paralogs is greater for the var chinensis clade than the var microsporus.

In terms of protein numbers, there are two types of *Rhizopus microsporus* genomes, those that have around 11,000 proteins, such as Rm13 and Rm14 (var microsporus); and those that have between 17,600 and 19,500 proteins, such as Rm17, Rm59 and Rm21 (var chinensis). Interestingly, host *Rhizopus* strains are found with both small and big genomes.

Protein, family numbers and genome sizes correlate with the *R. microsporus* groups found in our phylogeny. The ancestor *R. oryzae* has a genome size similar to the var chinensis clade, this suggest that the var microsporus clade could have suffered a genome reduction. However host strains are found in both clades, therefore the host trait seems to be independent of differences in genome size and protein numbers.

Protein domains comparison: *Rhizopus* host vs non-host

To gain insights into what makes host *Rhizopus microsporus* strains suited for symbiosis we compared their genomes with those of non-host strains.

This analysis involves the following genomes:

Rm13 & Rm14 & Rm17 vs Rm59 & Rm21

We compared protein domain counts between these two sets. We previously compared protein families according to sequence similarity, however our most significant differences lacked annotations and we couldn't say much about them. We then decided to compare protein domains because they are easier to link with a function, we ignored those Pfam domain families without an assigned function DUF (Domains of Unkown Function). To compare the domain counts we used R package edgeR (See methods section). *R. microsporus* strains vary in genome size (Figure 24), edgeR corrects for these differences just like it normalizes different RNA-Seq library sizes. We tested 3610 Pfam domains. The resulting domains of this comparison are show in Table 26.

Table 26. Top differences in protein domains among host and non-host Rhizopus microsporus strains

Protein domain (Pfam)	logFC	logCPM	P-value	FDR
PIF1-like helicase	-1.6	9.38	6.94e-05	0.25
Transposase	-1.3	11.4	0.00015	0.27
Autophagy protein Apg5	4.6	7.62	0.0002	0.29
Helix-turn-helix domain	-1	10.3	0.0003	0.29
Homeodomain-like domain	-1	9.79	0.0007	0.56
Nuclear pore localisation protein NPL4	3.8	7.33	0.006	1

Losses in host strains: PIF1 helicases

PIF1-like helicase domains have the best P-value (6.94e-05) among all domains tested. However, due to the number of Pfams tested we have high false discovery rates (FDR), therefore we cannot rule out that this result was given by chance. Nevertheless PIF1-like helicases have an interesting phylogenetic distribution (Figure 25). These domains are abundant in the *Rhizopus* genus, with up to 32 proteins in *Rhizopus oryzae*. However they display smaller numbers in host strains.

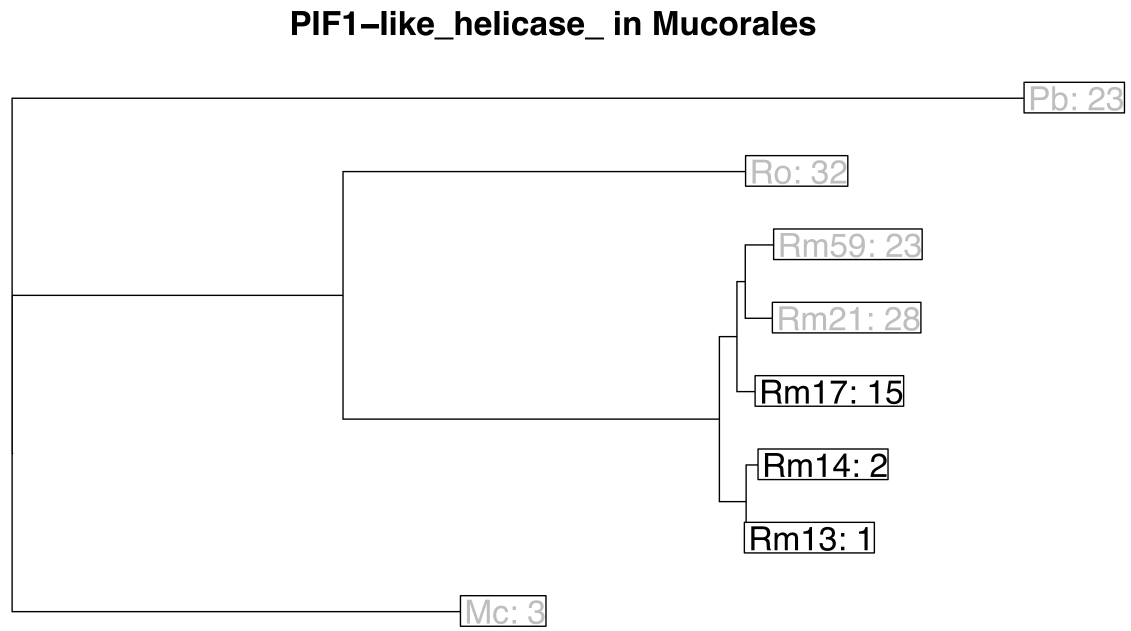


Figure 25. PIF1-like helicases in Mucorales. Host strains and non-host strains are shown in black and grey respectively.

A possible link of Pif1 helicases and the *Rhizopus-Burkholderia* symbiosis is not obvious.

Protein domains comparison: Mucorales vs Dykaria

We noticed that Mucoralean genomes have more Ras, Chitin synthases and SET domain containing proteins than some Dykaria genomes. These punctual observations motivated us to make a broader comparison between Mucorales and Dykaria protein domain contents. The chosen Dykaria representatives are mentioned in Methods.

We compared protein domain counts, the same approach used for the *R. microsporus* host and non-host comparison.

Reductions in Mucorales

Reduced domains in Mucorales relative to Dykaria are shown in Table 27. Full tables are available in the digital supplementary material (Supporting_material/2.Fungal_comparative_genomics/mucorales_vs_dykaria).
For each Pfam domain we also added information on whether it is enriched in up-regulated or down-regulated genes due to endosymbiont presence (Chapter III). We added this information with the aim of integrating the comparative genomics and RNA-Seq analyses.

Table 27. Selected reductions in Mucorales

Pfam domain	FDR	Enriched in
Fungal specific transcription factor domain	4.81e-68	Down-regulation
Fungal Zn(2)-Cys(6) binuclear cluster domain	2.48e-39	Down-regulation
KR domain	3.59e-23	Up-regulation
Beta-ketoacyl synthase, C-terminal domain	7.91e-11	Neither
Condensation domain	2.31e-09	Neither
Cellulase	5.24e-06	Neither
Fungal cellulose binding domain	1.78e-07	Neither

Natural products synthesis clusters

The condensation domain (PF00668) is under-represented in Mucorales relative to Dykaria (FDR 2.3e-09). This domain is characteristic of Non-ribosomal peptide synthases (NRPS) (Keller et al. 2005). In a similar way Mucorales have less keto reductase and Beta-ketoacyl synthase domains than Dykarial genomes (Table 27). These domains are found in polyketide synthases (PKS) (Keller et al. 2005). Keto reductase domains are enriched in up-regulated genes due to endosymbiont presence. These up-regulated keto reductases can’t be PKS, as they lack PKS essential domains such as ketoacyl CoA synthase, acyltransferase and acyl carrier protein domains. The keto reductases expressed by *R. microsporus* are likely to be involved in fatty acid synthesis, a process that has similarities with polyketide synthesis.

Expansions in Mucorales

Expanded domains in Mucorales relative to Dykaria are shown in Table 28. Again, for each Pfam domain we added information on whether it is enriched in up-regulated or down-regulated genes due to endosymbiont presence (Table 28, third column).

Table 28. Selected expansions in Mucorales

Pfam domain	FDR	Enriched in
Chromo (CHRromatin Organisation MOdifier) domain	3.68e-15	-
GATA zinc finger	9.07e-12	-
RasGEF domain	6.33e-10	Down-regulation
Polysaccharide deacetylase	3.02e-10	Down-regulation
Ricin-type beta-trefoil lectin domain	2.76e-08	Down-regulation
Chitin synthase	8.3e-08	Down-regulation
PIF1-like helicase	1.94e-07	-
Leucine rich repeat (LRR)	1.03e-05	-
Ras family	0.0102	Up-regulation
SET domain	0.357	Up-regulation
LysM domain	0.904	Down-regulation

We found differences in TF distribution in Mucorales relative to selected model dykarial genomes. Dykaria have more Fungal Zn(2)-Cys(6) binuclear cluster domain (FDR 2e-39) and Mucorales have more GATA zinc finger (FDR 9e-12) and bZIP transcription factors (FDR 1e-06). There is a general trend for TF to be down-regulated in presence of *B. rhizoxinica*.

Ras family proteins and Ras regulatory proteins were already reported to be expanded in *R. orzyae* (Ma et al. 2009).

Putative bacterial sensing receptors

Mucorales display expansion of Leucine rich repeats and Ricin B lectins domains relative to Dykaria. These domains represent good candidates for bacterial sensing through interactions with microbe-associated molecules. Most *R. microsporus* ricin B lectins are down-regulated in presence of *B. rhizoxinica* (FDR 0.022, Table 24).

LysM domains are used by plants to sense symbiotic bacteria (Gust et al. 2012). LysM domain numbers are no different among Mucorales and Dykaria (Table 28). But interestingly, the majority of *R. microsporus* LysM domains are down-regulated when *B. rhizoxinica* is present (FDR 0.022, Table 24).

These result one of two scenarios: *B. rhizoxinica* down-regulates the expression of its host bacterial sensing receptors or alternatively, the host fungus down-regulates its bacterial sensing receptors to allow bacteria to inhabit the fungal cytoplasm.

Cell wall & compartmentalization proteins

Chitin synthases

It was already reported that Mucoralean genomes have more chitin synthases than dykarial genomes (Ruiz-Herrera & Ortiz-Castellanos 2010). We also find chitin synthases among expanded domains. Many chitin synthases are being down-regulated in endosymbiont presence; this suggests that chitin synthesis is more active in absence of *B. rhizoxinica* (FDR 6.3E-4, Table 24)

Chitin deacetylases

Chitosan is a major component of mucoralean cell walls and is produced by chitin deacetylase enzymes. Chitin deacetylases are part of the polysaccharide deacetylase Pfam model (PF01522). Polysaccharide deacetylase domains are expanded in Mucorales relative to Dykaria (FDR 3e-10).

The majority of the polysaccharide deacetylases tend to be down-regulated. However, a clade annotated to have chitin deacetylase activity displays a trend to be up-regulated, this effect is clearer in sexual development (Figure 26). This suggests that sporangia and zygospore formation require higher amounts of chitosan that hyphae.

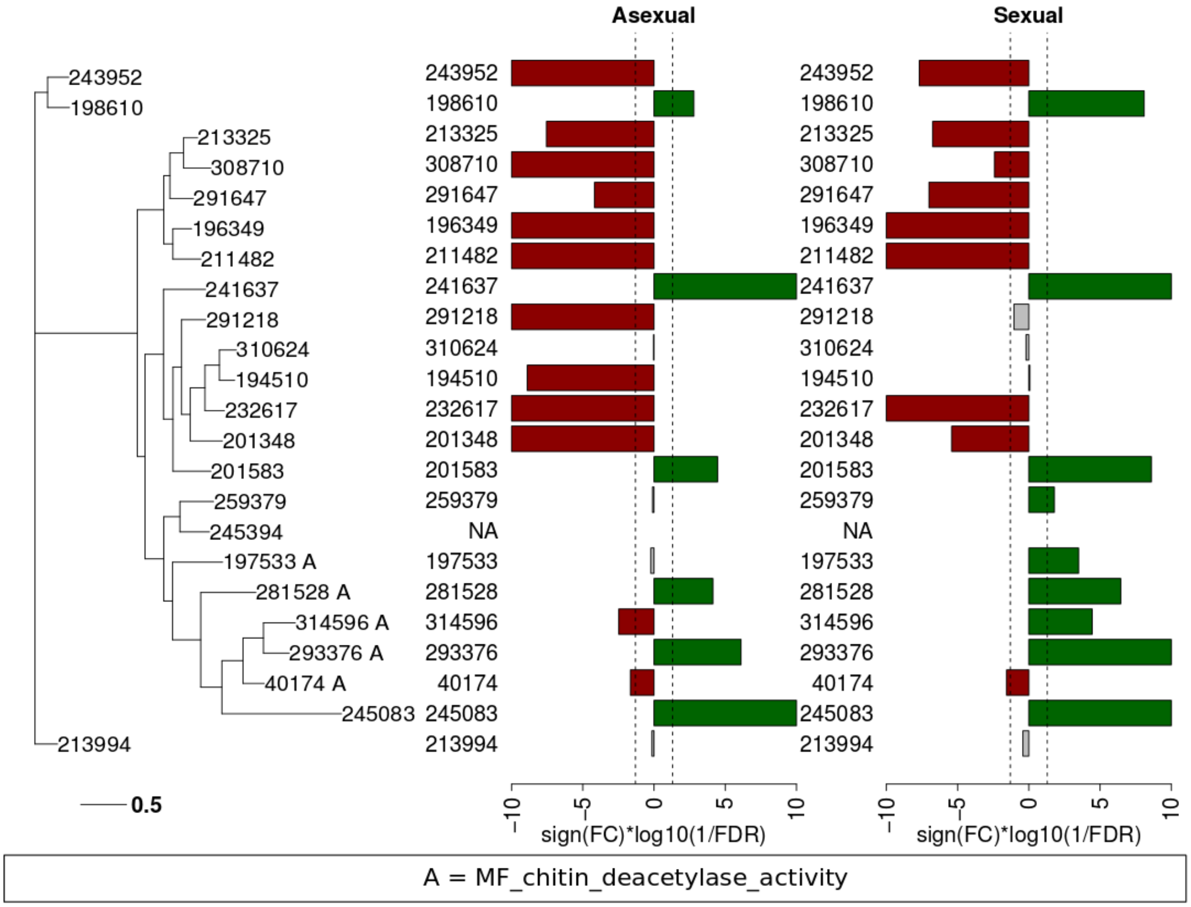


Figure 26. Polysaccharide deacetylases bayesian phylogeny and differential expression (DE) among asexual and sexual development. DE data is truncated to 10 and -10 for up-regulation and down-regulation respectively. Genes with positive and negative fold changes are shown in dark green and dark red respectively. Vertical lines represent thresholds for up and down-regulation with an FDR of 0.05

Septins

Septins are filament-forming GTP-binding proteins found primarily in fungal and animal cells. In fungi they are typically associated with septa, a partition dividing filamentous hyphae. They have also been found in hyphal tips and in septum of dividing yeast cells. In general they provide a certain degree of compartmentalization and are expressed in morphological transition boundaries.

