FUNGAL PHYLOGENOMICS. A GLOBAL ANALYSIS OF FUNGAL GENOMES AND THEIR EVOLUTION

Marina Marcet Houben

ISBN: 978-84-693-6431-4
Dipòsit Legal: T-1633-2010

Marina MARCET HOUBEN

FUNGAL PHYLOGENOMICS

A GLOBAL ANALYSIS OF FUNGAL GENOMES AND THEIR EVOLUTION

DOCTORAL THESIS

Supervised by Dr. Juan Antonio Gabaldón Estevan

Department of Biochemistry and Biotechnology

UNIVERSITAT ROVIRA I VIRGILI

July - 2010

Tarragona

UNIVERSITAT ROVIRA I VIRGILI
Departament de Bioquímica i Biotecnologia
c/ Marcel·lí Domingo s/n
Campus Sant Pere Sescelades
43007 Tarragona
Telèfon: 977 559 521
Fax: 977 558 232

I, Juan Antonio Gabaldón Estevan, group leader of the Comparative genomics group, in the Department of Bioinformatics and Genomics of the Center for Genomic regulation,

CERTIFY:

That the present study, entitled "Fungal Phylogenomics. A global analysis of fungal genomes and their evolution", presented by Marina Marcet Houben for the award of the degree of Doctor has been carried out under my supervision at the Department of Bioinformatics of the Centro de Investigación Principe Felipe, València, and the Department of Bioinformatics and Genomics of the Center for Genomic Regulation, Barcelona, and that it fulfils all the requirements to be eligible for the European Doctorate Label.

Barcelona, 1th of June 2010

AGRAÏMENTS

Estic davant de l'ordinador, com cada dia, escrivint el que serà l'última part d'aquesta tesis i casi no em puc creure que ja estigui. En aquest moment miro enrere i no puc evitar recordar moltíssims moments, bons i dolents, que han sigut part d'aquests anys d'aprenentatge. Ja us aviso per avançat que aquests agraïments seran llargs, ja que han sigut molts llocs on he treballat i molta gent a la que he conegut i a la que ara m'agradaria recordar.

Abans que res vull dedicar unes paraules al Toni, la persona que ha estat el meu supervisor, mentor i amic durant aquests últims anys. Hi han moltes característiques que fan del Toni un dels millors supervisors que un doctorant pot tenir. Des de la seva visió científica, al seu caràcter obert i alegre, i les seves idees i ganes d'innovar, ha sigut en moltíssimes ocasions una font d'inspiració i de suport. Moltes gràcies Toni per estar sempre aquí, per ajudar en els moments bons i especialment en els moments difícils, per donar-me a mi, i sospito que a tots els meus companys, la seguretat de saber que passi el que passi estas amb nosaltres i que aquesta feina és dels dos.

Tot i que en aquest volum només es recull la meva feina durant els últims tres anys, no crec que pugui fer uns agraïments complerts sense incloure els meus primers passos en aquest meravellós camp de la genòmica i la bioinformàtica. Us convido ara a que m'acompanyeu en un viatge en el temps, i recordeu amb mi aquells primers dies on fer un arbre filogenètic o un senzill script en perl era feina de dies i no de minuts.

Tarragona, Maig 2005- Agost 2006.

Si hi havia una cosa que tenia molt clara quan vaig acabar la carrera era que em volia dedicar a la bioinformàtica. Amb aquesta idea en ment va començar la meva aventura en el grup de genòmica comparativa de la URV. Moltes gràcies Santi per guiar-me durant tot el temps que vaig estar allí i fins i tot ara, quan ja fa anys que no estic a la URV, gràcies per fer-me de tutor i interessar-te per la meva feina.

Quan vaig entrar en el laboratori de bioinformàtica ja hi havia allí un grup molt ben avingut (i heu de reconèixer tots que una mica esbojarrat) que em va fer sentir benvinguda des del primer moment. Recordo moltíssims moments plens de rialles i bon humor, si és que amb persones com l'Albert, el Pere, l'Esther i la Montserrat, com pots esperar altra cosa? Gràcies Pere per les moltes discussions que hem tingut durant els anys, i compte, que un dia acabaràs trencant una taula!! Albert, crec que només t'he de dir una cosa: "Visca el Barça!!". Montserrat, moltes gràcies per ser una gran amiga i meravellosa persona, en la que puc confiar i se que sempre estarà allí. Esther, no canviïs mai, la teva capacitat per alegrar el dia a la gent es increïble... i aixo ho escric cinc minuts després de que m'hagis fet riure en mig del laboratori amb un missatge pel gtalk!! També vull agrair a tots els demés membres del lab (o encara és labo?) de bioinfo, Gerard, Pep, Eduard, moltes gracies pels bons moments que vam passar. Laura, tot i que no vam arribar a coincidir en el laboratori, moltes gracies pels molts sopars, festes i demés que hem compartit. Ja em perdonareu els respectius per posar-vos aquí, amb els "frikis". Gràcies Lidia per la teva franquesa i amistat, i Jordi Puxeu per la teva paciència i simpatia.

També vull recordar aquí tots als demés membres del departament de bioquímica, no els anomenaré a tots, ja que segur que me'n deixaria. De totes maneres hi ha algunes persones amb les que he passat molt bons moments. Per exemple aquells que vaig conèixer abans fins i tot de començar el doctorat. Recordes Gemma totes aquelles pràctiques de bioquímica? Sembla que a tu et van agradar més que a mi!! Si més no, moltes gràcies pels molts anys d'amistat i companyerisme, un d'aquests dies hem de fer un dinar virtual, pero aquest cop amb web-cam, sinó no té gràcia! David! Seguramente la persona dentro de estos agradecimientos a la que más tiempo hace que conozco. Cuanto nos reimos durante la carrera! Y que lástima que ya no estuviera en Tarragona cuando empezaste el doctorado, pero bueno, igualmente creo que lo hemos pasado muy bien! Sabina, Helena, Anabel, Montse, i molts dels altres dels experimentals, moltes gràcies pels bons moments i ànims per aquells que acabeu!

Si en una cosa s'ha caracteritzat la meva tesis ha sigut en els canvis. El primer potser no va ser molt important. De la facultat de química a la facultat de medicina. De Tarragona a Reus. De bioinformàtica... bé, a bioinformàtica, que us pensaveu? Tot i que estava camuflada en un laboratori experimental!

Reus, Septembre 2006- Març 2007

Reus va ser el lloc on vaig estar menys temps, pero no per això vaig deixar de coneixer a gent important alli. Vull agrair als que van ser els meus directors, l'Anton Romeu i José Luís Paternain, l'oportunitat que em van donar de treballar allí. També agrair als membres del departament el temps que vaig passar amb ells.

Però si amb una persona vaig compartir, i segueixo compartint ara, una amistat especial va ser amb la Tania. Quantes hores vam passar les dues, en un laboratori preparat per allotjar com a minim sis persones, ella amb els seus ratolins i jo amb el meu ordinador. Sempre pots esperar de la Tania que t'aparegui amb un comentari de lo més inversemblant, i no pots evitar pensar: en quin món he anat a parar ara? Moltes gràcies Tania per la teva companyia i els bons (i mals) moments que vam passar a Reus. També recordar a Espe i a la Marta, de fàrmaco.

Hi han moments en que penses: aquesta decisió pot canviar la meva vida. I en molts aspectes així va ser quan vaig deixar enrera Reus per anar a València. Però molts cops els canvis son bons i pots mirar enrera sense lamentar-te.

València, Abril 2007 - Agost 2008

Aquí es donde empezó el trabajo que presento en esta tesis. Durante mi estancia en el departamento de Bioinformática del CIPF conocí a muchísimas personas, y no es para menos dado el tamaño del departamento (debíamos oscilar entre los 20-30 bioinformáticos). Muchos son los que compartieron horas de trabajo conmigo y quiero agradecerles a todos el tiempo que pasé allí. Especialmente quiero recordar a Paco, que con su buen humor era capaz de alegrar el día a cualquiera y a Eva, por su caracter directo y simpático. También a la "invasión" de italianos: Giuseppe, Giulia, Daniela y Mariella, con los que compartí muchos momentos divertidos y a los que recuerdo con mucho cariño.

Y no, Salva y Jaime, no me olvido de vosotros (como podría?) pero puesto que me acompañasteis en la última etapa de mi pequeño viaje os invito ahora a venir a nuestro destino final.

Barcelona, Septembre 2008 - ?

És justament des d'aquest laboratori des d'on estic escrivint aquests agraiments, amb Salva i Diego treballant al meu darrera i el Toni sortint i entrant del seu despatx.

Muchas gracias Salva, por ser un buen compañero y maravilloso amigo. Recuerdas ese primer dia en Barcelona, tu en un hotel de mala muerte y yo en un piso vacío, cuando cogimos y nos fuimos a cenar por allí a pesar de que apenas nos conocíamos. Y las muchas veces que se ha repetido eso durante el último año y medio, donde hemos compartido trabajo y diversiones en miles de momentos.

Jaime, que pasa tronco? Que bien lo hemos pasado, tanto aquí como en Valencia. Muchas gracias por iniciarme en el mundo de los filomas y enseñarme muchas de las cosas que ahora se.

I des d'aquells primers moments el grup ha anat creixent: Javi, Diego (no se si llegarás a leer esto puesto que nos dejas en unos dias, pero muchas gracias por todo) i Leszek (thanks for all the fun moments, and for the ones that I hope will come in the future!).

I would also like to thank all the members of the FunPath consortium, some of which I have only met at the FunPath meetings (thanks, they were very interesting) and some of them which I know a little bit better. Thanks Karl, Walter and Tobby for the month I spent in Viena, it was very insightful.

A special thanks goes to Christophe and all the members of his group for the four months I spent in Paris, I had a great time even though it was freezing cold! I will always remember fondly how Christophe would come into the lab saying: “I have an idea”, it still makes me laugh even now. Thanks Ute for the many days we spent wandering around Paris, we’ll have to repeat it some time.

I aquí arribo, al final d’aquesta aventura anomenada tesis, i mirant enrera veig que a través de tots els canvis hi han constants, persones que sempre han estat amb mi, i és a ells als que dedico aquestes últimes paraules. A la meva germana, que encara que no ho sàpiga va ser un dels factors que em va animar a anar a València i que tot i la distància que ens separa ara, continua sent per mi una persona molt important. Al Vicent, que tot i a penes coneixe’m em va ajudar moltíssim durant tot el temps que vaig passar a València.

I molt especialment agraeixo als meus pares tot el seu suport i els seus consells. Per estar sempre allí quan els necessito, durant els moments bons i els dolents. Que al final ja ni s’inmutaven quan començava una conversa amb la frase: “Crec que en uns mesos me’n vaig a...”. Gràcies per l’interès que sempre heu mostrat en tot el que faig, encara que a vegades us soni a xino! Pares, moltes gràcies per tot, us estimo.

Cada amigo representa un mundo dentro de nosotros, un mundo que tal vez no habría nacido si no lo hubiéramos conocido.

Harcourt Brace

Als meus pares

CONTENTS

THESIS OUTLINE

This PhD thesis focusses on the use of phylogenomics to elucidate the evolution of fungal species. Here I outline the different parts that can be found in this volume.

The introductory part of the thesis is formed by **Chapter 1** and **Chapter 2**. In Chapter 1, an overview of the current knowledge on fungi is displayed with an emphasis on fungal evolution. The second chapter focuses on the main phylogenomic methods that have been applied throughout the present thesis. These two chapters are not essential for the understanding of the thesis but will provide sufficient background information for readers that are not familiar with some of the topics treated in this thesis.

Chapter 3 and **Chapter 4** are focused on the analysis of species trees. Chapter 3 presents an analysis about the robustness of the fungal species tree. This analysis is based on the use of extended phylogenetic methods to identify nodes in the fungal species tree that are in conflict with gene trees and that can therefore be source of errors in posterior analyses. The yeast phylome was used as the source of gene trees for this comparison. In addition, a comparison between two phylogeny-based orthology prediction methods (tree reconciliation and species overlap) is presented in this chapter. In Chapter 4, treeKO will be presented. TreeKO is a tree comparison program that, contrary to existing software, has been specifically designed for the comparison of multi-gene trees, and is therefore able to deal with pairs of trees that have different species content and that contain multiple duplications.

Chapter 5 and **Chapter 6** deal with two phylogenomic analyses concerning fungal species. Chapter 5 presents a high throughput search for inter-kingdom horizontal gene transfer events in 60 fully sequenced fungal genomes. We were able to identify more than 235 cases of horizontal gene transfer events, involving more than 700 genes, from prokaryotes to eukaryotes using an automatic pipeline. Additionally we observed that these events were unevenly distributed through the different fungal groups. Chapter 6 focuses on the evolution of the oxidative phosphorylation pathway in fungi. The analysis identified many duplications that occurred during the evolution of this metabolic pathway in fungi. Both chapters focus on events that may influence the gene tree topology and render it different from the species tree.

Chapter 7 is dedicated to the work done within the FunPath Consortium. The FunPath is european consortium created to elucidate the virulence mechanisms of the human fungal pathogen, *Candida glabrata*. The chapter also contains some of the work done in collaboration with Christophe d'Enfert's lab, which is part of the FunPath consortium, and will deal with the application of tree reconstruction methods using SNPs detected in several newly sequenced strains of *Candida albicans*.

The last chapter, **Chapter 8**, provides a summarizing discussion of the main points treated during this thesis.

Part I

Introduction

CHAPTER 1
THE FUNGI: OVERVIEW OF AN ESSENTIALLY UNEXPLORED KINGDOM.

Introduction

Fungi is one of the eukaryotic groups with the highest number of fully sequenced genomes. This kingdom comprises a large diversity of species, including mushrooms, yeasts, molds, rusts, smuts, puffballs, truffles, morsels and other less well-known organisms. The exact number of fungal species is unknown but estimates set this value around 1.5 million (Hawksworth, 1991) being 700.000 a conservative, lower estimate (Schmit and Mueller, 2007). Currently, only 70,000 species have been described, thus representing a mere 5-10% of the total diversity (Mueller and Schmit, 2007). While fossil records are scarce, data suggests that fungi already existed around 400 to 500 million years ago, though their exact origins may be more ancestral (Taylor and Berbee, 2006).

Fungi are ubiquitous and are able to colonize a broad diversity of habitats, including extreme environments such as deserts and deep sea sediments. Although a higher diversity of fungal species is found in tropical regions, the optimal growth conditions tend to vary across different species making fungi adaptable to numerous environments. As seen in figure 1.1, fungal species can adopt multiple morphologies.

One of the main roles performed by fungi within their ecosystems is that of decomposers of organic matter, thus being an important part of the carbon cycle. Fungi decompose biopolymers in order to absorb their constituents as nutrients. They usually live on their own food source and are able to grow into new sources when their first location becomes insufficient. When a food source is depleted, sporulation is triggered.

Many fungi display both sexual and asexual reproductive mechanisms. In sexual reproduction, sex determination is controlled by a small genomic region. In these regions specific gene combinations account for different mating types (Fraser and Heitman, 2003). Despite initial thoughts that the use of mating-type loci was limited to Ascomycota and Basidiomycota, recent studies have identified homologous regions to these loci in Zygomycota and Microsporidia (Idnurm *et al.*, 2008; Lee *et al.*, 2009). This supports the hypothesis that sex

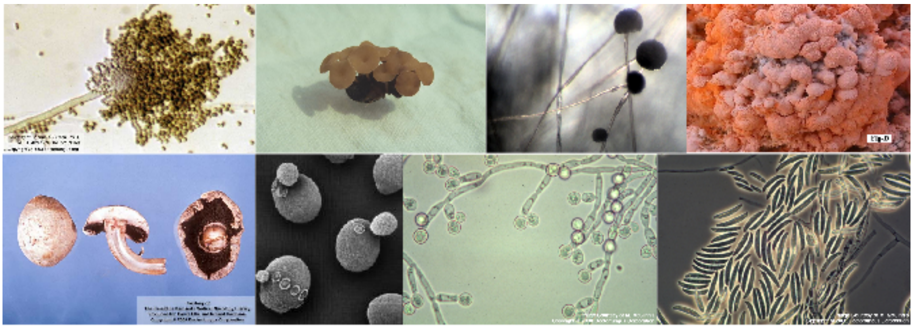

Figure 1.1: Example of fungal species. Images were taken from the Dr. Fungus webpage (http://www.doctorfungus.org/). The images represent, at the top, from left to right: *Aspergillus flavus*, *Sclerotinia sclerotiorum*, *Rhizopus oryzae* and *Neurospora crassa*. At the bottom, from left to right: *Agaricus bisporus*, *Saccharomyces cerevisiae*, *Candida albicans*, *Fusarium oxysporum*.

genes could have been a trait found in ancient fungi, which subsequently got lost in some lineages (Idnurm *et al.*, 2008). Fungi can also reproduce asexually. Yeast species, for instance, preferentially reproduce through budding, and in some species the sexual cycle has never been observed. For instance, *Candida glabrata* possesses all the genes that are known to be involved in sexual reproduction in yeast (Wong *et al.*, 2003), but only asexual reproduction of this species has been observed so far.

Fungal species can adopt numerous different life styles. Many of the known fungi are free-living saprobes, which obtain nutrients and other molecules from dead or decaying organic material in woods, soils, leafs, dead animals or animal excrements. Other species are symbionts that live in direct association with plants, animals or prokaryotes (i.e. lichens, mycorrhizae or endophytes). A large number of pathogenic species can also be found within the fungi. They possess different mechanisms that enable them to infect numerous animals and plants. While not primarily pathogenic, other fungi are able to synthesise secondary metabolites that can be toxic or carcinogenic, this is the case of aflatoxins secreted by Aspergillus species. While some fungal species cause important economical damage due to their pathogenesis of humans and crops, some others are beneficial and are exploited by biotechnological industries. For instance, *Saccharomyces cerevisiae* and *Kluyveromyces lactis* are used in the food industry while other species, such as *Yarrowia lipolytica,* are used in the production of antibiotics and other chemical compounds.

Some fungal species have been extensively studied and are considered model organisms. Probably the foremost species in this sense is the baker's yeast, *S. cerevisiae*. Indeed, this yeast is the first eukaryotic species for which a complete nuclear genome sequence was obtained (Goffeau *et al.*, 1996), and many efforts are still being done in order to fully comprehend the mechanisms that rule its biological processes. Other model organisms include the ascomycetes *Neu-*

rospora crassa, *Aspergillus nidulans* and *Schizosaccharomyces pombe*, as well as the basidiomycota *Ustilago maydis*. Results found in these species concerning genetics, physiological or morphological processes, are often extended to other fungal organisms, and are also the main source of functional information used for the annotation of newly sequenced genomes.

Fungal systematics: what do we know?

The advent of the genomic era led to many important discoveries regarding fungal evolution, which caused a complete revamping of the evolutionary history of these species. Morphological data was the main basis of taxonomic classification before the use of DNA and many of its postulates have been discarded since then. Probably the most relevant change was the discovery that fungi are more closely related to animals than to plants, to which they had been traditionally associated. The monophyly of fungi and metazoans within a clade known as opistokhonta is now strongly supported by both, phylogenetic and phylogenomic data (Wainright *et al.*, 1993; Bruns *et al.*, 1992; Lang *et al.*, 2002; Steenkamp *et al.*, 2006). Another important discovery was the fact that the group of Fungi, as described by morphological data was in fact polyphyletic and that the similarities in morphological characters found between the different groups were due to convergent evolution rather than because of a common ancestry. This led to the division between the “true fungi” and two other groups which included slime molds and oomycetes. Nowadays, the “true fungi” group holds the rank of Kingdom within the eukaryotic taxonomy, at the same level of animals and plants, being a sister group to the former. The phylogenetic position within the eukaryotes of the remaining two groups derived from the primary fungal group is not well defined yet, although slime molds are generally considered a sister branch of opisthokonts, forming a larger group known as unikonta (Keeling *et al.*, 2005).

The fungal kingdom can be further divided into several groups. Probably the broadest division is the one that separates fungi into the early diverging groups and the dikarya. The so called early diverging fungi traditionally included Zygomycota and Chitridiomycota. Nowadays these two groups are thought to be polyphyletic. Two new groups were also included with the advent of the genomic era. Genomic data precipitated the inclusion of Microsporidia into the fungi (Keeling *et al.*, 2000) and the separation of Glomeromycota from Zygomycota. The exact positions of these two last groups with respect to the others is still not clear. Microsporidia tend to adopt a basal position in phylogenetic analysis but it is unknown whether this results from methodological artifacts caused by their highly diverging sequences. The exact position of Glomeromycota in reference to Zygomycota and Dykaria is also still controversial (Redecker and Raab, 2006).

Dikarya is the fungal subkingdom that contains Ascomycota and Basidiomycota, the two best studied fungal groups. These have traditionally been considered sister groups based on their morphological characteristics (i.e. they have regularly septate hyphae and a dikaryotic stage in their life cycle), an association that has also been supported by numerous phylogenetic and phylogenomic studies (James *et al.*, 2006).

Phylogenomics and the fungal species tree:

Since the start of the genomic era an effort has been made in order to derive fully resolved phylogenies, including as many of the known species as possible. Methods for tree concatenation and super-tree construction are nowadays used in an attempt to elucidate the evolution of a growing number of species (Wolf *et al.*, 2001; Snel *et al.*, 2005). The fungal kingdom has been the focus of many phylogenetic studies for several reasons. One of the most important is that this kingdom is the best sampled eukaryotic group in terms of fully sequenced genomes. The reason for this can be found in the inherently smaller size of their genomes and their importance for human health and industry. The large amount of available sequence data has opened the doors for the use of phylogenomics to address the reconstruction of the still elusive fungal species tree, or Fungal Tree of Life. The difficulties in rendering the fungal species tree reside in part in the lack of knowledge regarding most of the fungal diversity. As mentioned above, only 5% of the estimated total of fungal species has been described so far. An even smaller percentage has information on the usual genomic markers while around 100 genomes have been sequenced, which represents 0,006% of the estimated number of fungal species. In light of this, any attempt at taxonomic classification will be tampered with the knowledge that it is limited to, and biased towards, our current awareness on fungi, and may change as more species are discovered.

Several species trees have been reconstructed over the last few years and many changes have been introduced to the previous, morphology based, taxonomic classification. Some of the largest analyses, in terms of number of species, include the work of (Lutzoni *et al.*, 2004) and (James *et al.*, 2006). The first one reconstructed several species trees using different loci. The largest of these trees, with 558 taxa, is based on two loci (nucSSU and nucLSU rDNA), whereas the smallest, including 103 species, was built on information provided by four loci. Representatives for the main known fungal groups were included. The tree presented by James et al., is based on 6 loci and shows the phylogenetic relationships between 170 species. This tree, unlike the one presented by Lutzoni et al., is fully resolved and offers insights into some problematic nodes such as the position of microsporidia or the lack of monophyly in Chitrids. The use of completely sequenced genomes in phylogenetic reconstructions allows for the use of larger amounts of data from a limited number of species which is often biased towards Ascomycotina. Even so, several large phylogenies have been reconstructed using data from fully sequenced genomes (Marcet-Houben and Gabaldón, 2009; Wang *et al.*, 2009; Fitzpatrick *et al.*, 2006). These trees, while slightly different in some nodes, mostly support a single topology (see chapter 3).

In Figure 1.2, we present the largest tree that has been reconstructed so far based on phylogenomics data. The tree represents the evolution of 102 different fungal species. It was built by concatenating 47 wide-spread proteins that displayed one-to-one orthology relationships and that were present in at least 90 species. Four outgroups were included: *Takifugu rubripes*, *Pediculus humanus*, *Drosophila melanogaster* and *Phytophtora infestans*. The outgroups were selected by searching for the candidate proteins in a database made of 206 completely sequenced eukaryotes using a Smith-Waterman search (Altschul *et al.*, 1997), we chose those outgroups that contained hits in most proteins and minimized the number of duplications.

The resulting tree is similar to the ones published before (Fitzpatrick *et al.*, 2006; Marcet-Houben and Gabaldón, 2009; Robbertse *et al.*, 2006; Wang *et al.*, 2009). Microsporidia, represented by four species here, occupy a basal position in fungi with a long branch that shows how divergent these species are compared to the other fungal species. Blastocladiomycetes, which was untill recently considered a Chitrid, does appear as an independent group in our tree, having diverged before the speciation of Chitrids. Even so, the low number of fully sequenced species representing these groups makes it difficult to infer any robust conclusions on their evolution. This is reflected in the many clades that result from the analysis performed by James et al. James *et al.* (2006) where Chitrids, Zygomycota, Blastocladiales and Microsporidia do not present clear monophyletic separations. We hope that in the future, new genomes will be sequenced within the early divergent fungi and that the relationships between them will be clearly elucidated.

There are three main groups recognized in Basidiomycota: Pucciniomycotina, Ustilagomycotina and Agaromycotina, though their exact phylogenetic classification is still uncertain. The most supported hypothesis places Agaromycotina and Ustilagomycotina as sister groups (Lutzoni *et al.*, 2004; James *et al.*, 2006; Marcet-Houben and Gabaldón, 2009). In contrast with this hypothesis, Ustilagomycotina are basal in Basidiomycota in our current tree. While aLRT support is high for this node, in chapter 3 we show that this position has a low phylome support and that this lack of support is often translated in difficulties in establishing unequivocally how a group of species has evolved.

Another debated distribution is that of the four Pezizomycotina groups. It is well established that Leotiomycetes and Sordariomycetes are sister groups

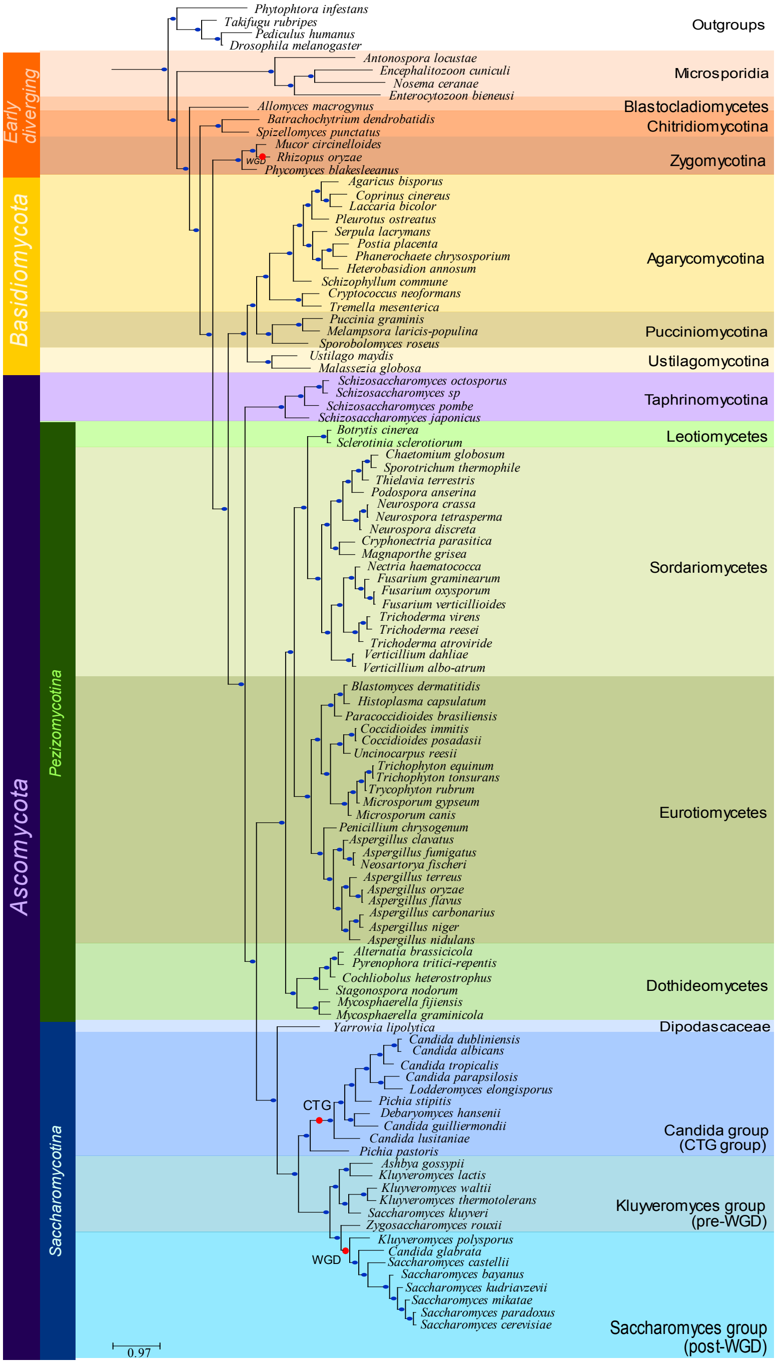

Figure 1.2: Fungal species tree containing 102 completely sequenced fungal species and four outgroups. The tree was obtained by concatenating 47 wide-spread single copy proteins.

but no consensus has been reached regarding the positions of Eurotiomycetes and Dothideomycetes. James et al. for instance presented Dothideomycetes species in a basal position for Pezizomycotina. In a concatenated tree presented by Fitzpatrick et al. (Fitzpatrick *et al.*, 2006) Eurotiomycetes were basal to Pezizomycotina while trees reconstructed using the super tree methodology presented Dothideomycetes and Eurotiomycetes as sister groups. This last distribution has also been found in other trees (Wang *et al.*, 2009; Marcet-Houben and Gabaldón, 2009). In our current tree, the first hypothesis, in which Dothideomycetes is the first diverging group of species, is supported. In recent years numerous genomes belonging to Dothideomycetes species have been sequenced, increasing the representation of organisms within this clade from one to six, this may confer more reliability to the phylogeny represented in figure 1.2 as it is made with a larger representation of species within the clade of interest. Even so, sequencing members of other, unrepresented, clades in Pezizomycotina may offer further resolution to this matter.

Yet another discussed node in the phylogeny of fungi is the relative positions of *C. glabrata* and *S. castellii*. We will address this matter further in chapter 4, but the distribution found in this tree is in line with most other phylogenetic methods where *C. glabrata* appears to have diverged from the Saccharomyces clade before *S. castellii*, which is in direct contradiction to data inferred from gene order conservation (Gordon *et al.*, 2009).

Completely sequenced fungi

The early diverging species

Very few species of fungi belonging to the early diverging groups, or lower fungi, have been sequenced so far. For several years, the only available genomes have been the microsporidian *Encephalitozoon cuniculi*, the two zygomycota *Rhizopus oryzae* and *Phycomyces blakesleeanus* and the chitrid, *Batrachochytrium dendrobatidis*. Interest in the exact phylogenetic location of Microsporidia and the project centered on the Origin of Multicellularity conducted by the Broad Institute (http://www.broadinstitute.org/) have attracted the attention of the scientific community to this group of species, resulting in a slow but steady growth in the number of sequenced genomes available.

Microsporidia are obligate, intracellular parasites that affect a wide variety of animals, ciliates and some apicomplexa. They lack typical mitochondria which is one of the reasons they were thought to be ancient eukaryotes pre-dating the mitochondrial endosymbiosis. However, it is now recognized that they possess highly derived forms of mitochondria (mitosomes), which are the result of secondary adaptations to anaerobic environments (Gabaldón and Huynen, 2004). Their adaptation to parasitic life-style has difficulted the elucidation of their phylogenetic position. Two studies involving alpha- and beta-tubulins placed the Microsporidia within the Fungal group (Edlind *et al.*, 1996; Keeling and Doolittle, 1996). Since then, additional evidence has been found that supports this relationship (Keeling *et al.*, 2000). The first, and for a long time only, complete genome available for this group is that of *E. cuniculi* (Katinka *et al.*, 2001), an unicellular parasite that infects numeruos animals. Several additional microsporidian genomes have been published recently (Cornman *et al.*, 2009; Corradi *et al.*, 2009). The scientific community hopes that the additional information will finally resolve the exact phylogenetic location of this interesting group, whether it is within any of the existing fungal groups (Keeling *et al.*, 2000; James *et al.*, 2006) or as a sister clade to fungi (Tanabe *et al.*, 2005). Currently in our lab we are undertaking a phylogenomic analysis to address this issue.

The two groups that have traditionally been regarded as early diverging within the fungal kingdom are the Chitrids and the Zygomycetes. Chitrids have very characteristic morphological traits. For instance they are mainly aquatic and possess flagelated zoospores. From an economic standpoint, they are not very interesting, which results in a lack of financing towards genome projects. For a long time the only genome sequenced was that of *B. dendrobatidis*. This species has gained importance in recent years due to its role in the eradication of amphibian populations (Pounds *et al.*, 2006). Two other groups are part of the early diverging fungi. The Glomales, which were traditionally considered part of the Zygomycota, are now thought to have evolved independently. Currently, they are considered to be more closely related to dikarya and are given the category of a new phylum: Glomeromycota, although some doubts remain about this assignment. Some of the most well known Zygomycota are molds that can be found in spoiling bread or fruits. The genomes of three species are currently available. *R. oryzae*, is the principal causative agent of mucoromycosis, which is lethal in immunocompromised hosts (Roden *et al.*, 2005; Ma *et al.*, 2009). *Mucor circinelloides*, closely related to *R. oryzae*, has recently been sequenced due to the interest generated to its use in the biotechnological industry. This species is able to generate large quantities of lipids that can be easily converted into biodiesel and may one day replace the use of plant oils for the production of this compound. In contrast, the interest in *P. blakesleeanus* is centered on its ability to respond to environmental stimulus like light, chemicals, touch, gravity and others. In this sense this species has served as a model organism for the investigation of genes that regulate responses to stimulus (Idnurm *et al.*, 2006).