Despite being considered coenocytic (non-septated), Mucorales have many septin genes. In fact, they have more of septin domain-containing proteins than those present in model ascomycetes and basidiomycetes (Table 28).

Septins are over-represented in up-regulated genes due to *B. rhizoxinica* presence. This over-representation holds in both sexual (p-value 1.7E-5) and asexual (p-value 4.6E-4) development. The up-regulatation of septins in both reproductive phases of *Rhizopus* suggests a need for compartmentalization during sporangia and zygospore formation. Septins may restrict membrane-associated proteins to specific regions acting as diffusion barriers.

Septins are frequently present in the morphological boundaries in several fungal species such as *Candida albicans*, *Aspergillus fumigatus*, *Ustilago mayidis* and *Saccharomyces cerevisiae* (Bridges & Gladfelter 2014). Based on the knowledge from these species, we propose some places for septin localization in *Rhizopus microsporus*: the hyphal tips, and the interphase between columnella and the spore formation section (Figure 27). Finally, septa are occasionally formed in *Rhizopus* hyphae; this is also a candidate spot for septin localization. During zygospore formation, we suspect that septins could also be located in the delimiting zone of the newly developing zygospore.

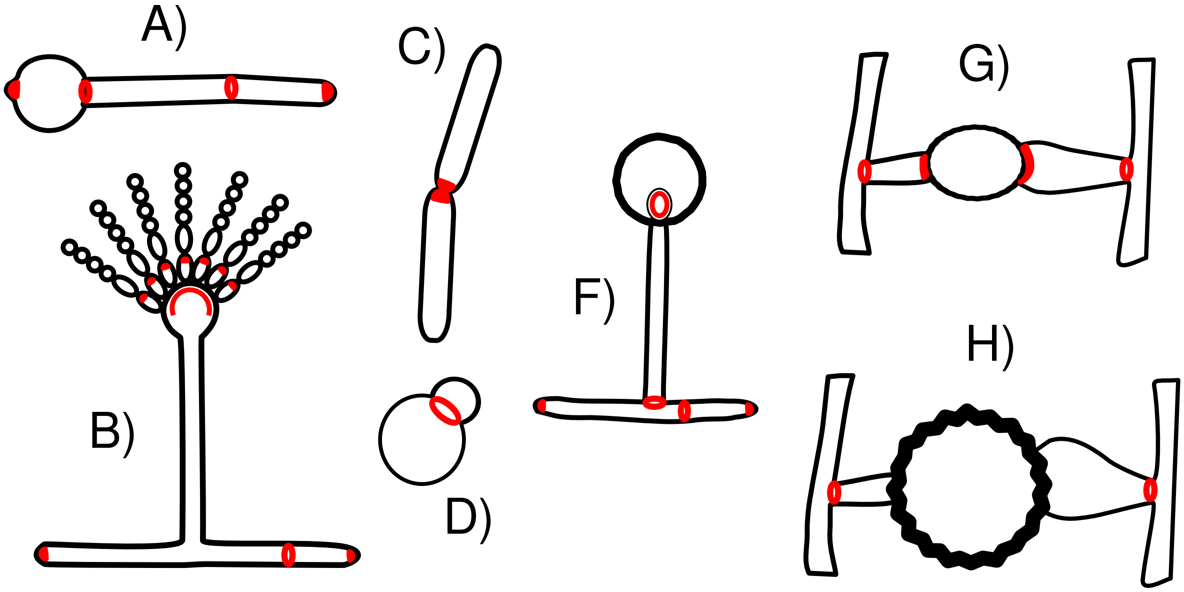


Figure 27. Proposed sites for septin localization in R. microsporus based on knowledge from other fungi. A) Candida albicans, B) Aspergillus fumigatus, C) Ustilago maydis, D) Saccharomyces cerevisiae, F) Rhizopus microsporus sporangium, G) Two Rhizopus microsporus mating, H) Zygospore. Septin localization is represented by red color

Chapter V: Results of *B. rhizoxinica* comparative genomics & expression inside host

Aims

There are two goals for our *Burkholderia* comparative genomics:

- To propose candidate genes relevant for *B. rhizoxinica* - *R. microsporus.* (Main goal)
- To determine the closest free-living relative of *B. rhizoxinica* (Secondary goal)

Burkholderia phylogeny and genomic features

A goal for our comparative analysis was to build a phylogeny with *Burkholderia* genomic data. A phylogeny of the *Burkholderia* genus will help us in two different ways: 1) to determine the closest free-living relative of *Burkholderia rhizoxinica*, and 2) to analyze the distribution of proteins, domains and functions under an evolutionary framework.

To build the *Burkholderia* phylogeny we first identified the core proteins of all analyzed genomes. This consisted of performing protein sequence comparisons searching for best bidirectional hit groups using the BBH-Star software. Using this method we found a core of 541 proteins shared in all 44 genomes. We then aligned individual ortho-groups, filtered poorly aligned sites, concatenated all alignments into a single one and built a phylogeny with a bayesian approach. We rooted our phylogeny manually with *Ralstonia* genomes as the out-group. For a more detailed description see the methods section.

We combined our resulting phylogeny with protein families according to sequence similarity (see Methods), protein numbers and GC content for each genome (Figure 27).

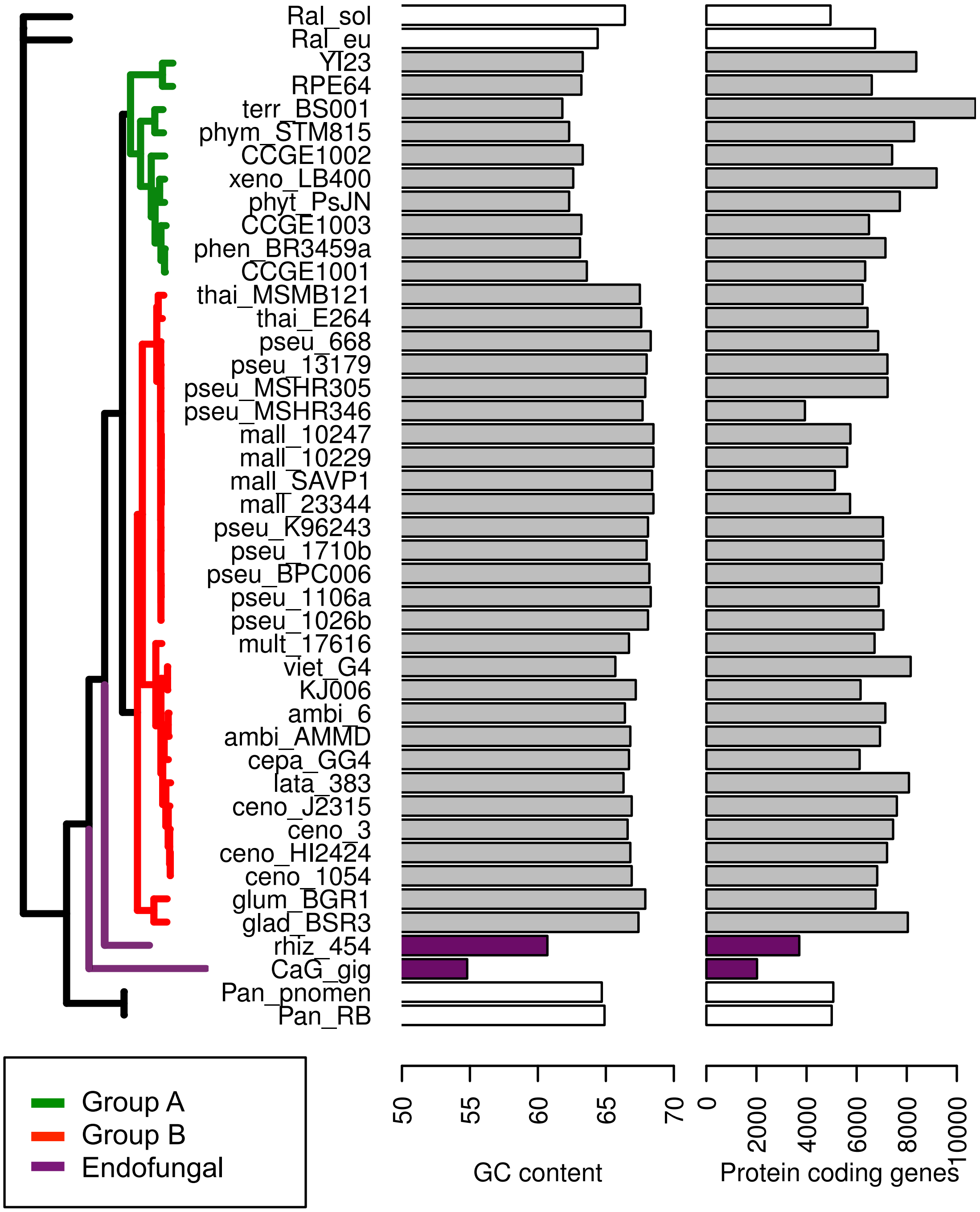


Figure 28. Phylogeny and genomic features of the Burkholderia genus. Short names were defined for each genome (see Methods). Bars represent GC content and number of total proteins. Endofungal bacteria GC and protein numbers are highlighted in purple. Free-living Burkholderia spp. information in barplots is shown in grey, outgroup bars are shown in white.

We found two main clades for the *Burkholderia* genus. Group A (Figure 27, green) includes plant-associated and environmental strains. Group B (Figure 27, red) includes mainly human and plant pathogens. Both groups are consistent with previous reports (Estrada-de los Santos et al. 2013).

According to our phylogeny *B. rhizoxinica* and *Ca.* G. gigasporarum are outgroups to the *Burkholderia* genus. The immediate interpretation is that both fungal endosymbionts have a distant common ancestor to all other members of the *Burkholderia* genus. However, this interpretation should be taken cautiously as endosymbiont have genome traits that may result in phylogenetic reconstruction artifacts. Factors such as heterogeneity of nucleotide compositions among species, rate variation across lineages and within-site rate variation contribute to systematic errors in phylogenetic reconstruction (Jeffroy et al. 2006). These sources of error don't disappear if more data is added and can result in statistically supported, but wrong phylogenomic trees. It is worth noting that our phylogeny nodes are statistically supported.

In our phylogeny, *B. rhizoxinica* and *Ca.* G. gigasporarum display long branches (Figure 27), suggesting that they also have fast-evolving rates. Some authors recommend the removal of odd species, for example fast-evolving microsporidia should be never be used to represent fungi (Jeffroy et al. 2006). Unfortunately, *B. rhizoxinica* is the main reason why we constructed the phylogeny and we cannot exclude it from the analysis.

B. rhizoxinica and *Ca.* G. gigasporarum have typical characteristics of endosymbiont genomes. Intimate endosymbionts tend to loose genes because its host provides nutrients and a stable environment. As a result of this friendly environment endosymbionts may lose complete metabolic pathways and functions (McCutcheon & Moran 2011). *Ca.* G. gigasporarum has the smallest genome with only 2,000 protein-coding genes while *B. rhizoxinica* has 3,780 protein-coding genes. The mean number of proteins for all other *Burkholderia* genomes is approximately 7,000, almost twice the number of proteins present in *B. rhizoxinica*.

Genomes affected by a reduction tend to have lower GC content. *Ca.* G. gigasporarum has the lowest GC content (54.8%), followed by *B. rhizoxinica* HKI 454 with a GC content of 60.7%. We detect a difference in GC content of members of groups A and B. Group A members have a mean GC of 62.9% while members of the B group have a mean GC of 67.4%. The difference in GC content between groups A and B was reported previously (Estrada-de los Santos et al. 2013).