Basidiomycota

Basidiomycota comprise nearly 40% of the described fungi. Within this group we find those species that are able to produce mushrooms (fruiting bodies) as well as numerous plant symbionts. Mushroom producing fungi are probably among the most well known representatives of this kingdom. While the genomic sequence of the common mushroom, *Agaricus bisporus*, has only recently been made available at the Joint Genome Institute (http://www.jgi.doe.gov/), the sequences of other mushrooms have been available for several years. For instance, *Coprinus cinereus* is a saprophytic mushroom with a limited edible value, but which has been used as a model for sexual reproduction in Basidiomycota, with the aim of extrapolating the results found in this mushroom to other, more economically important, species (Kües, 2000; Kamada, 2002). On the other hand, *Schizophyllum commune*, belongs to the same group as the common mushroom and has been the focus of much research due to its unusual sexuality pattern, as it contains thousands of different mating types (Fowler *et al.*, 2001). This species is one of the few known filamentous basidiomycetes that cause infections in immunocompromised humans. It can also play an important role in bioremediation by uptaking minerals, for instance uranium and cadmium (Merten *et al.*, 2004).

The economic importance of Basidiomycota resides not only in the commercialization and consumption of mushrooms, but also in their ability to produce economically important products. Numerous pathogenic species can also be found within this group. *Cryptococcus neoformans* is an opportunistic pathogen that can cause central nervous system and pulmonary diseases in inmunocompromised hosts and the yeast *Malassezia globosa* has been associated with dandruff and can cause more serious skin problems. In plants, the Ustilagomycotina, are able to infect some of the most important crops including corn, barley, wheat and sugarcane (Bakkeren *et al.*, 2008). *Ustilago maydis*, the corn smut pathogen, has been used as a model organism for many years in molecular plant pathogeny research. Moreover, it has been suggested that *U. maydis* could be a good model to study animal basic cell processes when baker yeast is unable to cover them. For example, part of the DNA repair machinery in animals has no homologs in yeast, but it has been found in *U. maydis* (Steinberg and Perez-Martin, 2008; Münsterkötter and Steinberg, 2007).

Ascomycota

Three monophyletic groups have been described within Ascomycota: Taphrinomycotina, Pezizomycotina and Saccharomycotina. Pezizomycotina and Saccharomycotina are very well supported sister groups and together they represent more than 70% of the total number of sequenced species. The phylogenetic position and monophyly of Taphrinomycotina, is still under discussion (Kuramae *et al.*, 2006a). The clade was first defined based on the phylogenetic reconstruction using rRNA (Nishida and Sugiyama, 1993), and ended up grouping a very diverse set of species, including Schizosaccharomyces and Pneumocystis. The statistical support of the group was not very high and phylogenies derived from more than one gene have been contradictory. Most phylogenomics analyzes using nuclear proteins place Schizosaccharomyces as basal to Pezizomycotina and Saccharomycotina. On the other hand, phylogenies using mitochondrial sequences, tend to group Schizosaccharomyces within the Saccharomycotina. In a recently published paper, Liu and collaborators (Liu *et al.*, 2009) provide support for the monophyly of Taphrinomycotina, and place it as basal to Saccharomycotina and Pezizomycotina. Their analysis of mitochondrial sequences also showed that the addition of new Taphrinomycotina species to their trees lowered the support of the grouping of Schizosaccharomyces within the Saccharomycotina.

Currently four species belonging to this group have been sequenced. The yeast *Schizosaccharomyces pombe* has been used for many years as a model species for the study of cell cycle and cell biology. *Schizosaccharomyces japonicus* and *Schizosaccharomyces octosporus* are two very close relatives of *S. pombe*, which were sequenced with the aim of furthering the knowledge on this model species. On the other hand, *Pneumocystis carinii* is a pathogen that is found in the lungs of mammals and can cause serious pneumonia to inmunocompromised people.

Pezizomycotina is one of the largest groups of fungi both in terms of described species, more than 32,000, and in terms of the number of completely sequenced genomes. Nowadays, 76 genomes plus several different strains for some species have been sequenced. Pezizomycotina are mainly haploid and filamentous, though some species are known to be dimorphic. They live in a large diversity of habitats and ecological niches (Spatafora *et al.*, 2006). The phylogenetic relationship of the classes found within Pezizomycotina remains unresolved. Only four of these classes have representatives with fully sequenced species: Leotiomycetes, Sordariomycetes, Eurotiomycetes and Dothideomycetes. One of the few well established phylogenetic relationships is the association of Leotiomycetes and Sordariomycetes as sister groups.

The Leotiomycetes is a diverse group from a morphological standpoint, with great differences in fruiting bodies, which range from the large, brightly colored one found in Cyttaria to the small, dark dots produced by Lophodermium (Wang *et al.*, 2006b). One of the main difficulties in classifying the species within this group is the lack of a clear connection between anamorphs and teleomorphs. Many fungi are known by their teleomorphic stage but the anamorph is not known while in some cases the opposite is true (Wang *et al.*, 2006a). *Botrytis cinerea* and *Sclerotinia scleriotiorum* are both fully sequenced plant pathogens. *S. scleriotiorum* has a large range of hosts, up to 400, including many economically important crops, like soybean and pea, or oil seeds, like canola and sunflower (Hegedus and Rimmer, 2005).

In contrast to Leotiomycetes, its sister group, Sordariomycetes, is one of the largest classes within the Ascomycota with 3,000 known species and 18 completely sequenced species. Numerous well known pathogenic species can be found within this group such as the plant pathogens *Magnaporthe grisea* and Fusarium or the human pathogens *Chaetomium globosum* and *Nectria haematococca*. In addition, *C. globosum* can produce mycotoxins that are lethal to animals.

Sequencing efforts during the last few years seem to have been centered around the Dothideomycetes. All the genomes sequenced so far belong to plant pathogenic species. For instance *Stagonospora nodorum* infects wheat and other related cereals (Hane *et al.*, 2007). Many phylogenetic trees only include this representative of the Dothideomycetes. The uncertainty present in those trees has shown that more data is needed in order to fully understand the position of this group within the Pezizomycotina. More sequences are nowadays available and yet the uncertainty remains, indicating a need for sequencing species of the remaining classes of Pezizomycotina instead of increasing the sampling of already represented groups.

The last group of Pezizomycotina with sequenced species is the Eurotiomycetes. Sequencing efforts for this group have been clearly centered around the Aspergillus species. A total of eight different Aspergillus have been sequenced, ranging from the human pathogens *Aspergillus fumigatus* and *Aspergillus flavus* to the rarely pathogenic, but economically interesting, *Aspergillus niger* and *Aspergillus oryzae (Denning* et al.*, 2002)* which are used in industry to produce cytric acid and other proteins or to produce fermented foods like sake or miso. Also enclosed in this group is the saprobe *Aspergillus nidulans*, who has long been one of the model organisms and was thought to be the only Aspergillus species with a sexual cycle. Recent research has shown that other Aspergillus species, like *A. fumigatus*, also can reproduce sexually (O'Gorman *et al.*, 2009).

The last large group of Ascomycota species is the Saccharomycotina, a monophyletic group that mostly comprises yeast species. They are usually found in damp or wet habitats that are rich in organic materials. They are very important in many biotechnological areas, for instance, the baker's yeast, *S. cerevisiae*, is used in bread, beer and wine making. Numerous pathogens can also be found within this group, for instance the human pathogen *Candida albicans* or the plant pathogen, *Ashbya gossypii*. Many species have been described as being involved with arthropods in which the yeasts provide vitamins and enzymes in exchange of an efficient habitat. *S. cerevisiae* has been often the focus of numerous studies because the basic mechanisms of DNA replication, chromosomal recombination, cell division, gene expression, and metabolism are generally conserved between yeast and higher eukaryotes (Castrillo and Oliver, 2004).

More than 20 complete genomes are available for this group. The most divergent species included is *Yarrowia lipolytica*, which is often used as an outgroup for studies based on Saccharomyces. *Y. lipolytica* can be found in lipidrich environments, it is mainly non-pathogenic and it has been used in industry for the production of citric acid (Casaregola *et al.*, 2000).

In this overview we will divide the remaining species of this group in three different sub-groups that are often distinguished in literature: the Candida group, the Kluyveromyces group and the Saccharomyces group. The first one, also known as CTG group, comprises all the species that are closely related to *C. albicans*. The second one includes the three Kluyveromyces species in addition to *A. gossypii* and *Saccharomyces kluyveri*. This group is often also referred to as the Pre-whole genome duplication group (Pre-WGD) as it is the group of species that diverged from Saccharomyces just before the whole genome duplication (WGD) (Wolfe and Shields, 1997). The last group is formed by the species that underwent a WGD, they are closely related to *S. cerevisiae* and therefore have been the focus of numerous studies.

Candida or CTG group: One of the main distinctive features of the Candida group, is the change in the codification of the CTG codon from leucine to serine. This change occurred at the base of the group formed by the candidas and is characteristic for all its members. Candida species are opportunistic pathogens that can cause important infections in inmunocompromised hosts. *C. albicans* is the main causative agent for candidiasis. The effect of antifungals that specifically target this species has caused an emergence of infections caused by non-albicans candidas in recent years. One example of this are the infections produced by *C. glabrata*, which is much more closely related to *S. cerevisiae* than to *C. albicans* (see chapter 7). Also, some Candidas are rarely observed as pathogens, but under certain conditions they can be found as the prevalent strains. For instance, *Candida guilliermondii*, while not usually found causing infections, is considered the main cause of fungemia in cancer patients. While mostly known because of their pathogenesis, some of the species within the Candida group are non-pathogenic. *Debaryomyces hansenii* for instance is a marine yeast that can tolerate high salinity levels. Other non-pathogenic species within the Candida group are *Lodderomyces elongisporus* and *Pichia stipitis*, which is known for its ability to ferment xylose.

Kluyveromyces group: Probably the two most well studied species within this group are *Ashbya gossypii* and *Kluyveromyces lactis*. The first one is a plant pathogen while the second one is a saprobe which is often found in lactic products as it is able to grow on lactose. Little is known about the ecology of the three other species that conform this group (*Kluyveromyces waltii*, *Saccharomyces kluyveri* and *Kluyveromyces thermotolerans*) and they were sequenced mainly due to their proximity to *S. cerevisiae*. This group of species has often been used as a reference to compare to Saccharomyces genomes as they are thought to retain most of the genomic structure the species had before undergoing the whole genome duplication.

Saccharomyces clade: In nature Saccharomyces species have been found in the bark of some trees and in fermenting fruits or other high sugar environments. They have a high tolerance to alcohol, many of them grow anaerobically and are able to ferment glucose to alcohol. Due to the importance of *S. cerevisiae* as a model species, the genomes of several closely related species have been sequenced. Probably one of the most interesting species within this group is *C. glabrata* which, unlike the others, is an opportunistic human pathogen that can cause candidiasis in inmunocompromised hosts. The fact that this species is so closely related to numerous non-pathogenic species, coupled by the differences observed in the infection mechanism when compared to *C. albicans*, points to the fact that *C. glabrata* developed its pathogenesis independently. One of the main phylogenetic questions that remains unresolved within this group is the relative position of *C. glabrata* and *S. castellii* (see chapter 4). Phylogenetic analysis usually shows a poorly supported topology in which *C. glabrata* was the first species to diverge from the main Saccharomyces group, but results based on synteny support that the first species to diverge was *S. castellii* (Gordon *et al.*, 2009).

Final remarks

The fungal kingdom presents a high diversity and we have barely started to discover the many entities that are part of it. Even though current estimates show that we are only aware of about 5% of the fungi, the many differences that have already been shown in terms of morphology, ecology and adaptation to different environments is amazing. The fungal kingdom is, without doubt, one of the most interesting groups of species that still remain mostly undiscovered. Fungal species will play a very important role in the genomic era. The conformation of their genome, with enough simplicity to facilitate its investiagtion but also with enough complexity to find numerous parallels to more complex organisms, conforms one of the main attractions for their use in experimental studies. The current availability of more than 100 completely sequenced genomes has also opened the doors for comparative genomics studies that will help understand how these organisms have been able to adapt to so many different environments. One of the main questions that still remains is which is the exact topology of the fungal species tree. Hopefully, our growing knowledge will help clarify those points in the fungal species tree that still remain unresolved.

CHAPTER 2

A BRIEF INTRODUCTION TO PHYLOGENOMICS

From phylogenetics to phylogenomics

Phylogenetics aims at establishing the evolutionary relationships between organisms. Since the publication of “The origin of species” (Darwin, 1859), species evolution has been represented in the form of a hierarchical tree where extant species are located at the tips of the branches whereas bifurcating nodes represent their common ancestors. Initially, phylogenies were based on morphological data, an approach that is highly susceptible to homoplasy effects. The increasing availability of biological sequences has facilitated the replacement of morphological characters as the basis for phylogenetic reconstruction. The first trees based on sequences were built using single molecular markers such as 16sRNA (Zuckerkandl and Pauling, 1965) but it was soon obvious that different markers may produce different phylogenies and, therefore, they were not suitable for establishing the evolutionary relationship between species. The possibility of combining several different markers to reconstruct a single species tree is a straightforward solution, which is now favored with the growing availability of molecular data.

Phylogenomics represents the intersection between the fields of evolution and genomics (Eisen and Fraser, 2003). Increasingly larger groups of genes are treated together in an attempt to gain a broader view of the events that have shaped the evolution of species. This incresing amount of data will enforce another way of working, with automation being a key factor in any phylogenomics study. Automatic pipelines have been designed to generate high quality phylogenetic trees with minimal manual imput. In addition, programs also need to be designed to automatically process the information present in phylogenetic trees, since the large amount of data makes manual processing unfeasible.

Some examples of phylogenomics studies comprise the construction of species trees using multiple genes (chapter 3), the establishment of orthology and paralogy relationships between genes (chapter 3) or the use of phylogenetic trees to establish evolutionary events such as duplications (chapter 6) or horizontal genes transfers (chapter 5).

Phylogenetic reconstruction

Phylogenetic pipeline

Often, the basis of phylogenomics studies consists of phylogenetic analysis at a large scale, both in terms of number of trees and in terms of the number of species included. In order to cope with the increasing amount of data, pipelines have been designed that are able to reconstruct large amounts of phylogenetic trees (Huerta-Cepas *et al.*, 2007; Hubbard *et al.*, 2007; Gabaldón *et al.*, 2008). A balance between accuracy and speed has to be achieved in order to produce accurate trees at a reasonable speed. Most phylogenetic pipelines can be divided into three steps: selection of putative homologous sequences, reconstruction of a reliable multiple sequence alignment and reconstruction of a phylogenetic tree that represents the evolutionary relationships of the sequences involved.

The pipeline we have used throughout this thesis was first applied in the reconstruction of the human phylome (Huerta-Cepas *et al.*, 2007). First a Smith-Waterman search (Smith and Waterman, 1981) is performed using an initial protein which will be referred to as seed protein. This search is performed against a locally installed database that contains either a downloaded database (for instance from NCBI (http://www.ncbi.nlm.nih.gov/) or Uniprot (http://www.uniprot.org/)) or a manually created database. The results are then filtered, and only those hits that have an e-value below a given threshold and have a long enough continuous aligned region are taken as homologs. In the second step Muscle v3.6 (Edgar, 2004) is used to make the multiple sequence alignment. This alignment is trimmed using trimAl (Capella-Gutiérrez *et al.*, 2009) in order to remove poorly aligned regions. PhyML (Guindon and Gascuel, 2003) is then used to derive the phylogenetic trees. In this pipeline a NJ tree is first reconstructed using BIONJ (Gascuel, 1997) as implemented in PhyML, then this tree is used as starting point to reconstruct the maximum likelihood trees. Up to four different evolutionary models are used (typically JTT, WAG, VT and Blosum62). The evolutionary model best fitting the data is then determined by comparing all the likelihoods according to the AIC criterion (Akaike, 1973).

Phylome reconstruction and phylomeDB.

While the described pipeline could be applied to one single protein family it was specifically designed to reconstruct phylomes. Phylomes are defined as the complete collection of phylogenetic trees for each gene in a genome (Sicheritz-Pontén and Andersson, 2001). They can be used for numerous studies including the prediction of orthology and paralogy relationships (Huerta-Cepas *et al.*, 2007; Marcet-Houben and Gabaldón, 2009), the tree based inference of function for newly sequenced genomes (Consortium, 2010) and to search for poorly supported nodes in a species tree (Marcet-Houben and Gabaldón, 2009).