As *B. rhizoxinica* doesn't group clearly with any other *Burkholderia* member we cannot establish its closest free-living relative. According to Estrada-de los Santos and collaborators, *B. rhizoxinica* belongs to group A of *Burkholderia*, which includes plant mutualistic and environmental strains. Based on a multilocus sequence analysis the apparent closest free-living relative is *B. kururiensis*. Five genes were used in this analysis: *atpD*, *gltB*, *lepA*, *recA* and 16S rRNA. These genes were aligned and concatenated to build a phylogeny (Estrada-de los Santos et al. 2013).

Conclusions from our phylogeny:

- *B. rhizoxinica* displays typical characteristics of a genome being reduced.
- With our model it is not feasible to identify the closest free-living relative of *Burkholderia rhizoxinica*
- Reconstructing the evolutionary history of endosymbionts is particularly challenging.

We next compared the functions, subsystems and protein domain distribution of *B. rhizoxinica* with those of free-living *Burkholderia* genus members in order to propose genes relevant to symbiosis.

Expanded functions in *B. rhizoxinica* vs free-living relatives

B. rhizoxinica is clearly suffering a genome reduction (Figure 29), therefore we would expect a general trend to loose genes and functions.

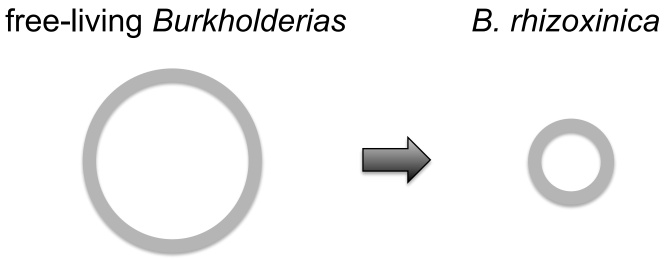


Figure 29. Cartoon of genome reduction in B. rhizoxinica. Circle size represents genome size.

We would expect that those genes kept by *B. rhizoxinica* are important for its survival. In a more extreme scenario, if a function is over represented in an endosymbiont with respect to free-living relatives, it may represent an adaptation to the host environment (Figure 30).

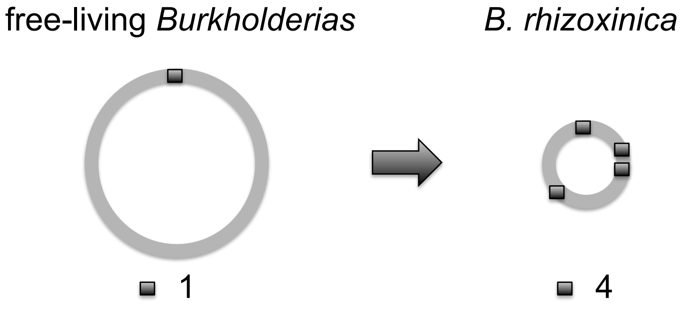


Figure 30. Cartoon of expanded functions in B. rhizoxinica. Circle size represents genome size. Black squares represent genes belonging to an expanded functional category

We looked for expanded functions in *B. rhizoxinica* relative to free-living *Burkholderia*. To this end we used RAST proposed gene functions, subsystems and our Pfam annotations to build functional category distribution matrices. Finally, we compared their distribution between organisms with edgeR (for more details see methods section).

We chose edgeR for this analysis for two main reasons:

First, we are not certain that we have a correct phylogeny of the *B. rhizoxinica* and the *Burkholderia* genus. Therefore, a phylogenetic guided comparative approach such as Count (Csuos 2010) or CAFE (De Bie et al. 2006) may lead to wrong interpretations.

Second, edgeR provides a simple way to get function comparisons with statistical significance values. EdgeR analyses are frequent in our laboratory so it represents a well-known tool. This knowledge comes handy in troubleshooting.

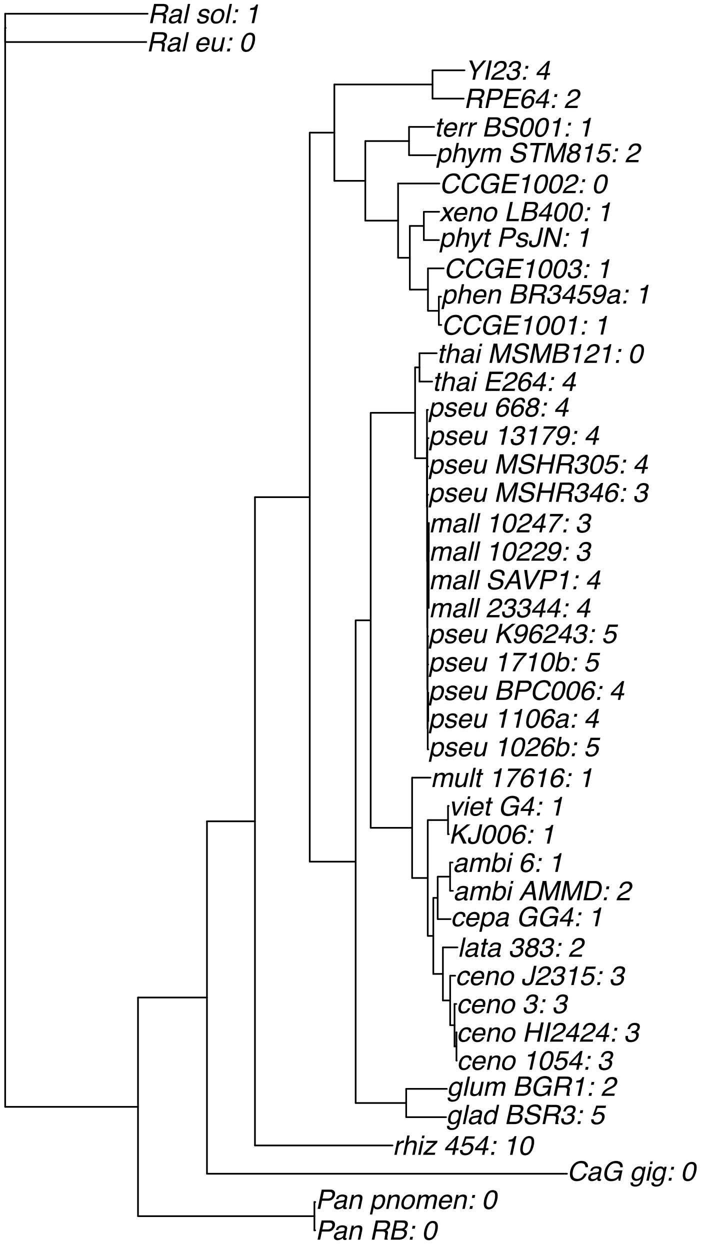
Statistically significant results for expanded functional categories are shown in Table 29. Complete tables for expanded functional categories are available in the digital supporting material (Supporting_material/3.Burkholderia/Comparative_genomics/B_rhizoxinica_vs_free-living and endofungal_vs_free-living). We will next discuss in more detail the categories of non-ribosomal peptide synthases and toxin-antitoxin replicon stabilization systems as these categories have the best significance statistical values (crossref Table 29) values.

Table 29. Expanded domains, subsystems and functions in B. rhizoxinica compared with free-living relatives.

Description	Category	logFC	logCPM	LR	PValue	FDR
Non-ribosomal peptide synthetase modules	proposed function	2.5	9.8	20	7E-06	0.002
Toxin-antitoxin replicon stabilization systems	subsystem	2.9	10.4	51	7E-13	8E-11
Coenzyme F420 synthesis	subsystem	4.6	8.8	28	8E-08	2E-06
Conjugative transfer	subsystem	1.9	10.8	18	2E-05	0.0003
T4-like virus tail tube protein gp19	Pfam	5.8	8.2	40	2E-10	2E-07
Phage tail sheath protein	Pfam	4.4	8.6	33	8E-09	2E-06
Helix-turn-helix domain	Pfam	1.4	10.7	11	0.0008	0.03
DNA polymerase III alpha subunit (EC 2.7.7.7)	Pfam	2.2	10.1	17	3E-05	0.008

Non-ribosomal peptide synthetases

The genome sequence of *B. rhizoxinica* revealed that it harbors 14 different non-ribosomal peptide synthetases (NRPS) (Lackner, Moebius, Partida-Martinez, et al. 2011). We compared the number of NRPS found by the antiSMASH server in the *Burkholderia* genus. To aid the visualization of this comparison, we added the number of NRPS to our previously computed "core" genome *Burkholderia* phylogeny (Figure 31). We found that the number of NRPS between *Burkholderia* is highly variable, ranging from 0 to 10. This variability could be due to NRPS being horizontally transferred. Notably, *B. rhizoxinica* has 10 NRPS; these are more NRPS than any free-living relative, despite being subjected to genome reduction. The average number of NRPS in free-living *Burkholderia* is only 2.5. The other available endofungal genome *Ca.* G. gigasporarum, has no NRPS, suggesting that *B. rhizoxinica* and *Ca.* G. gigasporarum have different secondary metabolite synthesis potential.



NRPS according to antismash

Figure 31. NRPS distribution in the Burkholderia genus according to antiSMASH

All *B. rhizoxinica* NRPS are expressed inside *R. microsporus* (Figure 32). Some of them are highly expressed. These expression patterns support conservation of all 10 NRPS in B1, B3 and B7, representing three out of eight known *R. microsporus* endosymbionts. The conservation in three different strains adds to the evidence for NRPS being involved in symbiosis, as we expect the three strains to be subject to genome erosion.

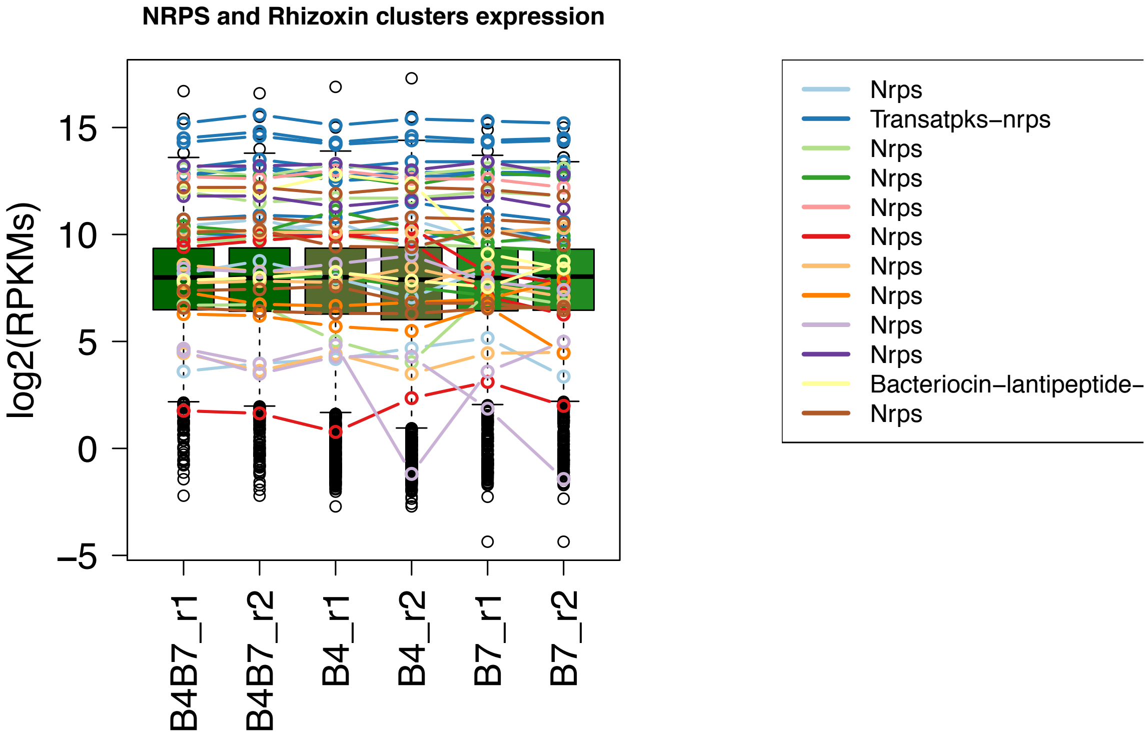


Figure 32. Non-ribosomal peptide synthetases (NRPS) and Rhizoxin cluster expression (Transatpks-nrps). Each cluster is colored differently

The high number of NRPS in *B. rhizoxinica*, their apparent conservation and expression suggests that they are relevant for symbiosis with *R. microsporus*. It is worth mentioning that we found evidence for the expression of 80% *B. rhizoxinica* genes inside *R. microsporus*.

We also detect high levels of expression for the Rhizoxin biosynthesis cluster (transatpks-nrps in Figure 32). Rhizoxin is already known to play an important role in this symbiosis by conferring the fungus with the capacity to infect rice. It is interesting that the Rhizoxin cluster shows such high expression levels when cultured in PDA and in absence of rice seedling. We cannot rule out a role of the Rhizoxin molecule beyond conferring pathogenicity to the fungus. *B. rhizoxinica* is depleted in transcriptional regulators (Lackner, Moebius, Partida-Martinez, et al. 2011); another possible explanation for high expression of Rhizoxin cluster is that it may have lost regulators, resulting in a constitutive high expression.

Our fungal comparative analysis revealed that Mucorales have less polyketide synthases and NRPS associated domains than Dykaria representatives. These domains include condensation, keto-reductase, enoyl-(acyl carrier protein) reductase, acyl transferase domain, among others (Table 27). The presence of *B. rhizoxinica* in *R. microsporus* host strains expands the natural product repertoire relative to all other Mucorales.

Toxin-antitoxin systems

Toxin-antitoxin (TA) systems are two or more linked genes that code for a poison protein and a corresponding antidote. These systems were first discovered in bacterial plasmids and help ensure their inheritance when the host bacterium divides. One of the two linked genes codes for a stable toxin protein that kills the bacterium or stops its growth. The antidote partner is labile and needs to be produced constantly to nullify the effects of the toxin. If the plasmid harboring the TA is inherited then both toxin and antitoxin are produced and bacteria live. However, when the plasmid is lost, the toxin and antidote present in the cytoplasm are inherited. Eventually the unstable antidote is degraded and the toxin is free to kill the bacterium.

Toxin-antitoxin systems are more abundant in endofungal bacteria than in any free-living *Burkholderia* (Table 29). In Figure 33 we show the number of TA proteins present in all analyzed genomes relative to their total protein coding genes. Interestingly *B. rhizoxinica* and *Candidatus* Glomeribacter gigasporarum have a number of TA proteins despite being the smallest genomes.

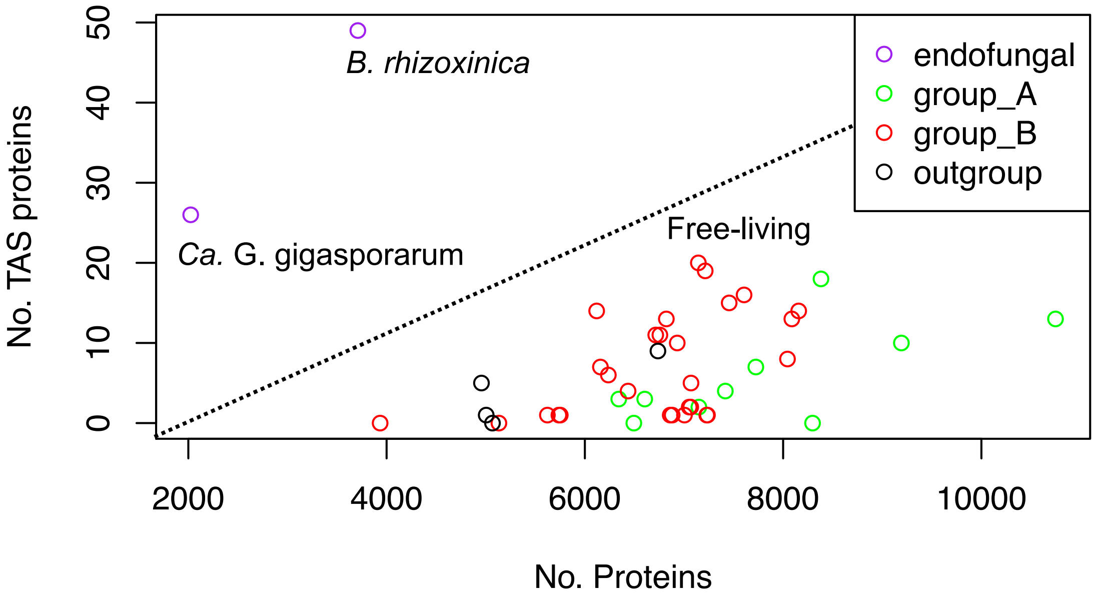


Figure 33. Toxin-antitoxin proteins are over-represented in endofungal bacteria. Toxin-antitoxin proteins in the Burkholderia genus relative to total protein number

We found VapC toxins in endofungal bacteria, which are characterized by having the PIN domain PF01850 (Gerdes et al. 2005). VapC toxins have endonuclease activity (van Melderen 2010). We found 11 and 8 proteins with the PIN domain in *B. rhizoxinica* and *Ca.* G. gigasporarum, respectively. These domains are enriched in endofungal bacteria relative to free-living *Burkholderias* (FDR 1.7e-07).

The majority of *B. rhizoxinica* TA proteins (33 out of 44) are expressed inside *R. microsporus*. Similar numbers of toxin and antitoxin genes are expressed, being 17 and 16 respectively. Interestingly, we found similar expression levels of toxins and antitoxins by comparing their expression distributions (Figure 34).

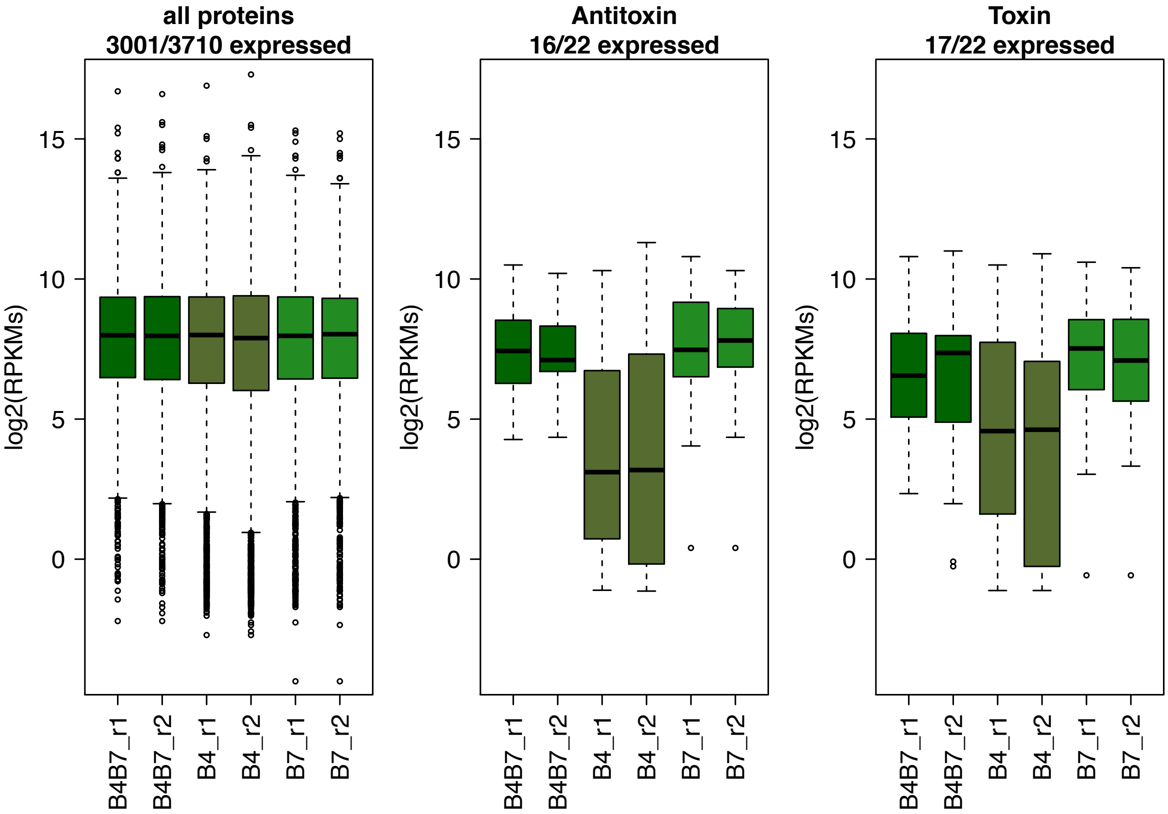


Figure 34. Toxin-antitoxin proteins are expressed inside host Rhizopus microsporus. Three sets of boxplots represent the expression distributions of all B. rhizoxinica proteins, antitoxins and cognate toxins. Expression is measuered as RPKMs (reads per kilobase per million). B4, B. rhizoxinica B4 strain. B7, B. rhizoxinica B7 strain. B4B7, library containing both B. rhizoxnica strains. R1 and r2 represent the number of replicate and are colored the same.

TA proteins are particularly abundant in both plasmids (Figure 35). Interestingly, a type IV secretion system (T4SS) is coded in pBRH02 plasmid. TA loci are encoded in the same plasmid, close to the T4SS. It is quite revealing that TA and T4SS share a gene neighborhood, suggesting that T4SS could be used to secrete TA proteins to the cytoplasm of *R. microsporus*.

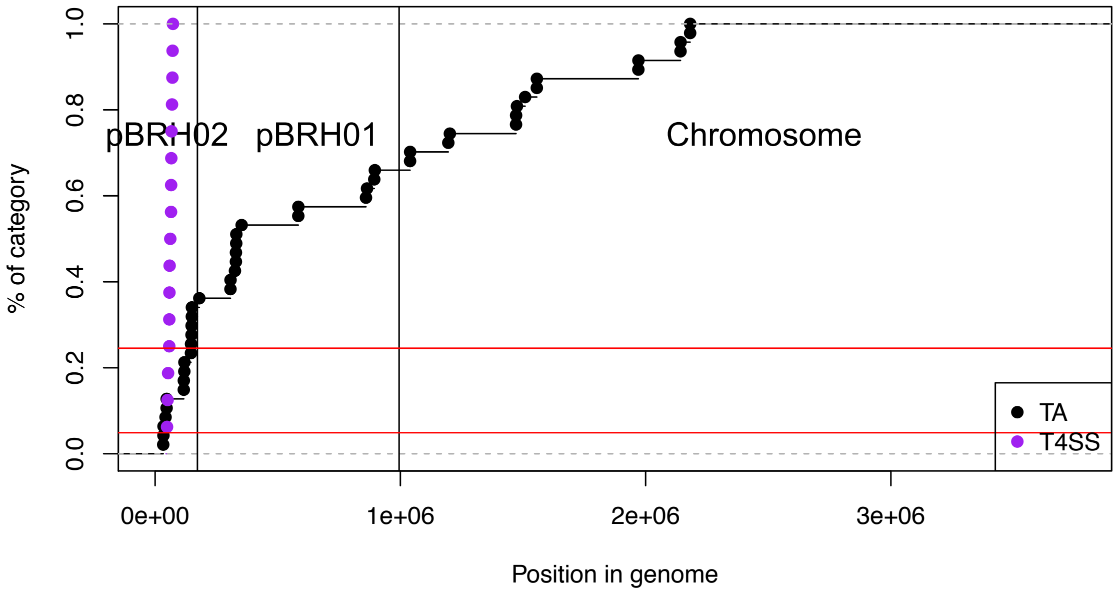


Figure 35. Toxin-antitoxin (TA) systems and Type IV secretion system (T4SS) distribution in B. rhizoxinica replicons. Cumulative distributions of TA (black) and T4SS (purple) proteins, vertical lines denote boundaries between different replicons. Horizontal red lines indicate the number of expected TA genes expected by the number of genes in each replicon

We found evidence for expression of all T4SS components inside *R. microsporus. B. rhizoxinica* B4 endosymbiont displays lower T4SS expression levels than the B7 strain (Figure 36). This finding represents an example of differences of gene expression between B4 and B7 strains. The lower expression of *B. rhizoxinica* B4 type IV secretion system correlates with a lower expression of toxin-antitoxins (Figure 34). Similar plots for other functional categories can be found in the digital supplementary material (Supporting_material/3.Burkholderia/B_rhizoxinica_expression_inside_R_microsporus).

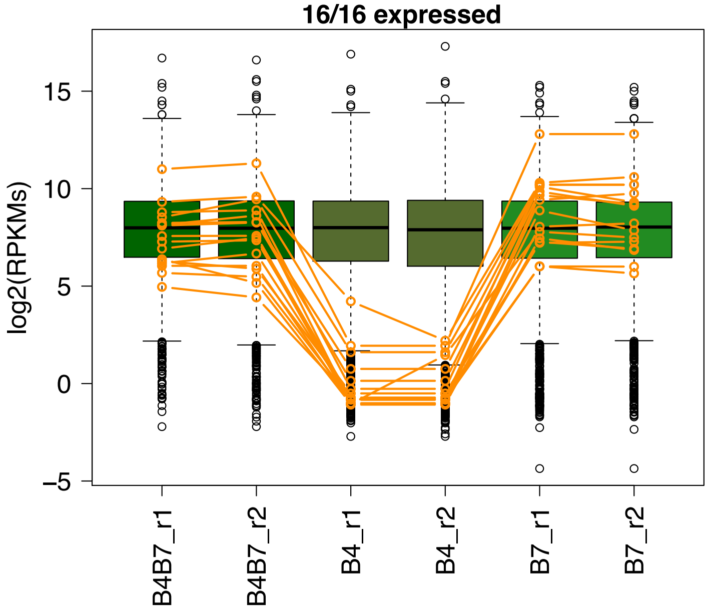


Figure 36. Type IV secretion system genes expression levels. Boxplots represent the expression distributions of all B. rhizoxinica proteins. Expression is measuered as RPKMs (reads per kilobase per million). The orange lines represent the expression levels for Type IV secretion system protein components.