The reconstruction of phylomes produces a large amount of data. In order to make this information available to the whole community, phylomeDB (www.phylomedb.org) (Huerta-Cepas *et al.*, 2008) was designed in our lab. PhylomeDB provides users with the opportunity to actively visualize phylogenetic trees or to download all the data for more intensive use. PhylomeDB has several browsing methods which include the search of proteins using Ids and the possibility to perform a blast search against the proteomes stored in the database. Several external Ids are covered in the database, easing the access to the data.

While the number of public phylomes is still relatively small (12), the number of proteomes included in the database is much larger (945). Even though proteomes not used as seed are not fully covered, many of their proteins can still be present in other phylogenetic trees. The work performed during this thesis has contributed to the reconstruction of 7 phylomes (four of *Saccharomyces cerevisiae*, one of *Candida glabrata*, one of *Candida albicans* and one of *Schistosoma mansoni*).

Combining information from various genes to derive a single phylogeny

Phylogenetic trees based on a single gene family are useful in many areas. However, they present the drawback of displaying different topologies depending on the family used to reconstruct the tree (Castresana, 2007). This situation is especially problematic when the aim of the analysis is to define a single phylogeny that represents the evolutionary relationships among a group of species, the so called species tree. A possible solution to this problem is to integrate the phylogenetic information from various genes into a single tree (Altekar *et al.*, 2004). We can distinguish between two kind of approaches. The first kind comprises those methods that do not use sequence data to infer the tree but rather are based on genome characteristics such as gene content or gene order (Snel *et al.*, 2005). The second kind directly uses sequences, either to create a super alignment (Brown *et al.*, 2001) from which to derive the tree, or to combine different trees created from individual genes into a single super-tree (Bininda-Emonds, 2004).

Here we will use the gene concatenation method as it is one of the most cited in the literature and is considered highly reliable (Delsuc *et al.*, 2005). It must be noted that comparisons between super-tree approaches and concatenations have not detected large differences in terms of topologies (Dutilh *et al.*, 2007; Fitzpatrick *et al.*, 2006). The rationale behind gene concatenation is that genes in a genome have undergone the same evolutionary history. Therefore, if taken together, they should be able to amplify the phylogenetic signal they share. In order to use gene concatenation, each gene should, ideally, be present in single copy in all species considered. As a result, the number of genes that can be used in such analyses decreases as the number of species included grows. To attenuate this effect, genes absent in a few genomes can be included by introducing gaps in the missing species and methods to select one gene from few recent paralogs can be applied.

Major drawbacks of this method include a low number of appropriate genes when large groups of species are considered. Another concern is related to data heterogeneity. In this approach all genes are treated together under the same models, assuming homogeneity in the data, which is known to be untrue (Bull *et al.*, 1993). Additionally, events such as lineage sorting or horizontal gene transfers can led to a different evolutionary history for a given gene. This method is very sensitive to such cases and care should be taken to exclude genes that are susceptible to have undergone such events (Wolf *et al.*, 2001).

Interpretation of large-scale datasets

Together with the high computational costs associated to large scale phylogenetics reconstruction, the biological interpretation of extensive collections of trees and alignments represents another important challenge in the field of phylogenomics. Inspection of thousands of phylogenies cannot be addressed manually and, therefore, automatic methods and algorithms for the interpretation of phylogenies are necessary. Here we give some examples of studies that can be undertaken and which have been used throughout the present thesis.

Phylogeny-based orthology prediction

Orthologs are genes that diverged after a speciation event (Fitch, 1970). When compared to paralogs, which evolved through a duplication event, orthologous pairs of genes show a higher tendency to perform the same function. This is the reason why predicting them correctly has become so important (Gabaldón, 2005).

There are many methods that predict whether two genes are orthologous to each other, most of them based on pair-wise similarities (Koonin, 2005). However, since the original definition of orthology is an evolutionary one (Fitch, 1970), a prediction based on phylogeny seems to be more appropriate (Gabaldón, 2008). There are two main approaches to derive orthology relationships from phylogenetic trees, namely reconciliation and species-overlap method. Reconciliation methods use a species tree as reference (Page and Charleston, 1997). When comparing a gene tree with the species tree, mismatching nodes are identified and considered to be duplication events which were subsequently followed by the necessary amount of gene loss to explain the resulting phylogeny. This approximation will render correct orthology predictions if the assumption that the gene and species trees are correct holds. An alternative approach is the so-called species-overlap method. In this case duplication nodes are only considered as such when their branches have shared species. While simple, this method performs well (Huerta-Cepas *et al.*, 2007) and has the advantage that the only information that is needed from the species phylogeny is the one used to root the tree. In a comparison between the two methods we showed that the species overlap algorithm had a higher sensitivity while the positive predictive value of both methods remained similar (Marcet-Houben and Gabaldón, 2009).

Tree comparison and topological pattern search

Natural questions that may arise when inspecting large datasets of phylogenetic trees include how similar a group of trees are from each other or which fraction of trees provide support for a specific topology. There is a large variety of programs and metrics that are able to compare two trees. Perhaps the quartet (Estabrook *et al.*, 1985) and Robinson and Foulds (Robinson and Foulds, 1981) distances are the most commonly used. The quartet distance counts the number of subtrees formed by four leaves that differ between two trees, whereas Robinson and Foulds distance is based directly on the edge structure of the trees and their induced bipartitions. While usually distance methods are centered around the topological comparison of trees, some programs also include information regarding branch lengths (Soria-Carrasco *et al.*, 2007). Their drawback is that they usually can not be directly applied on trees with different evolutionary rates. A serious problem involving the comparison of trees is that most algorithms are limited to trees with the same taxa. This situation is unrealistic as events such as gene loss or duplication often produce relationships between genes of different species that are not one-to-one. The straightforward solution to deal with the different amount of taxa in two trees is to prune the two trees until they contain the same amount of taxa. Then distances are corrected to take this deletion into account. The matter of duplications is slightly more complicated. In the Topd/FMTs (Puigbò *et al.*, 2007) method they address the problem by randomly deleting duplicated genes and comparing all the resulting trees with each other. While mathematically sound, the method does not account for orthology and paralogy relationships when prunning the tree (i.e. the prunned tree may contain a mixture of orthologous and paralogous genes) and by randomly choosing the sequences to delete it risks comparing multiple different trees even if the initial trees are identical. A clear example of this problem occurs when comparing a tree with several duplications with itself. The expected distance would be 0, but due to the randomness of the prunning step, the distance can often reach near random values (unpublished observations). In chapter 4 we present our algorithm of tree comparison (treeKo, http://treeko.cgenomics.org), which was designed with this specific problem in mind.

Related to the problem of comparing two trees is the need to develop algorithms that identify specific topological patterns within the trees. Patterns can go from the simplest case scenario in which we want to retrieve those trees in which one species is grouped specifically with another, without any other species between them, to much more complicated patterns that can even have duplication events involved. Some groups have implemented algorithms to search for specific topological patterns (Esser *et al.*, 2004; Gabaldón and Huynen, 2003) that are based on the examination of all the possible tree partitions. Dufayard and colleagues (Dufayard *et al.*, 2005; Gouret *et al.*, 2009) have implemented a similar algorithm that allows the user to define specific scenarios with the help of a graphical interface.

Transference of function to newly annotated proteins

The genomics era has brought about a large amount of sequence data that needs to be processed. Even with high-throughput experiments, thousands of coding sequences remain without an experimental validation. Even in model species, such as *Escherichia coli* and *Saccharomyces cerevisiae*, on which scientists have focused their efforts for decades, the function of many proteins still remains unknown or is poorly defined (Hu *et al.*, 2009; Peña-Castillo and Hughes, 2007). One way to overcome this lack of experimental information regarding the functionality of newly sequenced proteins is by using computational means to predict the putative function of a protein based on previous knowledge obtained in other species. Similarity between sequences, as predicted by best blast hits, was at first used to automatically transfer annotations from one protein of a species to another. While this method has been used extensively, it is full of potential pitfalls. Relationships between proteins of different species are not necessarily one to one, due to duplication events. The consequence is that if we transfer the function of a protein based only on similarity we may end up transferring the function of a paralogous protein. Another danger is the fact that a relationship of homology (or orthology) between two proteins does not necessarily mean they will have the same function at all levels. For instance there are large variations of substrate affinities within families of transporters or metabolic enzimes.

The use of phylogenies in order to transfer functional annotations is beneficial twofold. On one hand it clearly depicts the evolutionary relationship between proteins, easily differentiating between paralogs and orthologs. It also allows the identification of proteins that have several co-orthologs, which will immediately imply that the annotation can not be trusted without further analysis. On the other hand, it will allow the transference of functions not only from one species to another, but rather of all the species in the tree that have an annotation. This can either raise our confidence in the prediction, when all the annotations agree on the same function, or it can reveal when a protein changed its function over time. If we expand this to encompass a whole phylome, we will possess a useful tool which will help in the annotation of a newly sequenced genome. This methodology has already been successfully used in the annotation of the pea aphid genome (Consortium, 2010; Huerta-Cepas *et al.*, 2010b), which has been the first newly sequenced genome to be functionally annotated with a phylogeny based strategy, and the *Schistosoma mansoni* genome (unpublished data). Moreover we have used it in the functional analyses of *C. glabrata* genome, performed within the context of the FunPath consortium (see chapter 7).

Final remarks

Phylogenomics can be used for multiple purposes. The elucidation of the evolutionary relationship of a group of species, the transference of function from one protein to another, the detection of evolutionary important events, they are all based on the construction of one or, more often, numerous phylogenetic trees. The increase in the number of phylogenies used in a given analysis has led to the necessary automation of the whole process of tree reconstruction (Gabaldón *et al.*, 2008). A balance between speed and accuracy has been necessary in order to ensure the viability of the data used in further analysis. There exist different ways in which to deal with the data produced by phylogenies. Due to the size of the datasets involved, manual exploration is often put aside in favor of an automatic approach. Several applications of phylogenomics are comprised within this volume. Each chapter will involve a study in which phylogenomics tools are applied to a group of fungal species in order to solve different biologicaly relevant questions.

AIMS

- Reconstruct a fungal species tree that reflects the evolution of all completely sequenced fungal species.
- Asses, on a genome-wide basis, the robustness and congruence of each node in the fungal species tree.
- Develop novel tools that enable truly genome-wide scales in the comparison of gene trees and species trees.
- Assess the extend of some evolutionary mechanisms that may cause topological variation in fungal gene trees (i.e. gene loss, gene duplication, horizontal gene transfer).

Part II

Results

CHAPTER 3

THE TREE VERSUS THE FOREST: THE FUNGAL TREE OF LIFE AND THE TOPOLOGICAL DIVERSITY WITHIN THE YEAST PHYLOME

Marcet-Houben, M. and Gabaldón, T. (2009). The tree versus the forest: the fungal tree of life and the topological diversity within the yeast phylome. PLoS One, 4(2), e4357.

Introduction

The advent of the genome era and the availability of a growing number of fully-sequenced genomes have changed the way in which biologists study the evolutionary relationships among groups of organisms. For instance, the use of phylogenetics in the context of whole genomes, a field known as phylogenomics (Eisen, 1998), allows for the combination of evolutionary signals from various genes into a single tree. It has long been observed that phylogenetic trees built from different genes may provide conflicting topologies. Thus, the use of multiple gene approaches is a way to average out these discrepancies in order to provide a single topology that is expected to reflect the true evolutionary relationships more accurately. In recent years, the use of multi-gene approaches, and especially gene concatenation, is becoming the method of choice in most studies aiming to elucidate the evolutionary relationships among a group of species (Delsuc *et al.*, 2005). Such approaches are, however, not free from criticism. For instance, it has been argued that they use the information derived from a small fraction of the genes in a genome and, therefore, cannot represent the actual diversity of evolutionary histories within a genome (Dagan and Martin, 2006). Indeed, initial genome-wide phylogenetic studies have shown that the topological diversity encountered across a genome is high (Huerta-Cepas *et al.*, 2007; Rasmussen and Kellis, 2007). Besides questioning the validity of species trees, these findings have raised doubts regarding the possible sources for the high topological variability and the implications for large-scale phylogenetic inferences such as the prediction of orthology relationships.

Here we address the question of whether species trees constructed with standard alignment concatenation approaches do fairly represent the topologies that can be found in gene phylogenies across a genome. Conversely, we test whether the topological information found across all genes in a genome can be used to identify conflicting nodes and provide alternative reliability values in species trees. We test these ideas by using molecular data from fungal genomes, the group of eukaryotic organisms that is best sampled in terms of fully sequenced genomes (Galagan *et al.*, 2005). Currently, more than 60 fungal species have been sequenced, including many human pathogens as well as other species of industrial or agricultural interest. This has facilitated that the evolutionary relationships among fungi have been addressed by means of phylogenomic methods, being gene concatenation the most widely used (Kuramae *et al.*, 2006b; Fitzpatrick *et al.*, 2006; Robbertse *et al.*, 2006). To assess the extent of congruence between trees based on concatenated alignments and individual phylogenies, we compare the topology of phylogenies of genes encoded in the yeast genome with fungal species trees reconstructed from the concatenated alignments of widespread proteins present across different sets of fungal species. Our results show that, despite the large topological diversity of the yeast phylome, most nodes in the species tree do represent genome-wide supported evolutionary relationships. Some conflicting nodes, however, concentrate most of the topological variations found between gene and species trees. We propose to incorporate such information in the tree of life in the form of genome-wide levels of topological support, thereby identifying conflicting nodes. Finally, some of the possible causes for the existing topological diversity within a genome and its implications for orthology prediction are discussed.

Materials and Methods

Sequence data

Proteins encoded in 60 fully-sequenced fungal genomes were downloaded from several databases (see table 3.1). Additionally, genomes from *Homo sapiens* and *Arabidopsis thaliana* were downloaded from ensembl. The final database comprises 626,834 unique protein sequences.

Yeast phylome reconstruction

We used the pipeline described in (Huerta-Cepas *et al.*, 2008). In contrast to family-based methods in which first sequences are clustered into groups based on pair-wise comparisons, for instance using MCL clustering, the phylome approach uses one genome as a seed to find putative homologs, just as a phylogeneticist would do to reconstruct the evolution of a protein of interest. This approach maximizes the coverage over the seed genome and being independent of the parameters of the clustering algorithm (Huerta-Cepas *et al.*, 2007). For each *Saccharomyces cerevisiae* “seed” protein a Smith-Waterman (Smith

	Code	Organism name	Source	Proteins	Trees	T60	T21	T12a	T12b
Saccharomyces sensu stricto	Sce	Saccharomyces cerevisiae	SGD	5811 (86%)	5804 (100%)	■	■	■	■
	Spa	Saccharomyces paradoxus	Fungal Genomes	5356 (95%)	5396 (93%)	■	■	■	
	Smi	Saccharomyces mikatae	Fungal Genomes	5136 (90%)	5263 (91%)	■	■	■	
	Sku	Saccharomyces kudriavzevii	Fungal Genomes	5027 (83%)	5114 (88%)	■	■	■	
	Sba	Saccharomyces bayanus	SGD	5382 (91%)	5464 (94%)	■	■	■	
Saccharomyces complex	Sca	Saccharomyces castellii	YGOB	5253 (92%)	5165 (89%)	■	■	■	
	Cgl	Candida glabrata	Genolevures	4866 (93%)	5032 (87%)	■	■	■	
	Kpo	Kluyveromyces polysporus	YGOB	5004 (91%)	5129 (88%)	■	■	■	
	Ago	Ashbya gossypii	NCBI	4314 (91%)	4815 (83%)	■	■	■	■
	Kla	Kluyveromyces lactis	NCBI	4542 (85%)	5030 (87%)	■	■	■	
	Kwa	Kluyveromyces waltii	YGOB	4482 (85%)	4986 (86%)	■	■	■	
	Skl	Saccharomyces kluyveri	Fungal Genomes	4267 (74%)	4736 (82%)	■	■	■	
Candida cluster	Cal	Candida albicans	Candida Genome Database	4063 (66%)	4291 (74%)	■	■	■	■
	Cdu	Candida dubliniensis	Fungal Genomes	3952 (59%)	4165 (72%)	■	■		
	Ctr	Candida tropicalis	Broad Institute	4089 (65%)	4185 (72%)	■	■		
	Lel	Lodderomyces elongisporus	Broad Institute	3788 (65%)	4082 (70%)	■	■		
	Pst	Pichia stipitis	JGI	4162 (71%)	4277 (74%)	■	■		
	Dha	Debaryomyces hansenii	Integr8	4132 (65%)	4323 (74%)	■	■		
	Cgu	Candida guilliermondii	Broad Institute	4045 (68%)	4202 (72%)	■	■		
	Clu	Candida lusitaniae	Broad Institute	3831 (64%)	4180 (72%)	■	■		
Dipodascaceae	Yli	Yarrowia lipolytica	Integr8	3882 (59%)	4009 (69%)	■	■		■
Leotiomycetes	Bci	Botrytis cinerea	Broad Institute	3503 (21%)	3625 (62%)	■			■
	Ssc	Sclerotinia sclerotiorum	Broad Institute	3554 (24%)	3762 (65%)	■			
Sodariomycetes	Mgr	Magnaporthe grisea	Broad Institute	3566 (27%)	3680 (63%)	■			
	Ncr	Neurospora crassa	Broad Institute	3392 (31%)	3708 (64%)	■			■
	Cgo	Chaetomium globosum	Broad Institute	3207 (28%)	3458 (60%)	■			
	Pan	Podospora anserina	Fungal Genomes	3525 (27%)	3637 (63%)	■			
	Tre	Trichoderma reesei	JGI	3663 (40%)	3799 (65%)	■			
	Gze	Fusarium graminearum	Integr8	3905 (33%)	3755 (65%)	■			
	Fox	Fusarium oxysporum	Broad Institute	4375 (24%)	3780 (65%)	■			
	Fve	Fusarium verticillioides	Broad Institute	3965 (27%)	3736 (64%)	■			
	Nha	Nectria haematococca	JGI	4671 (29%)	3802 (66%)	■			
Dothideomycetes	Sno	Stagonospora nodorum	Broad Institute	3883 (23%)	3783 (65%)	■			■
	Mfi	Mycosphaerella fijiensis	JGI	3472 (33%)	3606 (62%)	■			
Eurotiomycetes	Hca	Histoplasma capsulatum	Broad Institute	2993 (32%)	3415 (59%)	■			
	Cim	Coccidioides immitis	Broad Institute	3361 (32%)	3680 (63%)	■			
	Ure	Uncinocarpus reesii	Broad Institute	3159 (40%)	3491 (60%)	■			
	Acl	Aspergillus clavatus	TIGR	3873 (42%)	3855 (66%)	■			
	Afu	Aspergillus fumigatus	Broad Institute	3907 (40%)	3837 (66%)	■			
	Nfi	Neosartorya fischeri	TIGR	4063 (39%)	3845 (66%)	■			
	Ani	Aspergillus nidulans	NCBI	3629 (38%)	3729 (64%)	■			■
	Ang	Aspergillus niger	Broad Institute	4157 (29%)	3809 (66%)	■			
	Aor	Aspergillus oryzae	Integr8	4090 (33%)	3653 (63%)	■			
	Afl	Aspergillus flavus	Broad Institute	4144 (32%)	3770 (65%)	■			
	Ate	Aspergillus terreus	Broad Institute	3865 (37%)	3710 (64%)	■			
Taphrinomycotina	Spb	Schizosaccharomyces pombe	Integr8	3212 (64%)	3442 (59%)	■			■
	Sja	Schizosaccharomyces japonicus	Broad Institute	3076 (59%)	3287 (57%)	■			
	Pca	Pneumocystis carinii	Fungal Genomes	814 (20%)	965 (17%)	■			
Basidiomycota	Cci	Coprinus cinereus	Broad Institute	3210 (23%)	3310 (57%)	■			
	Cne	Cryptococcus neoformans	Integr8	3130 (47%)	3366 (58%)	■			■
	Lbi	Laccaria bicolor	JGI	3779 (18%)	3350 (58%)	■			
	Ppl	Postia placenta	JGI	4952 (28%)	3291 (57%)	■			
	Pch	Phanerochaete chrysosporium	JGI	3241 (32%)	3216 (55%)	■			
	Pgr	Puccinia graminis	Broad Institute	2966 (14%)	2956 (51%)	■			
	Sro	Sporobolomyces roseus	JGI	2713 (49%)	3024 (52%)	■			
	Uma	Ustilago maydis	Broad Institute	2858 (43%)	3287 (57%)	■			
Zygomycota	Ror	Rhizopus oryzae	Broad Institute	4782 (27%)	3193 (55%)	■			■
	Pbl	Phycomyces blakesleeanus	JGI	4482 (30%)	3386 (58%)	■			
Chitridiomycota	Bde	Batrachochytrium dendrobatidis	Broad Institute	2774 (31%)	2842 (49%)	■			
Microsporidia	Ecu	Encephalitozoon cuniculi	Integr8	698 (36%)	920 (16%)	■			■

Table 3.1: Species included in the 60-species phylomes and their genomic coverage. For each species, "proteins included" column indicates the number of proteins present in trees of the yeast phylome and the percentage they represent; "trees" column indicates the number of trees in the phylome with proteins from that species (and the percentage from the phylome it represents). "Source" indicates the database from which the protein data for that species was retrieved. The last four columns indicate if this species was considered in each of the indicated species trees (shadowed boxes). Sources included are: JGI (http://www.jgi.doe.gov), Broad Institute (http://www.broad.mit.edu), YGOB (http://wolfe.gen.tcd.ie/ygob/), SGD (www.yeastgenome.org), Fungal genomes (http://fungalgenomes.org), Genolevures (http://cbi.labri.fr/genolevures/), integr8 (http://www.ebi.ac.uk/integr8), Candida genome database (http://www.candidagenome.org), NCBI (http://www.ncbi.nlm.nih.gov).

and Waterman, 1981) search was used to retrieve, from the abovementioned database, a set of proteins with a significant similarity (E-val<10–3). Only sequences that aligned with a continuous region representing more than 33% of the query sequence were selected. These sequences are considered putative homologs and are aligned with MUSCLE 3.6 (Edgar, 2004). Positions in the alignment with gaps in more than 10% of the sequences were trimmed as described in (Huerta-Cepas *et al.*, 2007). Neighbour Joining trees were derived using scoredist distances as implemented in BioNJ (Gascuel, 1997). PhyML aLRT version (Guindon and Gascuel, 2003; Anisimova and Gascuel, 2006) was used in to derive Maximum Likelihood (ML) trees. Four different evolutionary models were used for each seed sequence (JTT, WAG, Blosum62 and VT). In all cases, a discrete gamma-distribution model with four rate categories plus invariant positions was used, estimating the gamma parameter and the fraction of invariant positions from the data. The evolutionary model best fitting the data was determined by comparing the likelihood of the used models according to the AIC criterion (Akaike, 1973). The resulting 22,352 alignments and 111,760 phylogenetic trees for the four different generated phylomes can be publicly accessed in phylomeDB (Huerta-Cepas *et al.*, 2008) (http://www.phylomedb.org).

Reconstruction of the fungal trees of life

To reconstruct the T60 fungal species tree we proceeded as follows. Based on the orthology relationships derived from the yeast phylome (see below), we selected 69 proteins that were present in at least 58 of the 60 fungal organisms and show one-to-one orthology relationships in these species. The alignments of these proteins were concatenated into a single alignment, which was then trimmed to remove positions with gaps in more than 50% of the organisms. The resulting alignment comprises 31,123 amino acid positions. The tree was constructed using a Maximum Likelihood approach as implemented in the PhyML program (Guindon and Gascuel, 2003), using a discrete gamma-distribution model with four rate categories plus invariant positions. The gamma parameter and the fraction of invariant positions were estimated from the data. The evolutionary model used for the analysis was WAG, as it was the model best fitting 61 of the 69 individual alignments.

The same procedure was also applied for the T21, T12a and T12b trees. Also using WAG as a model. In all cases the alignments were trimmed to eliminate columns with gaps in more than 50% of the positions. T21 is derived from a concatenation of 1,137 protein families present in all 21 species. The final alignment included 28,3974 amino acid sites. T60 and T21 showed similar topologies with only the relative clustering of *Debaryomyces hansenii* and *Candida guillermondii* differing between the two trees. T12a included

2,007 proteins present in all species and 580,514 positions. And, finally T12b comprised 217 widespread proteins and 95,528 positions. Support values were computed by bootstrap analysis of 100 replicates, unless indicated otherwise. The topologies in these two trees are fully compatible to that of T60.

Simulations of sequence evolution

We used Rose (Stoye *et al.*, 1998) to generate simulated sequences from 50 yeast proteins that were chosen randomly among the ones used in the construction of T21. The simulations included insertions and deletions with a probability of 0.03. The other parameters for the simulation were the ones described in (Talavera and Castresana, 2007). We also used the same strategy to infer the patterns of rate heterogeneity of the seed proteins. In short, we used TreePuzzle (Schmidt *et al.*, 2002) assuming a 16 rate gamma distribution and for each position in the alignment we took the category and associated relative rate that contributed the most to the likelihood. These rates of heterogeneity were used by rose to model the evolution of the seed sequences along the T60 tree. The resulting simulated sequences were used to create a maximum likelihood tree using the WAG evolutionary model. Additionally, a species tree from the concatenated alignments was also reconstructed.

Inference of duplication and speciation events and benchmark of orthology assignments

We used two alternative phylogeny-based methods to derive orthology relationships on the 60-species phylome. First, we used a previously described species-overlap algorithm (Huerta-Cepas *et al.*, 2007) to map duplication and speciation events on the trees. In short, the algorithm starts at the seed protein used to generate the tree and runs through the internal nodes of the tree until it reaches the root. Trees were rooted at the midpoint. At each node, two daughter tree partitions are defined. If the two partitions share any species, the node is defined as a duplication node. Otherwise the node is defined as a speciation node. Once all the nodes have been classified, the algorithm establishes the orthologous and paralogous relationships between the seed protein and the rest of the proteins included in the tree.

Next, a strict tree-reconciliation algorithm was used (Zmasek and Eddy, 2001). In this case, every tree of the phylome is compared to the topology in the species tree by comparing the specific sets of species contained by all tree splits. The strict reconciliation algorithm maps the gene tree to the species tree and any incongruence is explained in terms of the minimal set of duplication and gene-loss events necessary to derive the observed gene tree topology from the one proposed in the species tree. These inferred duplication events are marked

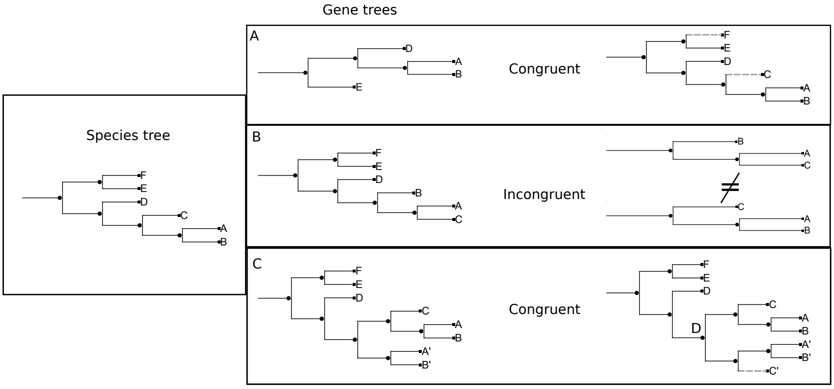

Figure 3.1: Explanation on how the topology scanning algorithm works in different situations. Figure 3.1a shows a tree that is congruent with the species tree despite the loss of two species. Figure 3.1b shows a tree that is not congruent with the species tree due to the rearrangement of two of its branches. Since no species overlap exists we do not consider this node as a duplication node as there is no evidence for such an assumption. Figure 3.1c represents a congruent tree with a duplication node as predicted by the species overlap algorithm.

on the tree and orthology and paralogy relations are derived accordingly.

The orthology predictions derived from the phylome with the two strategies explained above, were compared to those made in the YGOB database (Byrne and Wolfe, 2006). We used this reference set to compute the number of true positives (TP), false positives (FP) and false negatives (FN) yielded by our method. For each method the sensitivity, S = TP/(TP+FN), and the positive predictive value, P = TP/(TP+FP) were computed.

Topology scanning algorithm

The strategy used here to search for specific topologies within the phylome is based on an algorithm described earlier (Gabaldón and Huynen, 2003). Perl scripts were written to implement this tree scanning algorithm to the specific scenarios considered here. In brief (see figure 3.1 for more details), the algorithm proceeds sequentially throughout all internal edges of the tree, starting from each of the external nodes of the tree and proceed towards the root. Trees were rooted at the most distantly related species present in the tree, according to the topology in T60.

At each internal node, two daughter partitions are generated and the species present in each such partition are tracked. The specific order in which the species appear in the tree can then be compared to specific scenarios. This algorithm was used to compare all the trees in a given phylome with the topology of the corresponding species tree. The algorithm considers only topological relationships among orthologous sequences. The phylome tree was considered to have a topology not compatible to that of the species tree if it contained a single species arrangement not found in the species tree and which could not be explained by gene loss events. Duplications found in the gene tree were always considered compatible and we only focused on the specific species arrangement within each partition resulting from the duplication. Proceeding in such way, we assure that only topological arrangements between orthologous genes are considered (duplications define paralogous relationships). Note that, since duplications may originate one-to-many or many-to-many orthology relationships it might be the case that the relationship with one co-ortholog is supporting the species tree topology while that with the other co-ortholog is rejecting it. Since we evaluate "full compatibility", these trees were not considered compatible.

Phylome support values

We define the "phylome support value" for a node as the percentage of trees in a phylome that present exactly the same topological arrangement of the partitions defined by its two daughter nodes. As indicated in figure 3.2B, the two daughter nodes can define three (A,B,C) or four partitions (A,B,C,D) that might display three or fifteen alternative topologies, respectively. To compute the phylome support value we used the topology-scanning algorithm described above.

In this case, for any specific arrangement of three or four groups (see figure 3.2B) of species defined by a given node of the species trees we search for compatible partitions in all the trees in the phylome. Trees that did not have at least one species from each of the groups involved in the topology were not considered, because they do not provide information on that topology. That is, if the support for the topology ((A,B)C) is evaluated, we can only consider trees that contain at least one sequence from each of the three groups. Note that, in contrast to bootstrap supports that are only informative on the support for a single partition, the "phylome support value" takes into consideration the specific arrangement between several partitions and it is thus more informative.

Results and Discussion

Growing the fungal species tree

Recently, several groups have proposed fungal species trees based on the concatenated alignment of proteins selected from fully-sequenced genomes (Kuramae *et al.*, 2006b; Fitzpatrick *et al.*, 2006; Robbertse *et al.*, 2006; Kuramae *et al.*, 2007; Cornell *et al.*, 2007). The various studies considered different sets of species but used a similar method to select genes that were single-copy and widespread in their respective sets. A natural consequence of this methodology is that the number of genes considered in the phylogenetic analysis diminishes as the number of genomes included grows. In this way, the study of Robbertse et al (Robbertse *et al.*, 2006), limited to 17 ascomycota species, comprised 781 protein sequences (195,664 positions) in the alignment, whereas those of Kuramae et al (Kuramae *et al.*, 2006b) and Fitzpatrick et al (Fitzpatrick *et al.*, 2006), included, respectively, 531 genes (67,101 positions) for 24 species and 153 genes (38,000 positions) for 42 species. Remarkably, all these phylogenies are largely similar, at least for the set of species that they all have in common. Exceptions to this overall agreement include the phylogenetic position of *Stagonospora nodorum*, the relative branching order of *Candida glabrata* and *Saccharomyces castellii*, and some relative positions within the Candida genus.

We used a similar approach to reconstruct a broader fungal species tree including 60 fungi with completely-sequenced genomes (see table 3.1). To achieve this, we built a concatenated alignment of 69 widespread proteins that were present in at least 58 of the 60 species used and displayed one to one orthology relationships (see Material and Methods). The removal of positions with gaps in more than 50% of the sequences resulted in a trimmed alignment of 31,123 amino acid positions, which was subsequently used for Maximum Likelihood (ML) pylogenetic reconstruction, using a 4-rates gamma distribution model. Figure 3.2 shows the resulting tree, which is fairly congruent with previous fungal species trees.

Additionally, and to investigate the possible effects that the taxonomic sampling and the number of sequences involved may have in the final topology, we reconstructed three more species trees based on different sets of species. First, a well-sampled tree focusing on the 21 species from the Saccharomycotina group was built from a concatenated alignment of 1,137 widespread proteins. Next, another tree was built from the concatenation of 2007 proteins from the 12 species that belong to the Saccharomyces genus. Finally, a tree with the same number of species but each one sampled from the main fungal clades, was built using 217 concatenated alignments (see Material and Methods). We will refer to these fungal species trees as T60, T21, T12a and T12b, respectively. No major differences were encountered in terms of the relative topologies for the species they have in common between the different trees, see figure 3.3.

One tree fits all? : pattern pluralism within the yeast phylome

Many authors interpret the high level of similarity among different species trees as an indication that the proposed phylogeny reflects the real evolutionary relationships of the species included. A question that remains under discussion, however, is how well this tree represents the topological diversity encountered among trees from all the genes encoded in a genome. To evaluate this, we

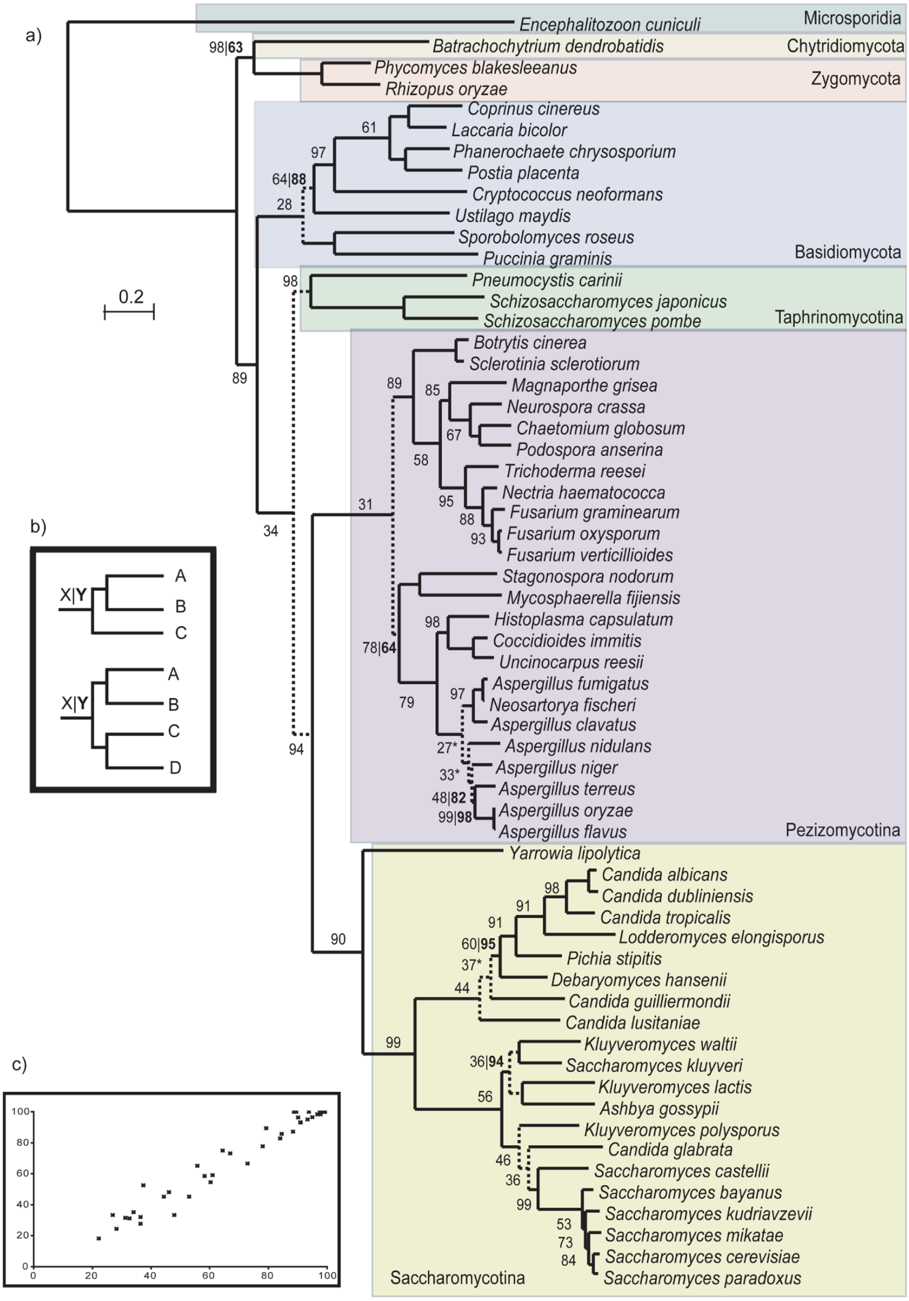

Figure 3.2: A) Phylogenetic tree representing the evolutionary relationships among the 60 fungal species considered in the study. The first number on each node indicates the phylome support for that node. An asterisk next to this number indicates that the topology obtained by the species tree is not the most common among the trees in the phylome. Whenever there is a second number (in bold), this indicates a bootstrap support lower than 100. Branches with dashed lines indicate evolutionary relationships that are supported by less than 50% of the trees in the phylome. B) Schematic representation of the two types of support values for the different nodes in the tree. C) Correlation between the fungal species tree topologies recovered by the individual trees included in the concatenated alignment (Y axis) and all the trees in the phylome (X axis). In both cases the fraction of trees that are compatible with a given topology, as computed with the topology scanning algorithm, is represented.

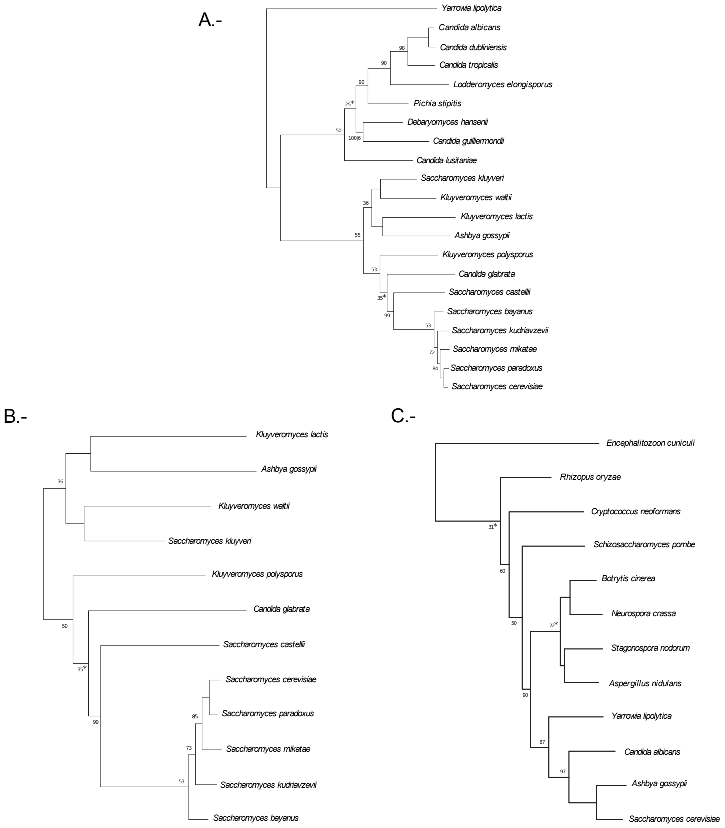

Figure 3.3: Fungal species trees. A) is the T21 species tree. In this case only 10 bootstrap repetitions were performed. B) represents the T12a fungal species tree. C) represents the T12b fungal species trees. Trees were reconstructed using the same methodology as the T60 species tree. Numbers represent the phylome support and the bootstrap values as explained in figure 3.2

reconstructed the complete collection of phylogenies of the genes encoded in the *S. cerevisiae* genome, that is, the yeast phylome. To do so, we applied a similar pipeline as the one used to reconstruct the human phylome (Huerta-Cepas *et al.*, 2007) (see Materials and Methods). We derived four versions of the yeast phylome that differ in their taxonomic scope and correspond to the species samples used in the species trees described earlier. The resulting 111,760 phylogenies and 22,352 alignments have been deposited in PhylomeDB (Huerta-Cepas *et al.*, 2008) (http://www.phylomedb.org; phylome codes SceP60, SceP21, SceP12a and SceP12b)

Some methodologies have been previously proposed to explicitly address the issue of concordance between species trees and individual gene trees (Ané *et al.*, 2007; Edwards *et al.*, 2007). Such approaches have been successfully applied to compare species trees of eight fungal species with the corresponding 106 phylogenies of widespread single-copy genes. However, these methods cannot account for gene phylogenies that include gene loss and duplication events and are not feasible for large datasets as the ones considered here. Here, we use a simple measure of concordance, which consists of evaluating whether the topology of each single gene tree is fully compatible with that proposed by the species tree (see Materials and Methods). Our results show that most individual gene phylogenies contain incompatibilities with the species tree. Of the 5804 trees of the phylome, only 410 (7.1%) were fully compatible with the topology in T60. Similar levels of congruence were observed for T21 (7.5%), whereas T12a showed a slight increase (12.3%). A marked improvement was observed in T12b (30.1%), suggesting that this set of distantly related species can be better resolved. These differences in congruence levels were generally similar when only partitions with high supports were considered (figure 3.4).

It must be noted that our measure for topological consistency (full compatibility) is highly stringent, since a single mismatch would render two trees inconsistent. Interestingly, when the consistency is evaluated for each single node in the species tree a different picture does emerge. Indeed, when the level of topological congruence is expressed for each specific internal node of the proposed species tree (figure 3.2), the result is a tree where most of the nodes (73%) show the topology that is most represented (>50%) among the trees in the phylome. Several conflicting nodes, in contrast, are supported by smaller percentages of the trees in the phylome. In three nodes the topology found by the tree of life is not even the most represented among the trees in the phylome (see figure 3.2). Nodes with low representation in the phylome do not always correspond to partitions that have low bootstrap values, indicating that bootstrap support in phylogenomic analyses can be misleading. These discrepancies cannot be explained by a topological bias in the sample used to reconstruct the species tree, since there is a high correlation between the topologies in the nodes of the sampled trees and that of the entire phylome (figure 3.2C). We

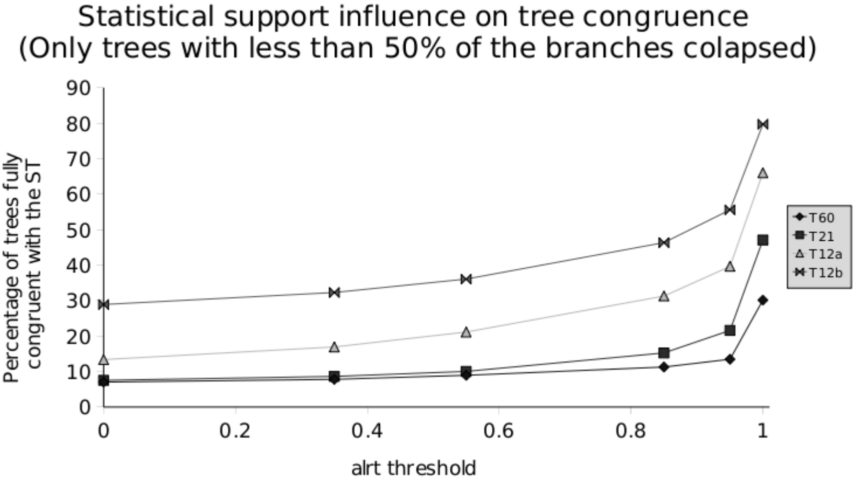

Figure 3.4: Percentage of trees in the phylome that are fully compatible with the topology of T60 (Y axis), at different statistical support thresholds in the nodes considered (X axis). The statistical support used is the minimum value of the approximate likelihood as computed by the Chi2-based paramteric methodology or the non-parametric branch support based on a Shimodaira-Hasenawa-like procedure (option -3 in PhyML aLRT). These methodologies are described in (Anisimova and Gascuel, 2006). The nodes with a statistical support lower than the given threshold were collapsed. We only considered trees in which less than 50% of the nodes were collapsed.

conclude from this analysis that, despite the high topological variation, species trees reconstructed from concatenated alignments do represent, at least for most of their nodes, the strongest phylogenetic signals observed along a genome. However, to properly reflect that some of the topologies are not widely supported by the majority of gene trees, we propose that these should be indicated by dashed lines. A reasonable cut-off could be set at 50%, as shown in figure 3.2. A more conservative decision could consist of collapsing these branches with low support, thereby introducing some polytomies. This will provide a less resolved species tree in which only dichotomies supported by a majority of the gene trees are shown.

Additionally, these under-represented nodes seem to correspond to topologies that are less robust to variations in taxonomic sampling. To assess this, we reconstructed nine additional species trees using randomly-chosen sets of 50, 40 and 30 species from our set. Combinations that did not contained the species *S. cerevisiae* and did not provide a set of at least 30 widespread proteins for the concatenation were discarded. In each case the tree was reconstructed from the concatenated alignment of the proteins that were widespread in the specific species sample. The three species trees with 30 species were fully congruent with T60. The remaining six trees did present slight topological variations in relation to T60 that mostly affected nodes with low support in the phylome. Of the 16 topological discrepancies with T60 found in these alternative trees, 13 (81%) affected nodes with support lower than 50%. The relative placement of *Debaryomyces hansenii* and *Aspergillus nidulans* within their respective groups and the position of Dothideomycetes species within the Pezizomycotina were the evolutionary relationships that were most affected by the taxonomic sampling.

Implications for phylogeny-based orthology prediction

Besides the reconstruction of species phylogenies, the existing high degree of topological variability in genome-wide data is likely to affect other applications of large-scale phylogenetic analyses. One of such applications is the large scale inference of phylogeny-based orthology predictions (Huerta-Cepas *et al.*, 2007; Ruan *et al.*, 2008; Gabaldón, 2008). Such phylogeny-based methods are being increasingly used and are considered more accurate than standard pairwise based methodologies (Gabaldón, 2008). There are two main approaches to infer orthology relationships from phylogenetic trees, namely reconciliation with the species tree (Zmasek and Eddy, 2001) and the use of species overlap information to ascertain whether a node represents a duplication or speciation event (Huerta-Cepas *et al.*, 2007). We previously suggested that species-overlap algorithms would be more appropriate to cope with the topological diversity in single-gene phylogenies (Huerta-Cepas *et al.*, 2007). To test this, we applied both a strict tree reconciliation method and our previously described species-

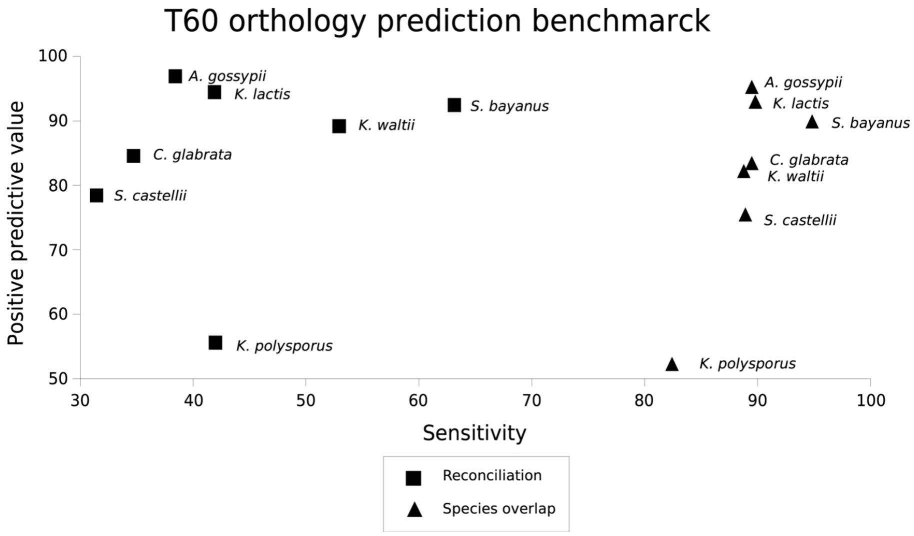

Figure 3.5: The synteny based and manually curated orthology predictions available at YGOB database (Byrne and Wolfe, 2006) is taken as a golden set to compute the number of true positives (TP), false positives (FP) and false negatives (FN) yielded by each method. For each method, the sensitivity S = TP/(TP+FN) and the positive predictive value P = TP/(TP+FP) are computed.

overlap algorithm to predict orthology relationships of all yeast genes. The orthology predictions from both methods were compared with the high-quality synteny-based orthology predictions from YGOB (Byrne and Wolfe, 2006). Although we observed no major differences in terms of positive predictive values between the two methods, there is a significant increase in terms of sensitivity when the species overlap algorithm is used figure 3.5. This algorithm correctly predicted 82–96% of the true orthology relationships as compared to 32–65% values reached by species reconciliation, indicating that a relaxed consideration of tree topology is more appropriate.