We were curious if any TA protein is encoded in the *R. microsporus* genomes. We identified those Pfam domains associated with TA proteins (Table 18) and searched for them in all Mucorales proteins. We didn't find *B. rhizoxinica* TA protein domains in any Mucoralean genome, regardless of being host or non-host. Our search for TA *Burkholderia* protein domains in *R. microsporus* discards the possibility of an endosymbiont-host toxin-antitoxin complementation.

We propose that toxin-antitoxin systems are relevant to *Rhizopus microsporus* – *Burkholderia rhizoxinica*. This proposal is based in the enrichment of TA proteins in endofungal bacteria (Figure 33) and on the majority of these genes being expressed inside its host, albeit to a lower level in endosymbiont B4 strain (Figure 34).

We envision a model where *B. rhizoxinica* produces stable toxins and labile antitoxins that end up in the cytoplasm of *R. microsporus* (Figure 37). The endosymbiont represents the source of both components. While the bacterium is present the toxin is inactivated. However, when the endosymbiont is removed, labile antitoxins decay and toxins are free to harm the fungus. This mechanism could contribute to the vertical inheritance of *B. rhizoxinica*.

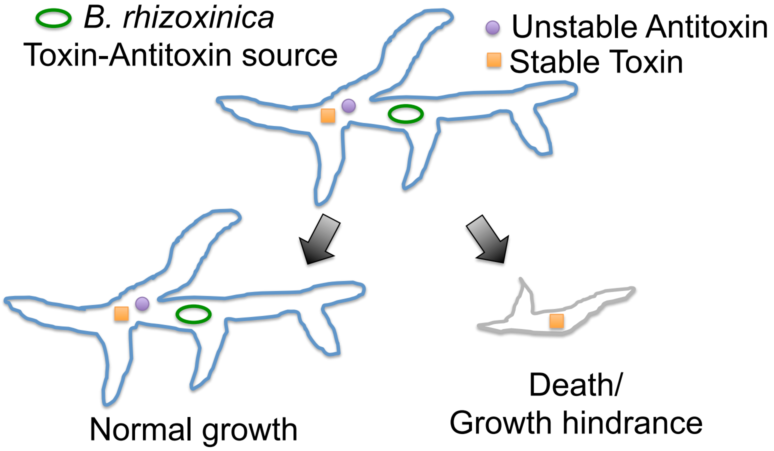


Figure 37. Model for toxin-antitoxin usage in the maintenance of R. microsporus – B. rhizoxinica symbiosis. Healthy R. microsporus mycelium is colored blue. Unhealthy/Death mycelium is colored grey. B. rhizoxinica is represented by a green empty oval. Toxin and antitoxins are represented by an orange square and a purple circle.

Chapter VI: Discussion

Narnaviruses in *Rhizopus microsporus*

Narnavirus distribution

Narnavirus reports are scarce, so far they have only been reported in *Saccharomyces cerevisiae* and in the oomycete *Phytophtora infestans* (Hillman & Cai 2013). Our analysis represents the first evidence for *Narnavirus* occurrence in Mucorales.

Our results support *Narnavirus* presence in ATCC52814 and absence in the ATCC52813 strain (Figure 11). This raises the question of how common are *Narnavirus* infections among *R. microsporus* strains? A PCR based approach could aid in the detection of *Narnavirus* in *R. microsporus* strains to determine its frequency. It would be interesting to determine if the viral infections are equally frequent in host and non-host strains.

Mycoviruses have two modes of reproduction: transmission by spores and hyphal fusion (anastomosis) (Nuss 2005). Viruses present in ATCC52814 could potentially infect ATCC52813 or any other compatible strain during mating. It would be interesting to determine if ATCC52813 x ATCC52814 mating descendants are still infected with *Narnavirus*.

Narnavirus phenotype

Narnavirus numbers increase in *S. cerevisiae* in stressful conditions such as Nitrogen starvation and heat shock (López et al., 2002). Some authors suggest that *Narnavirus* presence helps fungal hosts to cope with stress. The number of *Narnavirus* could be traced in *R. microsporus* growing in Nitrogen starvation or heat shock to determine if a similar phenomenon happens in this system.

A study on *Curvularia protuberata* discovered the presence of a mycovirus in this fungus. In this work authors generated virus-free fungal strains, and this helped to determine the phenotypic impact on its host (Márquez Luis M., Redman Regina S., Rodriguez Russell J. 2007). The elimination of the virus was achieved by freezing mycelia and lyophilization. The same procedure could be applied to ATCC52814 in order to obtain virus free isolates. If this is achieved, then the phenotypic properties of an ATCC52814 infected versus a cured strain could determine if the *Narnavirus* has any effect on its host.

Microinjection could be an interesting approach for *Narnavirus* studies. The cytoplasm of ATCC52814 cured (A14c) could be microinjected to ATCC52813. In A14c strain, *B. rhizoxinica* is absent but *Narnavirus* is present (Figure 11). This assay could test if the ATCC52813 strain is susceptible to *Narnavirus* infection.

The microinjection may be applied to any other *Rhizopus* strain, regardless of being host or non-host.

According to our results, *Narnavirus* RNA is the main contributor of the high GC peak (Figure 11) and the GC peak has more reads than those mapping to *B. rhizoxinica* (Figure 10). Therefore, viral RNA molecules potentially outnumber *B. rhizoxinica* transcripts in ATCC52814. Although read counts cannot immediately be translated into actual RNA molecules. It would be really interesting to know if *Narnavirus* particles have any influence on *B. rhizoxinica* biology.

If *Narnavirus* presence influences *B. rhizoxinica* or *R. microsporus* fitness, then we would be dealing with a fungus-bacterium-virus three-party symbiosis.

Differential expression analysis

Given the experimental design (Figure 4) it is important to note that observed differences in expression could be due to different developmental programs. Expression profiles of mycelia, sporangia and zygospores are likely to be very different. This implies that although some of the DE genes could be caused directly by the presence of the bacteria, many others could be the consequence of sampling very different developmental structures with their own expression repertoire.

Our analysis compares overall trends between *R. microsporus* plates. We lack more resolution like structure enriched RNA sequencing, such as sporangiophores, zygospores or young and old mycelium.

Our results show less expression of nuclear transcripts in presence of *B. rhizoxinica* (Figure 16). Structures such as sporangia, zygosores and spores were present when samples were taken in endosymbiont harboring libraries. Two different scenarios could explain the difference in nuclear transcripts: first, nuclei present in spores could be transcriptionally less active than those present in mycelium. Alternatively, the difference in nuclear transcripts could be explained by the presence of fewer nuclei in sporulating plates relative to one with plain mycelia. This suggestion sounds less likely if we consider that each fungal spore should harbor at least one nucleus to be able to germinate. Depending on its size, a Mucoralean sporangium may contain up to 100,000 spores (Richardson 2009). Therefore, we would expect as many as 100,000 nuclei per sporangium. This sounds like a lot of nuclei in a small space when *Rhizopus* is sporulating.

The endosymbiont seems to impact the regulation of *R. microsporus* gene expression at different levels. Differentially expressed genes include transcription factors, histones and proteins involved in chromatin assembly or disassembly. We suspect that some of these changes could be the result of *B. rhizoxinica* SET (BrSET) protein methylating some of its host histones. We found evidence for BrSET expression along with a Type III secretion system (T3SS) (Figure 21).

Burkholderia thailandensis secretes a SET protein to host cell lines via a T3SS and this influences the expression of host rDNA (Li et al. 2013). We suspect a similar process could be happening between *B. rhizoxinica* and *R. microsporus*.

We found a fungal GATA-SET protein that is up-regulated in presence of *B. rhizoxinica* (Figure 20). The GATA-SET architecture is new and exclusive to *R. microsporus* species (Table 25). This is something remarkable as there are 554 different protein domain architectures with at least one SET domain in the Pfam database. In eukaryotes, SET proteins frequently act in protein complexes, which have different outcomes in gene expression (Herz et al. 2013). It is intriguing to know if GATA-SET would form new protein complexes as it represents a new architecture. It would be important to determine its substrates and the level of methylation (mono-, di- or trimethylation).

Transcription factors represent the last link between signal transduction cascades and target genes expression. Our results evidence a strong trend towards down-regulation of TFs in presence of the endosymbiont (Figure 16, Figure 17). Considering the developmental stage, this translates to more TFs in plain mycelia vs sporulating mycelia.

We found that the down-regulation trend holds for different TFs classes defined by Pfam domains (Figure 17). In humans, chromatin immuno precipitation sequencing analyses have revealed that different transcription factor families display affinity for different DNA binding sites. For example, GATA transcription factors tend to bind to “GATA” sequences while homebox domains bind “TAAAA” sites (Jolma et al. 2013). These DNA preference differences could be conserved in fungi.

We also found an enrichment of LysM proteins in down-regulated genes due to *B. rhizoxinica* presence (Table 24). LysM proteins mediate bacterial perception in plants, both in pathogen and mutualistic interactions (Gust et al. 2012). These results, suggest that *B. rhizoxinica* down-regulates the expression of its host bacterial sensing receptors.

Better interpretations of our RNA-Seq results are limited by the general lack of studies in Mucorales, relative to Dykaria. A considerable amount of knowledge in fungi development has accumulated in Ascomycetes such as *Neurospora crassa* and *Aspergillus nidulans* (Park & Yu 2012). *Ustilago maydis* is perhaps the best studied Basidiomycete. A proper identification of ortholog genes between better studied fungal species and *R. microsporus* would aid the interpretation of its expression profile.

Fungal comparative genomics discussion

We built a phylogenetic tree of Mucorales fungi based on 46 proteins. We found two different *Rhizopus microsporus* groups, var *microsporus* and var *chinensis*. According to our tree, the host trait is paraphyletic because it is found in both *R. microsporus* clades (Figure 23).

The use of genomic data to build phylogenies has the advantage of providing a large number of loci. However, sequencing genomes is much more expensive than amplifying and sequencing specific markers. In our phylogeny we have five *R. microsporus* representatives, this is a small number relative to the study performed by Dolatabati, which analyzed 48 *R. microsporus* isolates (Dolatabadi et al. 2013). The limited number of strains in our study hinders the drawing of broader conclusions.

Rhizopus genomes have accumulated rapidly, from 2013 to 2015 the number of *Rhizopus* sequences increased from two to six. This is remarkable as the *Rhizopus* genus has always being overlooked in comparison with Dykaria models. So far the information for 451 Dykaria genomes is available at the Joint Genome Institute database, consisting of 296 Ascomycota and 155 Basidiomycota genomes. Four of the *Rhizopus* genomes were sequenced aiming to shed light on the *Rhizopus – Burkholderia* symbiosis. In a near future all eight fungus-bacterium genome pairs may be available. The *Rhizopus-Burkholderia* system is attractive enough to soon overcome the genome data limitation. However, the funding for new sequencing projects of *Rhizopus-Burkholderia* genomes will surely depend on the benefits we can get from them, therefore it is worth trying different bioinformatic approaches to mine from them as much information as possible.

Different Institutions assembled and annotated their genomes according to their standards. Ideally, a unique pipeline of assembly and annotation could make these data more comparable. In addition, the genomes provided are all in a draft stage. Comparative genomics analyses ideally require closed genomes. In this way, we would be sure that observable differences are real and not due to missing data. When working with draft genomes we can’t be completely sure that the absence of genes is due to lack of information or a bona fide loss. However draft genomes provide a mean to compare general gene contents, we performed this kind of general comparisons.

The host trait is found in *R. microsporus* strains with small and big genomes (Figure 24). Considering that *Rhizopus* does not easily incorporates foreign DNA in its genome (Xu et al. 2014), small genome strains would be preferable to construct loss of function mutants. Duplicated genes (paralogs) could replace the function of a mutated gene; the choice of a smaller genome would avoid large numbers of paralogs.

PIF1 helicases

Our analysis reveals an interesting distribution pattern for PIF1 helicases. They are expanded in Mucorales (Table 28) but reduced in host *Rhizopus* genomes (Figure 25).

PIF1 family helicases are involved in nuclear and mitochondrial genome maintenance (Bochman et al. 2010). PIF1 helicases are conserved in almost all eukaryotes. Dykarial fungi typically have two Pif1 helicases and metazoans have just one copy. Kinetoplast parasites have seven or eight Pif1 helicases (Bochman et al. 2010). According to our results some *Rhizopus* members have up to 32 Pif1 helicases. Host *Rhizopus* strains display the lowest numbers of Pif1 helicases in the genus.

Saccharomyces cerevisiae has two PIF1 helicases, ScPf1 and ScRrm3. *R. microsporus* ATCC52813 has only 1 PIF1 helicase; it would be worth to know if it resembles more ScPf1 or ScRrm3, which have different functions. This is particularly relevant as ScPf1 and ScRrm3 sometimes have opposing roles, as is the case for rDNA replication. ScPf1 inhibits and ScRrm3 promotes fork progression through rDNA (Bochman et al. 2010).

Despite a revision of the literature we couldn’t come up with a proposal for the role of PIF1 helicases in the *R. microsporus* – *B. rhizoxinica* interaction.

Reported differences between Mucorales and Dykaria

Some of the differences of protein domain contents we found were already reported between Mucorales and Dykaria. For example, an expansion of Ras proteins and chitin deacetylases was reported for *R. oryzae* (Ma et al. 2009). It was already reported that Mucorales have more chitin synthases than some Ascomycetes (Ruiz-Herrera & Ortiz-Castellanos 2010).

Regarding reductions, Mucorales are considered non-cellulolytic microorganisms (Richardson 2009). This is consistent with our findinding of fewer for cellulase and fungal cellulose binding domains in Mucorales (Table 27). These functional categories serve as positive controls and gives us confidence in our analysis. We expand this knowledge to other unreported domains such septins, ricin B lectins and PIF1 helicases.

The finding of higher numbers of septins in Mucorales is quite interesting as these fungi are non-septated. We also found lots of septins over-expressed in presence of *B. rhizoxinica*, a condition that correlates with more diverse morphologies than plain mycelia. Septins are generally present in morphological transition boundaries. In these spots septins can act as diffusion barriers of membrane-associated proteins (Bridges & Gladfelter 2014). This restriction could potentially influence the localization of other membrane-associated morphologically relevant proteins such as Ras and Rho GTPases.

Non-ribosomal peptide synthetases in Mucorales and *B. rhizoxinica*

Our analysis is consistent with a previous report in which Non ribosomal peptide synthetases (NRPS) are almost absent in Zygomycetes (Bushley & Turgeon 2010). This is supported by the difference in condensation domains, an essential domain of NRPS, between Mucorales and Dykaria (Table 27).

B. rhizoxinica has ten non-ribosomal peptide synthetases, more than any free-living *Burkholderia* (Figure 31). This is equivalent to half the number of total natural product (NP) gene clusters present in a typical actinomycete. This finding is quite remarkable as actinomycetes are famous for being a clade rich in NP gene clusters (Traxler & Kolter 2015). NRPS numbers are surprising given that *B. rhizoxinica* has such a smaller genome size (3.75 Mb) than actinomycetes (~8 Mb).

B. rhizoxinica NRPS are expressed inside *R. microsporus*, this adds to the evidence of NRPS being relevant for this bacterial-fungal symbiosis.

The richness of NRPS in *B. rhizoxinica* likely increases the repertoire of natural products in host strains relative to non-host *R. microsporus* and to Mucorales, giving the host strains a competitive advantage.

Toxin-antitoxin systems in *R. microsporus – B. rhizoxinica*

We found that toxin-antitoxin (TA) system proteins are enriched in endofungal bacteria in comparison with free-living relatives (Figure 33). The majority of TA genes are expressed inside *R. microsporus* (Figure 34).

Our comparative analysis is restricted to a single *B. rhizoxinica*, the B1 strain. It would be very interesting to have all eight genomes to see how similar they are. Do they all have the same classes of TA? Do they all have similar numbers of TA?

We propose that toxin-antitoxin systems are relevant to the maintenance of *Rhizopus microsporus – Burkholderia rhizoxinica* symbiosis. Two reports are key to our proposal:

First, a 2011 paper reports that bacterial TA toxins can harm eukaryotic cells (Audoly et al. 2011). In this study, microinjection of *Rickettsia felis* VapC toxins caused apoptosis in L929 mouse cells. Additionally, heterologous expression of VapC proteins in *Escherichia coli* and *Saccharomyces cerevisiae* revealed that *R. felis* proteins are toxic to both bacteria and eukaryotes. These detrimental effects are nullified if VapC is expressed along with VapB, its cognate antitoxin. *R. felis* VapC toxin has *in vitro* RNase activity.

Secondly, a 2013 paper revealed a possible role for TA systems in vertical transmission of bacteria in an eukaryote host (Socolovschi et al. 2013). In this study the number of TA genes in bacteria belonging to the order Rickettsiales was positively correlated with their vertical transmission in arthropod vectors such as flea and ticks.

Our results together with these reports suggest similar molecular mechanisms participating in the vertical transmission of Beta-proteobacteria–fungi and Rickettsiales-arthopods.

We found that some TA genes are in the same genomic neighborhood to a type IV secretion system (T4SS) (Figure 35). This is also the case in some other bacterial species such as *Bartonella rattaustraliani* (Saisongkorh et al. 2010) and *Rickettsia felis* (Audoly et al. 2011). It is tempting to speculate that the TA system toxins could be secreted via the T4SS. In this regard, a 2010 report found evidence for conjugation transfer of T4SS genes between *Bartonella* species inside *Acanthamoeba polyphaga* amoeba. The sequence of a *Bartonella* plasmid, pNH4, was determined and shown to harbor a T4SS and nearby TA genes. Interestingly, the authors found higher numbers of *B. rattaustraliani* cells in amoeba cytoplasm when comparing a plasmid conjugant strain pNH4(+) and a non-conjugant strain pNH4(-) (Saisongkorh et al. 2010). However, in this study it is impossible to directly link TA presence and greater intracellular survival. The higher intracellular numbers effect could be the result of any other gene present in the pNH4 plasmid.

All *B. rhizoxinica* T4SS components are expressed inside its fungal host (Figure 36). Therefore a complete T4SS could be used by *B. rhizoxinica* to transfer DNA or proteins into *R. microsporus*.

Interestingly, *Rickettsia* VapC toxins have RNase activity. RNase activity was assessed with *in vitro* RNA phage degradation (Audoly et al. 2011). If *B. rhizoxinica* VapC toxins do have RNase activity and if these toxins were found in *R. microsporus* cytoplasm, then host mRNAs and *Narnavirus* ssRNA may be susceptible to degradation. This could be possible given that ssRNA is sensitive to digestion by ribonucleases (Bozarth 1972). *B. rhizoxinica* VapC endonuclease activity could represent a mechanism to control *R. microsporus* expression and influence *Narnavirus* numbers.

However, we observe an opposite effect in our RNA-Seq data. Narnavirus are found in all *R. microsporus* ATCC52814 containing libraries and *B. rhizoxinica* B7, the endosymbiont of ATCC52814, has higher expression levels of its TA genes. Therefore *Narnavirus* presence correlates with higher expression of TA systems. This observation is based on few samples and in a single experiment, therefore it needs to be revisited.

The role of TA in *B. rhizoxinica* – *R. microsporus* symbiosis maintenance represents a promising research line in this system. We propose some experiments to evaluate TA function in the perspectives section of this thesis.

Conclusions

In this work we set the main goal of identifying candidate genes involved in the symbiosis between *Rhizopus microsporus - Burkholderia rhizoxinica*.

To achieve this main goal we perfomed the following analyzes:

1. We looked for differentially expressed *R. microsporus* due to the presence or absence of *B. rhizoxinica*.

We provide an overview of the processes being affected by the bacterial endosymbiont. Nuclear genes such as transcription factors and SET proteins are differentially expressed. We also contribute with a glimpse of some of the *B. rhizoxinica* genes expressed inside *R. microsporus*, such as BrSET and a type III secretion system.

While analyzing RNA-Seq data we found evidence for *Narnavirus* presence in *Rhizopus microsporus*. This finding incorporates a new player in the *Rhizopus microsporus – Burkholderia rhizoxinica* symbiosis.

2. We compared the genomes of host and non-host *R. microsporus* strains in an attempt to better understand the molecular basis of this trait. Additionally, we compared Mucoralean genomes with some dykaria model fungi genomes.

Our analysis could not detect statistcally-supported differences in protein domain contents of host and non-host *Rhizopus microsporus* strains. The closest protein domain to meet the statistical criterion was PIF-1 helicases, which are less abundant in host strains.

We provide an overview of differences between Mucorales and Dykaria protein domain contents. Our resulting list includes known differences such as chitin synthases, non-ribosomal peptide synthase domains, Ras proteins, Ras regulators, cellulases that serve as positive controls. We expand the knowledge of differences between Mucorales and Dykaria with other previously non-reported domain differences such as septins and leucine rich domains.

3. We compared the genomes of free-living *Burkholderia* spp. with the genomes of *B. rhzoxinica* and *Candidatus* Glomeribacter gigasporarum.

We found more non-ribosomal peptide synthetases and toxin-antitoxin systems in *B. rhizoxinica* than in any free-living *Burkholderia* genome.

Our study suggests that toxin-antitoxin systems may be used as a endosymbiont addiction mechanism in both *B. rhzoxinica* and *Candidatus* Glomeribacter gigasporarum.

Our final list of candidate genes for further studies is shown in Table 30. In this table we include organism source, gene identification number and the evidence to support it.

Table 30. Candidate symbiosis relevant genes. Br stands for B. rhizoxinica and Rm stands for R. microsporus.

Candidate genes	Org. source	IDs	Evidence
BrSET	Br	RBRH_00796	*In vitro* activity, expression inside host
RmGATA-SET	Rm	218959 (protein id)	Differentially expressed in presence of Br, interesting distribution pattern in *Rhizopus*
NRPS	Br	fig|32008.60.peg.2568 fig|32008.60.peg.2569 fig|32008.60.peg.2570 fig|32008.60.peg.2573	More NRPS in Br than any free-living relative, expressed in B4 and B7, located in Br chromosome
Toxin-antitoxin VapC	Br	fig|32008.60.peg.276 fig|32008.60.peg.277	TA systems over-represented in endofungal bacteria, expressed in B4 and B7, probable endonuclease, located in Br pBRH01 plasmid
Toxin-antitoxin VapC	Br	fig|32008.60.peg.2278	TA systems over-represented in endofungal bacteria, expressed in B4 and B7, probable endonuclease, located in Br chromosome
Toxin-antitoxin VapC	Br	fig|32008.60.peg.3553 fig|32008.60.peg.3554	TA systems over-represented in endofungal bacteria, expressed in B4 and B7, probable endonuclease, located in Br pBRH02 plasmid

Perspectives

Narnavirus

The discovery of *Narnavirus* in *R. microsporus* is explorative; therefore a large number of interesting experiments would be very interesting.

To quantify Narnaviruses through *R. microsporus* development and under different growth conditions. A detection of a difference in *Narnavirus* particles could give some hint if it has an impact on *R. microsporus* fitness. We propose qRT-PCR as a technique to achieve the quantification.

We don't know how stable are *Narnavirus* particles through *R. microsporus* generations. Therefore we propose to sub-cultivate ATCC52814 and to perform virus detection to determine the infection stability.

To determine *Narnavirus* presence among all possible *R. microsporus* strains. This could insights into the ecological role of *Narnavirus* presence.

It would be interesting to know if ATCC52813 x ATCC52814 mating descendants are also infected with the *Narnavirus* found in ATCC52814.

Finally, we suggest the microinjection of ATCC52814 *B. rhizoxinica* cured strain cytoplasm into ATCC52813. This experiment would shed light into the infective capabilities of *Narnavirus* on ATCC52813. The microinjection procedure could be applied to other *R. microsporus* strains. The inclusion of host and non-host strains for *B. rhizoxinica* could expand the scope of this symbiosis system.

SET proteins

We propose to perform an *in vitro* characterization of *R. microsporus* GATA-SET protein (RmGATA-SET). The goals are to determine if RmGATA-SET is able to methylate histones *in vitro* and to find its substrate specificities.

To determine the role of *B. rhizoxinica* SET protein, it would be relevant to construct a loss of function bacterium mutant of BrSET. If this protein is relevant for the *R. microsporus – B. rhizoxinica* symbiosis it should display a less fit phenotype than the wild-type endosymbiont.

Burkholderia endofungal bacteria evolution

We used the merged information of 541 protein-coding genes to to explore the evolution of endofungal bacterium of the *Burkholderia* genus. However, we are not completely certain of the correctess of the phylogeny. Each of the 541 genes contributes with phylogenetic signal. We could explore the contribution of different subsets of proteins and the overall consistency of our phylogeny with a Random Addition Concatenation Analysis (Narechania et al. 2012).

We used the Poisson distribution as background model for protein evolution. The usage of ProtTest could aid the choice of a better amino acid evolution model. The background evolution model could substancially change the results.

Toxin-antitoxin

To test the toxicity of *B. rhizoxinica* TA (BrTA) proteins in *R. microsporus* we propose to introduce free toxin in *R. microsporus* mycelium via microinjection. Including host and non-host *Rhizopus microsporus* strains in this experiment would determine if the toxin equally affects both kinds of strains. Heterologous expression in *E. coli* could be used to obtain purified toxin and to test the toxicity of BrTA in bacteria. The heterologous expression of BrTA in *Saccharomyces cerevisiae* could determine if these toxins can harm a eukaryote cell.

References

Akira, S., Uematsu, S. & Takeuchi, O., 2006. Pathogen Recognition and Innate Immunity. *Cell*, 124(4), pp.783–801. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0092867406001905.

Altschul, S. et al., 1997. Gapped BLAST and PSI- BLAST: a new generation of protein database search programs. *Nucleic acids Res*, 25(17), pp.3389–3402. Available at: http://nar.oxfordjournals.org/content/25/17/3389.short.

Alvarez-Venegas, R., 2014. Bacterial set domain proteins and their role in eukaryotic chromatin modification. *Frontiers in Genetics*, 5(APR), pp.1–8.