Lack of sufficient accuracy of current phylogenetic methods might explain a significant part of the topological diversity

Finally, we investigated some of the possible sources for the high topological variability observed. In principle, two main causes may be envisaged. First, some evolutionary processes such as horizontal gene transfer or gene duplication followed by differential gene loss may result in a divergent gene tree topology as compared to the actual species phylogeny. Alternatively, the topological variation might just be the result of insufficient accuracy of the methodology used. Two recent studies support the latter hypothesis by showing that different alignment reconstruction methods often result in different topologies (Wong *et al.*, 2008) and that trees reconstructed from longer alignments are more likely to conform to the species tree (Rasmussen and Kellis, 2007). In our case, we did not observe significant differences in terms of the length of the alignment, but our results confirmed that the use of different alignment methods significantly affected tree topology. For instance, when using the alternative programs MUSCLE (Edgar, 2004) and clustalw (Thompson *et al.*, 1994), only 7,22% of the trees had exactly the same topology. Moreover, we observed that the choice of the phylogenetic reconstruction method was also a source of variation. When comparing the trees produced using four alternative evolutionary models, we observed that only 9.9% of the trees presented the same topology in all models, and only 33% had two or more models pointing to the same topology. Thus, our results confirm previous findings (Wong *et al.*, 2008) that topological variation may result from alignment uncertainty and extend this conclusion to the case of uncertainty in the specification of an evolutionary model. Besides alignment uncertainty and model misspecification, many other methodological aspects such as the modelling of co-variation or the assignment of proportion of invariable sites are subject to uncertainty and thus may also affect the levels of topological variation. That the choice of different parameters or methodologies introduces topological variations in phylogenies reconstructed from exactly the same sequences and that the levels of variation are similar to those observed when comparing trees from different genes, suggest that the lack of sufficient accuracy of current phylogenetic methods is likely to be an important source for the observed topological variation. This is especially true when the methods are used automatically without carefully selecting the parameters. Alternatively, one might argue that the small overlap between the topologies resulting from the use of different models/alignment methods results from the fact that only one of the methods is accurate and able to reconstruct the underlying true phylogeny. To further assess the accuracy of the phylogenetic methods used here under a more controlled framework, we performed simulations of sequence evolution along the branches of the T60 tree. For this we used as a seed 50 yeast sequences and simulated their evolution using the program ROSE (Stoye *et al.*, 1998). Although in this case there is a true underlying phylogeny which is the same for all genes, in 70% of the cases, the phylogenetic reconstruction did not reconstruct the correct topology. A tree reconstructed from the concatenation of their alignments, however, was able to recover the original T60 topology.

Conclusions

Altogether, our results show that, despite high levels of topological variations, the gene-concatenation approach can fairly recover the strongest phylogenetic signals present across single-gene phylogenies. As a result, most of the nodes in such a species tree do represent topologies that are widely represented across the genome. Our analysis only reflects the topological variation found in the yeast phylome and thus phylogenies of genes not present in *S. cerevisiae* are not taken into account. However, we consider that the circa 6000 phylogenies used do provide a broad enough sample to assess the strength of the topology of the species tree. The fact that we found no significant bias in terms of node support for the set of widespread genes (Figure 3.2C), suggest that other species phylomes are likely to provide similar results.

Despite the overall high support for most of the nodes in the species tree, some partitions of the species tree include topologies that are poorly represented in the phylome. Additionally, such conflicting nodes are more prone to variations when different taxonomic samples are used and are therefore less certain to be correct. Levels of topological support across a complete phylome provide a direct approach to identify such conflicting nodes. This measure is completely independent of bootstrap analyses, which only provide information on the support of the different partitions from the alignment in which the tree is based. Thus, as we have identified in our analyses, high bootstrap supports do not necessarily indicate highly represented topologies. As a way to identify conflicting nodes and to incorporate genome-wide information on species trees, we propose to map gene-tree variability (phylome support) levels on the nodes of the species tree. This information could be used to mark, or eventually collapse, low represented (<50%) nodes so that our uncertainty on certain areas of the tree of life is properly represented. Moreover, our approach could be used to compare alternative phylogenomic approaches in terms of their representativeness across large samples of single-gene phylogenies. A firm candidate for this comparison is the super-tree approach, which combines information from single copy genes that should not necessarily be widespread (Pisani *et al.*, 2007). When used over fungal datasets, this approach has resulted in similar topologies to that produced by gene concatenation (Fitzpatrick *et al.*, 2006; Dutilh *et al.*, 2007), but the former were found to have less support in the literature (Dutilh *et al.*, 2007).

The high levels of topological variations in single-gene phylogenies combined with the uncertainty on the inferred species trees may mislead further phylogenetic analyses such as the inference of orthology. In this respect we have shown that a relaxed interpretation may overcome the pitfalls of a strict reconciliation algorithm. In this direction, reconciliation algorithms that incorporate uncertainty in the gene and the species trees (Berglund-Sonnhammer *et al.*, 2006) or species-overlap algorithms (Huerta-Cepas *et al.*, 2007; van der Heijden *et al.*, 2007) may represent promising alternatives to standard phylogeny-based methods to predict orthology. Finally, our results suggest that a significant part of the topological variation among gene-trees may result from methodological uncertainty. In this study we have used molecular data from fungal genomes. The conclusions raised here are likely to be valid for other eukaryotic phyla.

However, high levels of horizontal gene transfer across prokaryotic genomes, and perhaps certain unicellular eukaryotes, may invalidate the gene-concatenation approach as a means to infer a representative phylogeny. In such cases, besides the inherent levels of methodological noise discussed here, the topological variation in a genome will also reflect alternative evolutionary histories.

CHAPTER 4
TREEKO: A DUPLICATION-AWARE ALGORITHM FOR THE COMPARISON OF PHYLOGENETIC TREES

Marcet-Houben M and Gabaldón T. treeKO a duplication aware algorithm for the comparison of phylogenetic trees. (submitted)

Background

Phylogenetic trees represent evolutionary relationships among groups of species or biological sequences. The growing availability of sequence data from whole genomes, as well as the development of faster computers and more efficient phylogenetic programs, has facilitated the reconstruction of large collections of phylogenetic trees. In parallel, this has brought about the necessity of scaling up the analysis of phylogenetic trees to genomic scales (Gabaldón *et al.*, 2008). A recurrent analysis in phylogenetic studies is the comparison of the topologies of two or more phylogenetic trees. This is routinely used, for instance, to measure the support of tree partitions in a bootstrapping analysis or to compare alternative phylogenetic hypotheses (Felsenstein, 1985). Additional applications include, but are not limited to, the reconstruction of a single species tree from a number of individual gene trees (Bininda-Emonds, 2005), the evaluation of orthology inference (Altenhoff and Dessimoz, 2009), the detection of horizontal gene transfer events (Beiko and Hamilton, 2006), or the detection of host-pathogen co-evolution (de Vienne *et al.*, 2007).

Two main types of algorithms are available that compute topological distances between phylogenetic trees. A first class of algorithms uses the information directly from the topological arrangement of terminal nodes (i.e. leaves) in the trees. For instance, the popular Robinson-Foulds distance counts how many leaf splits are implied by only one of the compared trees (Robinson and Foulds, 1981). Similarly, the quartet distance is defined as the number of subsets of four leaves that are only implied by only one of the compared trees (Estabrook *et al.*, 1985). Finally, the so-called maximum agreement subtree is a similarity measure based on the largest subtree present in the two compared trees (de Vienne *et al.*, 2007). A second group of algorithms measure the minimal number of topological re-arrangements necessary to transform one topology into the other one. This is the rationale behind the transposition distance (Alberich *et al.*, 2008; Valiente, 2005), the prune and regraft distance (Wu, 2009) and the tree edit distance (Bille, 2005). Despite their extensively proven usefulness, all these algorithms share one limitation, namely the requirement that the mapping of leaves between the trees is complete and univocal (i.e. every leaf in one tree corresponds to only one leaf in the other tree). Thus, when applied to the comparison of gene family trees, these algorithms are unable to deal with trees that contain duplications (i.e. there is more than one gene per species) or losses (i.e. not all species are represented in both trees). As a result, only a reduced fraction of gene trees in a given phylogenomic study may be subject to analysis. For instance, Rasmussen and Kellis (Rasmussen and Kellis, 2007) found that for 9 fungal, and 12 drosophila species, only 739 and 5154 protein families, respectively, contained no duplications. Thus gene trees suitable for comparisons with the above mentioned methods will account for only 11% and 37% of their respective genomes. The fraction of tractable gene trees will decrease as the number of species in the set and their evolutionary distance increase.

Some algorithms have been designed to tackle this problem. For instance PhyloPattern (Gouret *et al.*, 2009) can search for trees containing a specific topological pattern, which may contain duplications. However this only serves to search for identical subtrees and does not provide a distance measure. A different approach was implemented by Puigbò et al. in the TOPD/FMTS program (Puigbò *et al.*, 2007), which is able to compare any pair of trees regardless of the number of duplications contained. While TOPD/FMTS does indeed provide an estimate of the distance, it has several drawbacks. First, whenever a species is represented by multiple sequences, TOPD/FMTS randomly prunes all but one to produce a single gene subtree without duplicated sequences. While this should ideally be done for each combination of duplicated genes, this becomes unfeasible for relatively small number of duplications. In such cases, TOPD/FMTS produces only a set of randomly chosen trees, which would provide an approximate, non-reproducible, distance measure. Besides this, the main limitation of TOPD/FMTS is that the inter-species orthology and paralogy relationships are not considered during the pruning process, resulting in pruned trees that contain a mixture of orthologous and paralogous sequences. These drawbacks hamper the interpretation of the distances provided by TOPD/FMTS. For instance, the comparison of identical trees containing some duplications will often provide distances greater than 0 (see comparative analysis below).

In order to address this important issue so that genome-wide collections of gene trees can be effectively compared, we have developed treeKO. TreeKO is a novel, duplication-aware algorithm that is able to compare two tree topologies regardless of the number of duplications and, at the same time, provide a distance measure that is evolutionarily meaningful.

Material and Methods

treeKO implementation.

treeKO was is implemented using the python programming language and the ETE programming toolkit. ETE and all its dependencies need to be installed, alternatively "a virtual machine" can be installed that already contains ETE and all its dependencies. TreeKO should then be downloaded within the virtual machine in order to be used. The input of treeKO are two bifurcated phylogenetic trees in which the species source for the different sequences is indicated (by default a three letters pre-fix is expected). The entry trees should be rooted or a rooting strategy indicated (midpoint rooting is used by default if an un-rooted tree is provided). Two files containing the two newick formatted trees to be compared need to be provided. An additional configuration file can be included in the treeKO command line in order to adapt some parameters to the users' trees. For more details check treeKO's web page, http://treeko.cgenomics.org/doku.php.

Strict distance

In order to assess the similarity between subtrees, both subtrees are pruned so that they contain the same species and then the RF distance is calculated for a pair of subtrees (d), as indicated by the following formula:

$$d = \frac{\left(\frac{RF}{RF_{max}}.r\right) + p}{r + p}$$

Where RF and RFmax represent the Robinson Fould distance between the two trees, and the maximum possible RF distance, respectively; r is the number of remaining leaves in the two subtrees after the pruning phase; and p the total number of leaves that were pruned. When comparing the two sets of subtrees from the original trees A and B, each subtree from tree A will be matched to the most similar subtree from tree B. Each subtree can only be matched once. If two pairs have the same distance, then other factors such as lower RF/RFmax ratio or a larger subtree size, in this order, are used to decide between pairs. Rejected subtrees have to be matched to a worse option or remain unpaired. Finally, a combined weighted distance is composed by two terms, one representing each initial tree (i and j), and is computed as follows:

$$D_{ij} = \frac{\left(\frac{\sum d.l}{\sum l}\right)_i + \left(\frac{\sum d.l}{\sum l}\right)_j}{2}$$

Where d represents the subtree distance between a pair of subtrees (see above) and l the number of leaves in the subtree. The unpaired subtrees are added to their corresponding term by assuming that d=1. A minimal distance of 0 will only be obtained when the two trees are identical in terms of topology, including the inferred duplication and loss events.

Speciation distance

In this case all subtrees from tree A are matched to their best subtrees in tree B, regardless of whether the best matched subtree has been previously matched. The subtree distance is calculated as explained above, but the normalized RF distance is not corrected by the number of pruned leaves (term p above), resulting in a simple normalized RF distance:

$$d = \frac{RF}{RF_{max}}$$

The weighted final distance between the two trees (Di,j) is computed as explained above. As a result, two trees with a speciation distance of 0 are not necessarily identical, but the inferred history of speciation events of the shared species will be the same (i.e they are fully congruent in terms of the inferred species tree).

Fungal species trees

The T12a and T60 fungal species tree published by Marcet-Houben and Gabaldón (Marcet-Houben and Gabaldón, 2009) was used as the reference tree topologies in the comparative analyzes. ETE was used to generate additional species tree topologies by swapping consecutive pairs of branches of the post-WGD species included in the tree (*S. cerevisiae*, *S. paradoxus, S. mikatae, S. kudriavzevii, S. bayanus, C. glabrata, S. castellii* and *Kluyveromyces polysporus*). A total of six alternative topologies were considered for each species tree.

Phylome to species trees comparison

The speciation distance implemented in treeKO was used to compare each tree in the P12a and P60 phylomes to each different species tree generating a distance distribution. The resulting distance distributions were compared with a t-test as implemented in R package.

Comparison with alternative methods

The phylomes described above, plus P12b and P21, based on different sets of species were also used to evaluate the number of trees that could be used by each tree comparison program. For the comparison between TOPD/FMTS and treeKO three sets of 100 trees were selected. The format was adapted using ETE for the TOPD/FMTS program. Each tree was compared to itself. TOPD/FMTS was run on default mode, calculating the split distance (equivalent to the RF distance) and generating a maximum of 100 subtrees, no random analysis was included.

Results and Discussion

Decomposing a tree into all possible subtrees by recursively splitting duplication nodes

The main rationale behind treeKo is the decomposition of the two gene-family trees to be compared into all possible subtrees so that every subtree is formed by the maximum number of sequences that are orthologous to each other, without including any paralogous sequence. This can be achieved by recursively splitting the gene family tree at each duplication node. Subsequently all subtrees produced by the two trees can be compared so that a weighted distance measure is produced (see below). We have implemented this algorithm using python and the ETE programming toolkit (Huerta-Cepas *et al.*, 2010a), and it is freely available here: http://treeko.cgenomics.org. The algorithm of tree decomposition is briefly described here (Figure 4.1), additional details can be found in the online documentation of treeKO.

Given a rooted tree, duplication nodes are detected by a species-overlap algorithm described earlier (Huerta-Cepas *et al.*, 2007), as implemented in ETE (Huerta-Cepas *et al.*, 2010a). In brief, this algorithm traverses the tree from every leaf in direction to the root. For every node, the species content of the two daughter branches are compared, nodes are considered speciation events if no species are shared, or duplication events otherwise. In a subsequent step, treeKO splits the original trees into single gene subtrees that contain only one of the paralogous partitions resulting from each duplication. To do so, treeKO traverses the tree and, for each duplication node, two subtrees are produced, each of which contains only one of the paralogous partitions that derive from the duplication. TreeKO will continue recursively on each resulting tree until all possible subtrees are generated from the combination of the different single gene partitions. Finally, the sets of subtrees produced by each input tree are compared to obtain a final topological distance.

Evolutionary-sound topological distances

Distances calculated by treeKO are based on the Robinson & Foulds distance (RF). In its current implementation, treeKO can compute two alternative distances measures strict distance and speciation distance (see Material and Methods). The strict distance is basically a weighted RF that, in addition, penalizes

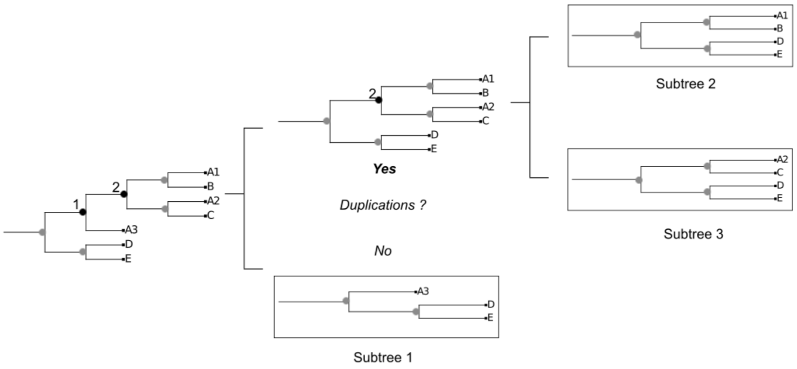

Figure 4.1: Subtree construction Example of how treeKO derives subtrees from a tree containing duplications. The initial tree (tree on the left) contains two duplication nodes (in black) marked as node 1 and node 2. treeKO splits the tree by node 1 and generate two different trees, each one of them containing one of the daughter partitions of node 1. This results in subtree 1 and an intermediate subtree that still contains a duplication (node 2). treeKO will then scan these subtrees for more duplications. In this case one of the subtrees has a second duplication and the subtree will be once again split and reconstructed, resulting in subtrees 2 and 3. treeKO will repeat this process until no resulting subtree contains further duplication nodes.

differences in evolutionary relevant events such as gene duplications and gene losses. In contrast, the speciation distance does not compute differences that can be attributed to duplication and loss events. This would be more appropriate for phylogenetic analyses that are focused on assessing the congruence of gene trees with a given species tree topology.