Anders, S., Pyl, P.T. & Huber, W., 2014. HTSeq – A Python framework to work with high-throughput sequencing data HTSeq – A Python framework to work with high-throughput sequencing data. , pp.0–5.

Audoly, G. et al., 2011. Effect of rickettsial toxin vapC on its eukaryotic host. *PLoS ONE*, 6(10).

Aziz, R.K. et al., 2008. The RAST Server: rapid annotations using subsystems technology. *BMC genomics*, 9, p.75.

Benjamini, Y. & Hochberg, Y., 1995. Controlling the False Discovery Rate : a Practical and Powerful Approach to Multiple Testing When researchers tend to select pursuing multiple the (statistically) and support of conclusions . An unguarded use in a greatly results of single-inference inc. *J.R Statist. Soc.B*, 57(1), pp.289–300.

Berger, S.L., 2007. The complex language of chromatin regulation during transcription. *Nature*, 447(7143), pp.407–412. Available at: http://www.nature.com/nature/journal/v447/n7143/pdf/nature05915.pdf.

De Bie, T. et al., 2006. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics (Oxford, England)*, 22(10), pp.1269–71. Available at: http://bioinformatics.oxfordjournals.org/content/22/10/1269.abstract.

Biterge, B. & Schneider, R., 2014. Histone variants: Key players of chromatin. *Cell and Tissue Research*, 356(3), pp.457–466.

Blakeslee, A.F., 1904. Sexual Reproduction in the Mucorineae. *Proceedings of the National Academy of Arts and Sciences*, 40(4), pp.205–319.

Bochman, M.L., Sabouri, N. & Zakian, V. a., 2010. Unwinding the functions of the Pif1 family helicases. *DNA Repair*, 9(3), pp.237–249. Available at: http://dx.doi.org/10.1016/j.dnarep.2010.01.008.

Bozarth, R.F., 1972. Mycoviruses: a new dimension in microbiology. *Environmental Health Perspectives*, 2(October), pp.23–39.

Brasier, C., 1983. A cytoplasmically transmitted disease of Ceratocystis ulmi. *Nature*, 305(5931), pp.220–223. Available at: http://www.nature.com/nature/journal/v305/n5931/abs/305220a0.html.

Bridges, A. a. & Gladfelter, A.S., 2014. Fungal pathogens are platforms for discovering novel and conserved septin properties. *Current Opinion in Microbiology*, 20, pp.42–48. Available at: http://dx.doi.org/10.1016/j.mib.2014.04.004.

Bryant, C.E. et al., 2010. The molecular basis of the host response to lipopolysaccharide. *Nature reviews. Microbiology*, 8, pp.8–14. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19946286.

Buist, G. et al., 2008. LysM, a widely distributed protein motif for binding to (peptido)glycans. *Molecular Microbiology*, 68(4), pp.838–847.

Bushley, K.E. & Turgeon, B.G., 2010. Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships. *BMC evolutionary biology*, 10, p.26.

Capuano, F. et al., 2014. Cytosine DNA methylation is found in Drosophila melanogaster but absent in Saccharomyces cerevisiae, Schizosaccharomyces pombe, and other yeast species. *Analytical chemistry*, 86(8), pp.3697–702. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4006885&tool=pmcentrez&rendertype=abstract.

Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution*, 17(4), pp.540–552.

Charif, D. & Lobry, J.R., 2007. Seqin R 1.0-2: project for statistical computing devoted to biological sequences retrieval and analysis. *Structural approaches to sequence evolution: Molecules, networks, populations*, pp.207 – 232.

Chowdhury, P.R. & Heinemann, J.A., 2006. The General Secretory Pathway of Burkholderia gladioli pv. agaricicola BG164R Is Necessary for Cavity Disease in White Button Mushrooms. *Applied and environmental microbiology*, 72(5), pp.3558–3565.

Csuos, M., 2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, 26(15), pp.1910–1912. Available at: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btq315.

Dale, C. & Moran, N. a., 2006. Molecular Interactions between Bacterial Symbionts and Their Hosts. *Cell*, 126, pp.453–465.

Dean, P., 2011. Functional domains and motifs of bacterial type III effector proteins and their roles in infection. *FEMS microbiology reviews*, 35(6), pp.1100–25. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21517912 [Accessed March 26, 2013].

Dehé, P.-M. & Géli, V., 2006. The multiple faces of Set1. *Biochemistry and cell biology = Biochimie et biologie cellulaire*, 84(4), pp.536–548.

van Diepeningen, A.D., Debets, A.J.M. & Hoekstra, R.F., 2006. Dynamics of dsRNA mycoviruses in black Aspergillus populations. *Fungal genetics and* *biology : FG & B*, 43(6), pp.446–52. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16546419.

Dolatabadi, S. et al., 2013. Diversity and delimitation of Rhizopus microsporus. *Fungal Diversity*. Available at: http://link.springer.com/10.1007/s13225-013-0229-6 [Accessed September 17, 2013].

Drinnenberg, I. a, Fink, G.R. & Bartel, D.P., 2011. Compatibility with killer explains the rise of RNAi-deficient fungi. *Science (New York, N.Y.)*, 333(6049), p.1592. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3790311&tool=pmcentrez&rendertype=abstract.

Du, J. et al., 2015. DNA methylation pathways and their crosstalk with histone methylation. *Nature Publishing Group*, 16(9), pp.519–532. Available at: http://dx.doi.org/10.1038/nrm4043.

Eddy, S., 1998. Profile hidden Markov models. *Bioinformatics*, 14(9), pp.755–763. Available at: http://bioinformatics.oxfordjournals.org/content/14/9/755.short.

Edgar, R.C., 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), pp.1792–1797.

Eissenberg, J.C., 2001. Molecular biology of the chromo domain: An ancient chromatin module comes of age. *Gene*, 275(1), pp.19–29.

Enright, a J., Dongen, S. V & Ouzounis, C. a, 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7), pp.1575–1584.

Escoll, P. et al., 2015. Targeting of host organelles by pathogenic bacteria: a sophisticated subversion strategy. *Nature Reviews Microbiology*. Available at: http://www.nature.com/doifinder/10.1038/nrmicro.2015.1.

Estrada-de los Santos, P. et al., 2013. Phylogenetic analysis of burkholderia species by multilocus sequence analysis. *Current microbiology*, 67(1), pp.51–60. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23404651 [Accessed August 8, 2013].

Fineran, P.C. et al., 2009. The phage abortive infection system, ToxIN, functions as a protein-RNA toxin-antitoxin pair. *Proceedings of the National Academy of Sciences of the United States of America*, 106(3), pp.894–9. Available at: http://www.pnas.org/content/106/3/894.short.

Finn, R.D. et al., 2014. Pfam: The protein families database. *Nucleic Acids Research*, 42(D1), pp.222–230.

Gentleman, R.C. et al., 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10), p.R80. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=545600&tool=pmcentrez&rendertype=abstract.

Gerdes, K., Christensen, S.K. & Løbner-Olesen, A., 2005. Prokaryotic toxin-antitoxin stress response loci. *Nature reviews. Microbiology*, 3(5), pp.371–382.

Grabherr, M.G. et al., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, 29(7), pp.644–652.

Gust, A. a. et al., 2012. Plant LysM proteins: Modules mediating symbiosis and immunity. *Trends in Plant Science*, 17(8), pp.495–502.

Herz, H.M., Garruss, A. & Shilatifard, A., 2013. SET for life: Biochemical activities and biological functions of SET domain-containing proteins. *Trends in Biochemical Sciences*, 38(12), pp.621–639. Available at: http://dx.doi.org/10.1016/j.tibs.2013.09.004.

Hillman, B.I. & Cai, G., 2013. *The Family Narnaviridae. Simplest of RNA Viruses.* 1st ed., Copyright &copy; 2013, Elsevier Inc. All Rights Reserved. Available at: http://dx.doi.org/10.1016/B978-0-12-394315-6.00006-4.

Hoffman, M.T. & Arnold, a E., 2010. Diverse bacteria inhabit living hyphae of phylogenetically diverse fungal endophytes. *Applied and environmental microbiology*, 76(12), pp.4063–75. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2893488&tool=pmcentrez&rendertype=abstract [Accessed March 28, 2013].

Holland, J.N. & Bronstein, J.L., 2008. Mutualism. , pp.2485–2491.

Horn, F., Guthke, R. & Hertweck, C., 2015. Draft Genome Sequences of Symbiotic and Nonsymbiotic Rhizopus microsporus Strains CBS 344 . 29 and ATCC 62417. , 3(1), pp.14–15.

Inohara, N. & Nunez, G., 2003. NODs: intracellular proteins involved in inflammation and apoptosis. *Nat Rev Immunol*, 3(5), pp.371–382. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12766759\nhttp://www.nature.com/nri/journal/v3/n5/abs/nri1086.html\nhttp://www.nature.com/nri/journal/v3/n5/pdf/nri1086.pdf.

Jeffroy, O. et al., 2006. Phylogenomics: the beginning of incongruence? *Trends in Genetics*, 22(4), pp.225–231.

Jolma, A. et al., 2013. DNA-binding specificities of human transcription factors. *Cell*, 152(1-2), pp.327–339. Available at: http://dx.doi.org/10.1016/j.cell.2012.12.009.

Kawaharada, Y. et al., 2015. Receptor-mediated exopolysaccharide perception controls bacterial infection. *Nature*. Available at: http://www.nature.com/doifinder/10.1038/nature14611.

Keller, N.P., Turner, G. & Bennett, J.W., 2005. Fungal secondary metabolism — from biochemistry to genomics. *Nature Reviews Microbiology*, 3(12), pp.937–947. Available at: http://www.nature.com/doifinder/10.1038/nrmicro1286.

Korotkov, K. V., Sandkvist, M. & Hol, W.G.J., 2012. The type II secretion system: biogenesis, molecular architecture and mechanism. *Nature Reviews*

Microbiology, 10(5), pp.336–351. Available at: http://dx.doi.org/10.1038/nrmicro2762.

Lackner, G., Moebius, N., Partida-Martinez, L.P., et al., 2011. Evolution of an endofungal lifestyle: Deductions from the Burkholderia rhizoxinica genome. *BMC genomics*, 12(1), p.210. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3102044&tool=pmcentrez&rendertype=abstract [Accessed March 1, 2013].

Lackner, G. et al., 2009. Global distribution and evolution of a Toxinogenic burkholderia-rhizopus symbiosis. *Applied and Environmental Microbiology*, 75(9), pp.2982–2986.

Lackner, G., Moebius, N. & Hertweck, C., 2011. Endofungal bacterium controls its host by an hrp type III secretion system. *The ISME journal*, 5(2), pp.252–61. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3105691&tool=pmcentrez&rendertype=abstract [Accessed March 1, 2013].

Lagesen, K. et al., 2007. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, 35(9), pp.3100–3108.

Langmead, B. & Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4), pp.357–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3322381&tool=pmcentrez&rendertype=abstract [Accessed July 10, 2014].

Lawrence, M., Gentleman, R. & Carey, V., 2009. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics*, 25(14), pp.1841–1842. Available at: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btp328.

Leone, M.R. et al., 2010. An unusual galactofuranose lipopolysaccharide that ensures the intracellular survival of toxin-producing bacteria in their fungal host. *Angewandte Chemie (International ed. in English)*, 49(41), pp.7476–80. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20718018 [Accessed August 15, 2013].

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), pp.2078–2079.

Li, T. et al., 2013. SET-domain bacterial effectors target heterochromatin protein 1 to activate host rDNA transcription. *EMBO reports*, 14(8), pp.733–40. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23797873.

López, V. et al., 2002. Occurrence of 20S RNA and 23S RNA replicons in industrial yeast strains and their variation under nutritional stress conditions. *Yeast (Chichester, England)*, 19(6), pp.545–52. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11921103\nhttp://www.ncbi.nlm.nih.gov/pubmed/11921103.

Ma, L.-J. et al., 2009. Genomic analysis of the basal lineage fungus Rhizopus oryzae reveals a whole-genome duplication. *PLoS genetics*, 5(7), p.e1000549. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2699053&tool=pmcentrez&rendertype=abstract [Accessed August 6, 2013].

Marchetti, M. et al., 2010. Experimental evolution of a plant pathogen into a legume symbiont. *PLoS Biology*, 8(1).

Márquez Luis M., Redman Regina S., Rodriguez Russell J., R.M.J., 2007. A Virus in a Fungus in a Plant: Three-Way Symbiosis Required for Thermal tolerance. *Science*, 315(April), pp.513–516.

Martin, W. & Kowallik, K., 1999. Annotated English translation of Mereschkowsky's 1905 paper "Über Natur und Ursprung der Chromatophoren imPflanzenreiche." *European Journal of Phycology*, 34(3), pp.287–295.

Marzluff, W.F., Wagner, E.J. & Duronio, R.J., 2008. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. *Nature Reviews Genetics*, 9(11), pp.843–854. Available at: http://www.nature.com/doifinder/10.1038/nrg2438.