The two distances have been implemented keeping in mind the different applications of tree comparisons. The strict distance, which penalizes differential gene loss and duplication patterns, would be more appropriate when searching for protein families with a similar history of duplication, loss, and speciation events. Such searches are common in studies of co-evolution and inference of protein function. For instance, correlated gene loss have been used to predict functional interactions between mitochondrial proteins (Gabaldón and Huynen, 2005). In contrast, the speciation distance would fit better in studies where the main focus is the underlying species phylogeny. There are many possible applications in which the availability to compare gene trees in the presence of duplications will present an advantage. Here we present one such case in which the genome-wide support to alternative species phylogenies is explored by measuring the distance of each alternative species topology to the complete collection of phylogenies for all genes encoded in a given genome.

A practical use of TreeKO: Assessing the genome-wide support of alternative species tree topologies

The evolution of twelve completely sequenced yeast species, encompassing the Saccharomyces and the Kluyveromyces clades, is mostly well resolved (Marcet-Houben and Gabaldón, 2009). Only the relative order of divergence of *Candida glabrata* and *Saccharomyces castellii* species remains unresolved. Most phylogenomic studies support an earlier splitting of *C. glabrata* (Marcet-Houben and Gabaldón, 2009; Wang *et al.*, 2009; Fitzpatrick *et al.*, 2006). In contrast, analysis of chromosomal gene order has shown that the number of inversions that occurred during the evolution of these species is minimized in a scenario in which *S. castellii* diverges before *C. glabrata* (Gordon *et al.*, 2009).

Most species trees reconstructed from phylogenomic methods are evaluated with bootstrapping techniques that assess how stable a given topology is to random re-samplings of the input alignment. Since the input alignment contains generally a small fraction of the genes present in a genome, gene sampling effects may result in highly supported topologies which are not representative of the evolution of a given genome. An alternative strategy to study the evolution of a genome is to analyze the complete collection of phylogenetic trees of all of its genes (i.e the phylome). Evaluating a given species topology over such genome-wide set of gene-trees would provide a more accurate measure on whether it is fairly representative at a genome-wide scale.

Although phylomes have successfully been used to determine which nodes in the fungal species tree are most congruent at genomic-scales (Marcet-Houben and Gabaldón, 2009), there is as yet no quantitative measure of the levels of similarity between a given species tree and a complete phylome. Here, we address the question of whether the distributions of speciation distances to a given phylome can be used to decide among alternative evolutionary scenarios. For this we used treeKO to compute the speciation distances of the yeast phylome (P12a dataset described in (Marcet-Houben and Gabaldón, 2009) and available at PhylomeDB (Huerta-Cepas *et al.*, 2008)) against different alternative species trees. Gene trees were rooted with midpoint so that no assumptions on the species tree topology were made a priori. Alternative topologies were derived from a reference species tree by swapping pairs of neighboring branches (i.e. interchanging the positions of Saccharomyces paradoxus and Saccharomyces mikatae or the positions of *C. glabrata* and *S. castellii* as shown in Figure 4.2). Resulting distance distributions were compared with a t-test.

As seen in figure 4.2, the swapping of the well-supported *S. paradoxus* and *S. mikatae* branches (alternative topology 2) resulted in significantly larger distances (p-value < 2x10-16), whereas inter-changing the controversial positions of *C. glabrata* and *S. castellii* (alternative topology 1) presented distances that were not significantly different (p-value = 0,0955).

We extended the analysis to all the topologies created by swapping consecutive pairs of post-whole genome duplication (post-WGD) species. In all cases, the distances obtained were significantly larger than the one obtained from the reference tree. The size of the trees affects the final distance and as such the same difference in trees of different sizes will contribute differently to the final distance. Similar results were obtained when the phylome and the species trees considered were based on a broader taxonomic range of 60 fungal species (see methods)

Thus, this test would discard all alternative topologies tested except the one that exchanges *C. glabrata* and *S. castelli* positions. These results are congruent with the ones observed before, where the *C.glabrata/S.castellii* node was the only one involving post-WGD species, that displayed a low phylome support (Marcet-Houben and Gabaldón, 2009), and indicate that gene family sampling may have influenced previous phylogenomic studies. We believe that this type of comparisons provides an alternative way to evaluate, in a statistically sound manner, the genome-wide support for alternative phylogenetic hypotheses.

Comparison to other distance measures

To show that these type of analyzes, ensuring a genome-wide coverage, are difficult with existing methods. We compared the performance of treeKO and that of 1) a standard RF measure, 2) a RF measure after a pruning phase to

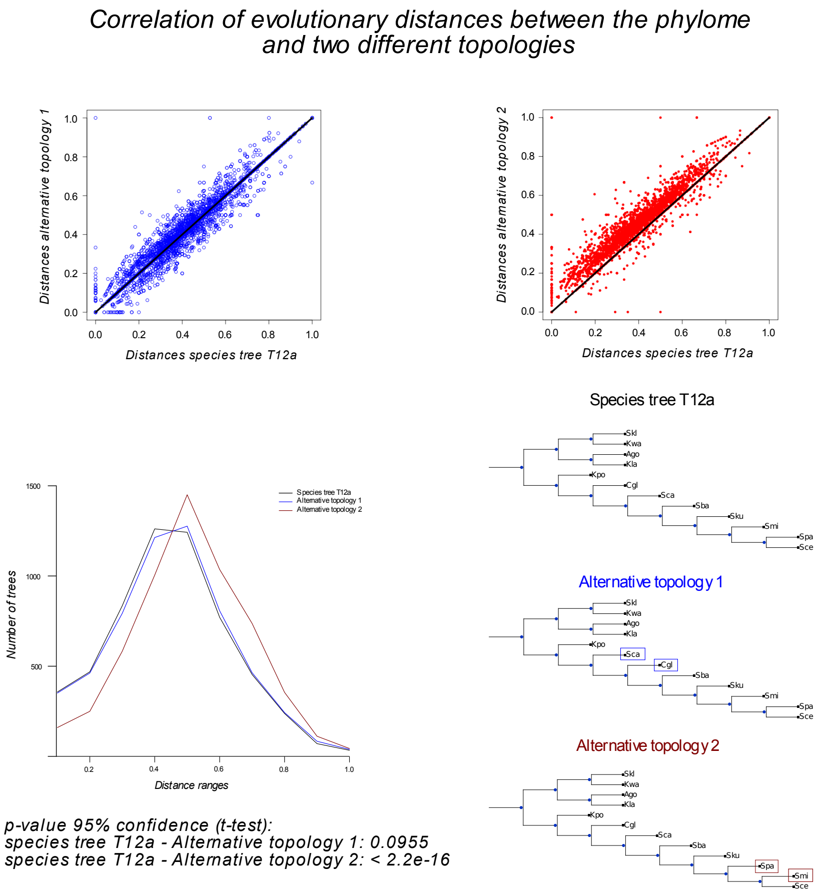

Figure 4.2: Distribution of distances between trees in P12a phylome and three alternative species trees. The bottom right part of the figure shows the three topologies used. The first one is the T12a tree while the other two represent changes in this topology. Alternative topology 1 represents a change in a poorly supported node while Alternative topology 2 represents a well supported node. The two upper graphs plot each distribution of distances of the alternative topologies against the reference T12a topology. The lower left panel represents the frequency graph for the three distance distributions.

Phylome	TreeKO	TOPD/FMTS	RF	RF + pruning
P60	100%	100%	0%	22%
P21	100%	100%	0%	36%
P12a	100%	100%	14%	38%
P12b	100%	100%	2%	27%

Table 4.1: Percentage of trees each program can compare to its species tree. Columns represent the four compared programs: treeKO, TOPD/FMTS, RF and RF with an initial pruning step. Rows represent each of the fours yeast phylomes with different taxonomic coverage that can be found in phylomeDB (Huerta-Cepas *et al.*, 2008).

delete species that were not present in both, gene tree and species tree, and, 3) measure provided by TOPD/FMTS. In all cases we evaluated the number of trees suitable to analysis in the yeast phylome according to each method. Additionally, for TOPD/FMTS, the distance obtained when comparing a tree to itself and the average computing time were calculated. An ideal method to perform genome-wide analyses would be fast, able to compare all gene trees and would produce a distance measure that is easy to interpret (e.g. identical trees would provide a distance of 0).

The main disadvantage of RF distances as implemented in standard programs such as Ktreedist (Soria-Carrasco *et al.*, 2007) or PHYLIP (Retief, 2000), is that only a minor fraction (~15%) of gene trees are suitable to analysis. This is improved by the inclusion of a pruning step (~40%). As expected, treeKO and TOPD/FMTS are able to achieve 100% coverage. As seen in table 4.1, when computing the percentage of trees each program can use, the values very significantly between larger phylomes or those that have a broader taxonomic scope.

To compare the performance of treeKO, TOPD/FMTS and Ktreedist we randomly took three sets of 100 trees each from the phylome and compared them against themselves. For Ktreedist only single gene trees were compared. The programs were compared in terms of time consumption and average distance between two identical trees. While distances calculated by treeKO and RF are null for identical trees, TOPD/FMTS reports an average split distance of 0,41. If we divide the trees in single gene trees and multi gene trees, we see that the distance calculated by TOPD/FMTS for the first group is 0 but that the average distance for the multi-gene trees is of 0,67. Non-0 distances for identical trees are difficult to interpret and, therefore, not desired.

Finally, as seen in table 4.2, RF is by far the fastest program of the three. For single gene trees there is not much difference between treeKO and TOPD/FMTS. On the other hand, for multi gene trees treeKO is 8-9 times faster than TOPD/FMTS. This could be attributed to the fact that the number of subtrees generated by treeKO is much lower than in TOPD/FMTS, even when

	TreeKO			**TOPD/FMTS**			**RF**		
	Set1	Set2	Set3	Set1	Set2	Set3	Set1	Set2	Set3
Percentage of trees compared	100%	100%	100%	100%	100%	100%	34%	43%	41%
Average time consumption per tree (s)	1.31	1.65	1.95	10.38	8.84	9.57	-	-	-
Average time consumption per single-gene tree (s)	1.09	1.09	1.15	2.12	2.26	2.07	0.06	0.07	0.05
Average time consumption per multi-gene tree (s)	2.62	3.22	3.42	22.00	21.05	22.11	-	-	-
Average distance	0	0	0	0,45	0,36	0,43	-	-	-
Average distance single gene trees	0	0	0	0	0	0	0	0	0
Average distance multiple gene trees	0	0	0	0,70	0,61	0,69	-	-	-

Table 4.2: Comparison between three tree comparison programs (treeKO, TOPD/FMTS and RF). Three sets containing 100 randomly chosen trees of the P12a phylome were used for comparison. Columns represent one of the sets of trees and a program. Rows contain data regarding the percentage of trees that were compared, the time consumption (expressed in seconds) and the average distance between pairs of identical trees. Data on separated by single-gene and multi-gene trees is also provided.

this program only takes 100 random subtrees into account. This is the result of considering duplication nodes as the splitting point for the subtree generation instead of taking each species independently.

Concluding remarks

We have developed treeKO, a tree comparison tool that is able to compare trees in the presence of duplication and loss events. By addressing the main limitations of previous approaches, treeKO enables the comparison of genome-wide phylogenetic datasets. treeKO provides two alternative RF-based distance metrics that are biologically sound and specifically adapted to applications such as the search for co-evolving protein families or the assessment of topological congruence in the inferred order of speciation events. The use of treeKO opens the door to novel phylogenomic analyses such as the one presented here that evaluates whether differences in genome-wide support to two alternative topologies are statistically significant.

CHAPTER 5
ACQUISITION OF PROKARYOTIC GENES BY FUNGAL GENOMES

Marcet-Houben M and Gabaldón T. 2010. Acquisition of prokaryotic genes by fungal genomes. Trends Genet. 26(1):5-8

Horizontal gene transfer in eukaryotes

Horizontal gene transfer (HGT), the exchange of genetic material between two species (Syvanen, 1985), was discovered 50 years ago (Akiba *et al.*, 1960), but it is the current wealth of genomic sequences that is revealing its real impact. Currently, it is widely accepted that HGT is one of the main evolutionary driving forces in prokaryotes, but its relevance for eukaryotes remains controversial (Kurland, 2005; Andersson, 2005; Keeling and Palmer, 2008). The traditional view that HGT in eukaryotes is virtually absent has been recently challenged by genome-wide analyses reporting the transfer of bacterial genes to organisms such as ciliates (Ricard *et al.*, 2006) or rotifers (Gladyshev *et al.*, 2008). The presence of acquired genes can sometimes be associated with important evolutionary adaptations. This is the case for a set of carbohydrate metabolism genes acquired by some ciliates (Ricard *et al.*, 2006) and fungi (Garcia-Vallvé *et al.*, 2000) during their adaptation to the ruminant gut. Despite such findings, most newly sequenced eukaryotic genomes are not screened for HGT events, and there is as yet no comprehensive analysis of the impact of this process at broad taxonomic levels.

Fungi are one of the best sampled eukaryotic groups in terms of fully sequenced genomes (Galagan *et al.*, 2005; Marcet-Houben and Gabaldón, 2009). Moreover, many fungi have a saprophytic or symbiotic lifestyle, which involves close interactions with bacteria, and lack some of the classical barriers to HGT such as the differentiation of germ line and soma. Therefore, they are theoretically more likely to undergo a higher frequency of HGT. Although several small studies have identified prokaryotic genes in fungi (Garcia-Vallvé *et al.*, 2000; Uo *et al.*, 2001; Hall and Dietrich, 2007; Gojković *et al.*, 2004), few entire genomes – mainly from Saccharomycotina species – have been systematically searched for HGT (Hall *et al.*, 2005; Fitzpatrick *et al.*, 2008). To assess the impact of inter-domain HGT in the fungal kingdom, we performed a comprehensive phylogenomic search for acquired prokaryotic genes in 60 fungal genomes. Our results show that HGT has affected most fungal clades to various degrees and suggest that inter-domain gene transfers are important for the evolution of fungi.

Materials and Methods

Sequence sources

Two different proteome databases were used to conduct this analysis. A fungal proteome database, containing all proteins for 60 completely sequenced fungi downloaded from different databases (see Table 3.1) and a second database, comprising all the proteomes found in the KEGG database as of February 2008 (Kanehisa *et al.*, 2008). This includes 595 bacterial and archaeal species and 36 eukaryotic species.

HGT candidate selection

We used blast (Altschul *et al.*, 1997) to detect proteins that were not widespread among fungal species. Starting from *Saccharomyces cerevisiae*, we recursively went through all fungal species performing blast searches using all proteins as seed, excluding those that had already appeared as hits in previous searches. Blast results were filtered based on an e-value threshold of 5e-04 and a continuous overlap threshold of 33% with the query protein. Proteins found in 10 or less species were selected as possible HGT candidates. These 126,694 proteins were then compared to the prokaryotic database using Blast. Proteins which produced more than 30 hits in prokaryotic species and not in other eukaryotes were selected for further study.

Phylogenetic analyses

We reconstructed phylogenetic trees based on the analysis of the HGT candidates and their closest 150 prokaryotic homologs. Phylogenetic reconstruction used the same pipeline described in the human phylome (Huerta-Cepas *et al.*, 2007). Briefly, the selected sets of homologous proteins were aligned with MUSCLE 3.6 (Edgar, 2004) and positions in the alignment with gaps in more than 10% of the sequences were trimmed with trimAl (Capella-Gutiérrez *et al.*, 2009) (http://trimal.cgenomics.org). Neighbor Joining trees were derived using scoredist distances as implemented in BioNJ. PhyML aLRT (Anisimova and Gascuel, 2006) version was used in to derive Maximum Likelihood (ML) trees. Four different evolutionary models were used for each seed sequence (JTT, WAG, Blosum62 and VT). In all cases, a discrete gamma-distribution model with four rate categories plus invariant positions was used, estimating the gamma parameter and the fraction of invariant positions from the data. The evolutionary model best fitting the data was determined by comparing the likelihood of the used models according to the AIC criterion (Akaike, 1973), and used in further analysis.

Evolutionary mapping of HGT events

To map HGT events on the fungal tree of life, the most parsimonious scenario was always considered. When the transferred gene was found in a single fungal species the transfer was assumed to have occurred in that specific lineage. Candidates with hits in multiple fungal species were analysed as to ascertain whether the fungal proteins were grouped into a monophyletic group. If this was not the case, it was assumed that the protein was transferred in multiple independent events. All events were mapped into the fungal species tree assuming that the common ancestor of all species present in the cluster was the initial recipient of the transferred gene. Alternatively, a ratio between HGT events and loss events could have been used to describe transference events (Snel *et al.*, 2002). For instance, using a ratio of 1:3 (one transference for three losses) a higher number of transfers was predicted, 313 events versus the 235 predicted selecting just monophyletic groups (Supplementary figure 6). However, this did not alter the main findings of our study.

Functional analyses of putative transferred genes

Functional prediction of the putative transferred genes was based on phylogeny. KEGG annotations of the prokaryotic sequences found in the sister branches of the transferred proteins were used to infer the putative function of