McCarthy, D.J., Chen, Y. & Smyth, G.K., 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic acids research*, 40(10), pp.4288–97. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3378882&tool=pmcentrez&rendertype=abstract [Accessed July 10, 2014].

McCutcheon, J.P. & Moran, N. a., 2011. Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology*, 10(1), pp.13–26. Available at: http://dx.doi.org/10.1038/nrmicro2670.

van Melderen, L., 2010. Toxin–antitoxin systems: why so many, what for? *Current Opinion in Microbiology*, 13(6), pp.781–785. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1369527410001670.

Meyer, F., Overbeek, R. & Rodriguez, a., 2009. FIGfams: yet another set of protein families. *Nucleic Acids Research*, 37(20), pp.6643–6654. Available at: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkp698.

Mistry, J. et al., 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic acids research*, 41(12), p.e121. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3695513&tool=pmcentrez&rendertype=abstract [Accessed August 4, 2014].

Moebius, N. et al., 2014. Active invasion of bacteria into living fungal cells. , pp.1–20.

Morelle, W. et al., 2005. Galactomannoproteins of *Aspergillus fumigatus*. *Eukaryotic Cell*, 4(7), pp.1308–1316.

Mosse, B., 1970. Honey-Coloured, Sessile. *Arch. Mikrobiol*, 145.

Narechania, A. et al., 2012. Random Addition Concatenation Analysis: A novel approach to the exploration of phylogenomic signal reveals strong agreement between core and shell genomic partitions in the cyanobacteria. *Genome Biology and Evolution*, 4(1), pp.30–43.

Nazir, R., Tazetdinova, D.I. & van Elsas, J.D., 2014. Burkholderia terrae BS001 migrates proficiently with diverse fungal hosts through soil and provides protection from antifungal agents. *Frontiers in Microbiology*, 5(598), pp.1–10.

Newman, M.-A. et al., 2007. Priming, induction and modulation of plant defence responses by bacterial lipopolysaccharides. *Journal of endotoxin research*, 13(2), pp.69–84.

Nürnberger, T. et al., 2004. Innate immunity in plants and animals: Striking similarities and obvious differences. *Immunological Reviews*, 198, pp.249–266.

Nuss, D.L., 2005. Hypovirulence: Mycoviruses at the fungal–plant interface. *Nature Reviews Microbiology*, 3(8), pp.632–642. Available at: http://www.nature.com/doifinder/10.1038/nrmicro1206.

Ogura, T. & Hiraga, S., 1983. Mini-F plasmid genes that couple host cell division to plasmid proliferation. *Proceedings of the National Academy of Sciences of the United States of America*, 80(15), pp.4784–4788.

Pandey, D.P. & Gerdes, K., 2005. Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. *Nucleic Acids Research*, 33(3), pp.966–976.

Paradis, E., Claude, J. & Strimmer, K., 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2), pp.289–290.

Park, H.-S. & Yu, J.-H., 2012. Genetic control of asexual sporulation in filamentous fungi. *Current opinion in microbiology*, 15(6), pp.669–77. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23092920 [Accessed August 19, 2013].

Partida-Martinez, L.P., Groth, I., et al., 2007. Burkholderia rhizoxinica sp. nov. and Burkholderia endofungorum sp. nov., bacterial endosymbionts of the plant-pathogenic fungus Rhizopus microsporus. *International journal of systematic and evolutionary microbiology*, 57(Pt 11), pp.2583–90. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17978222 [Accessed August 8, 2013].

Partida-Martinez, L.P., Monajembashi, S., et al., 2007. Endosymbiont-dependent host reproduction maintains bacterial-fungal mutualism. *Current biology : CB*, 17(9), pp.773–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17412585 [Accessed May 31, 2013].

Partida-Martinez, L.P. & Hertweck, C., 2007. A Gene Cluster Encoding Rhizoxin Biosynthesis in “Burkholderia rhizoxina”, the Bacterial Endosymbiont of the FungusRhizopus microsporus. *ChemBioChem*, 8(1), pp.41–45. Available at: http://doi.wiley.com/10.1002/cbic.200600393.

Partida-Martinez, L.P. & Hertweck, C., 2005. Pathogenic fungus harbours endosymbiotic bacteria for toxin production. *Nature*, 437(7060), pp.884–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16208371 [Accessed February 28, 2013].

Polak, M., 1996. Ectoparasitic Effects on Host Survival and Reproduction: The Drosophila-- Macrocheles Association. *Ecology*, 77(5), p.1379. Available at: http://www.jstor.org/stable/2265535?origin=crossref.

Quast, C. et al., 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research*, 41(Database issue), pp.D590–6. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531112&tool=pmcentrez&rendertype=abstract [Accessed July 10, 2014].

Richardson, M., 2009. The ecology of the zygomycetes and its impact on environmental exposure. *Clinical Microbiology and Infection*, 15(SUPPL. 5), pp.2–9. Available at: http://dx.doi.org/10.1111/j.1469-0691.2009.02972.x.

Ritchie, M.E. et al., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), pp.e47–e47. Available at: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv007.

Robinson, M.D., McCarthy, D.J. & Smyth, G.K., 2009. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), pp.139–140.

Robinson, M.D. & Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology*, 11(3), p.R25. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2864565&tool=pmcentrez&rendertype=abstract.

Rohm, B. et al., 2010. Toxin production by bacterial endosymbionts of a Rhizopus microsporus strain used for tempe/sufu processing. *International Journal of Food Microbiology*, 136(3), pp.368–371. Available at: http://dx.doi.org/10.1016/j.ijfoodmicro.2009.10.010.

Ronquist, F. et al., 2012. Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3), pp.539–542.

Ruiz-Herrera, J. & Ortiz-Castellanos, L., 2010. Analysis of the phylogenetic relationships and evolution of the cell walls from yeasts and fungi. *FEMS Yeast Research*, 10(3), pp.225–243.

Saisongkorh, W. et al., 2010. Evidence of transfer by conjugation of type IV secretion system genes between bartonella species and rhizobium radiobacter in amoeba. *PLoS ONE*, 5(9), pp.1–14.

Salvioli, A. et al., 2010. Endobacteria affect the metabolic profile of their host Gigaspora margarita, an arbuscular mycorrhizal fungus. *Environmental Microbiology*, 12(8), pp.2083–2095.

Scherlach, K. et al., 2006. Antimitotic rhizoxin derivatives from a cultured bacterial endosymbiont of the rice pathogenic fungus Rhizopus microsporus. *Journal of the American Chemical Society*, 128(10), pp.11529–11536.

Scherlach, K. et al., 2013. Biosynthesis and Mass Spectrometric Imaging of Tolaasin, the Virulence Factor of Brown Blotch Mushroom Disease. *ChemBioChem*, 14(18), pp.2439–2443. Available at: http://doi.wiley.com/10.1002/cbic.201300553.

Schmitt, I. et al., 2008. Evolution of host resistance in a toxin-producing bacterial-fungal alliance. *The ISME journal*, 2(6), pp.632–41. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18309361 [Accessed February 27, 2013].

Schmitt, M.J. & Breinig, F., 2006. Yeast viral killer toxins: lethality and self-protection. *Nature Reviews Microbiology*, 4(3), pp.212–221. Available at: http://www.nature.com/doifinder/10.1038/nrmicro1347.

Schwartz, R.M. & Dayhoff, M.O., 1978. Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. *Science (New York, N.Y.)*, 199(4327), pp.395–403.

Shilatifard, A., 2008. Molecular implementation and physiological roles for histone H3 lysine 4 (H3K4) methylation. *Current Opinion in Cell Biology*, 20(3), pp.341–348.

Soanes, D.M. & Talbot, N.J., 2010. Comparative genome analysis reveals an absence of leucine-rich repeat pattern-recognition receptor proteins in the kingdom fungi. *PLoS ONE*, 5(9), pp.1–10.

Socolovschi, C., Audoly, G. & Raoult, D., 2013. Connection of toxin-antitoxin modules to inoculation eschar and arthropod vertical transmission in Rickettsiales. *Comparative Immunology, Microbiology and Infectious Diseases*, 36(2), pp.199–209. Available at: http://dx.doi.org/10.1016/j.cimid.2013.01.001.

Stopnisek, N. et al., 2015. Molecular mechanisms underlying the close association between soil Burkholderia and fungi. *The ISME Journal*, pp.1–12. Available at: http://dx.doi.org/10.1038/ismej.2015.73.

Sublines, V. et al., 1986. Rhizoxin , a Macrocyclic Lactone Antibiotic , as a New Antitumor Agent against Human and Murine Tumor Cells and Their Rhizoxin , a Macrocyclic Lactone Antibiotic , as a New Antitumor Agent against. , 46(January), pp.381–385.

Suzuki, M.M. & Bird, A., 2008. DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics*, 9(6), pp.465–476. Available at: http://www.nature.com/doifinder/10.1038/nrg2341.

Traxler, M.F. & Kolter, R., 2015. Natural products in soil microbe interactions and evolution. *Nat. Prod. Rep.*, 32, pp.956–970. Available at: http://xlink.rsc.org/?DOI=C5NP00013K.

Wang, D. et al., 2013. Draft Genome Sequence of Rhizopus chinensis

CCTCCM201021 , Used. , 1(2), pp.1–2.

Weber, T. et al., 2015. antiSMASH 3.0--a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Research*, 43(May), pp.237–243. Available at: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv437.

Wen, Y., Behiels, E. & Devreese, B., 2014. Toxin-Antitoxin systems: their role in persistence, biofilm formation, and pathogenicity. *Pathogens and disease*, 70(3), pp.240–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24478112.

Wernegreen, J.J., 2012. Endosymbiosis. *Current Biology*, 22(14), pp.555–561.

Westermann, A.J., Gorski, S. a & Vogel, J., 2012. Dual RNA-seq of pathogen and host. *Nature reviews. Microbiology*, 10(9), pp.618–30. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22890146 [Accessed November 7, 2013].

Xu, S. et al., 2014. Efficient transformation of Rhizopus delemar by electroporation of germinated spores. *Journal of Microbiological Methods*, 103, pp.58–63. Available at: http://dx.doi.org/10.1016/j.mimet.2014.05.016.

Xu, X.L. et al., 2008. Bacterial Peptidoglycan Triggers Candida albicans Hyphal Growth by Directly Activating the Adenylyl Cyclase Cyr1p. *Cell Host and Microbe*, 4(1), pp.28–39.

Young, M.D. et al., 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome biology*, 11(2), p.R14.

José Roberto Bermúdez Barrientos

Born in August 18th, 1989. Mexican. Civil Status: Single

T: +52 477 1129137 E: jbermudez@langebio.cinvestav.mx

Bioinformatics Skills

Three years of experience working in Bioinformatic projects. Confortable in Unix operative systems and the command line interface. Coding in Perl and R languages. Accustomed to handle high throughput data. Experience using cluster computing. Analyzed Illumina RNA-Seq data to identify gene expression and discover differentially expressed genes. Reference based genotypic reconstructions (SNPs calling) using Illumina data. Familiarized with tools such as fastQC, bowtie2, samtools, HTSeq and R bioconductor.

Working Experience

Research Assistant at RNA Computational Genomics Lab [2012]. Under leadership of PhD Cei Abreu-Goodger. National Laboratory of Genomics for Biodiversity (LANGEBIO), CINVESTAV Irapuato, Mexico.

Teaching Experience

An introduction of RNA-Seq data analysis BSc level [2014].
Eight hours course with both theory and hands on practices. Given at ENES Leon, Mexico. From the National Autonomous University of Mexico. Part of the Bioinformatics subject.

Formation

MSc in Integrative Biology at LANGEBIO CINVESTAV, Irapuato, Mexico. With an expected degree date of August 2015.

BSc in Genomics Sciences [2012], graduated with honors from the National Autonomous University of Mexico (UNAM).

Publications

Vences-Guzmán MA, Guan Z, Bermúdez-Barrientos JR, Geiger O, Sohlenkamp C. 2012. Agrobacteria lacking ornithine lipids induce more rapid tumour formation. Environmental Microbiology. n/a, 895-90.

Vences-Guzmán MA, Guan Z, Escobedo-Hinojosa WI, Bermúdez-Barrientos JR, Geiger O, Sohlenkamp C. 2014. Discovery of a bifunctional acyltransferase responsible for ornithine lipid synthesis in Serratia proteamaculans. Environmental Microbiology. 1462-2920.

Languages

Spanish (mother language)
English, fluently

References

PhD Cei Abreu-Goodger. MSc thesis advisor.
cei@langebio.cinvestav.mx
PhD Laila Partida-Martínez. MSc thesis advisor.
laila.partida@ira.cinvestav.mx
PhD Christian Sohlenkamp. BSc thesis advisor.
chsohlen@ccg.unam.mx

Declaración de independencia

Por este medio declaro que yo he preparado este trabajo de tesis de forma independiente y sin ayuda externa. Especialmente declaro que he citado de forma correcta y explícita a los autores y trabajos en los que esta tesis se apoya, así como las contribuciones de las personas que coadyuvaron en su desarrollo.

Lugar: Irapuato, Guanajuato, México

Fecha: 28 de Octubre de 2015 Firma: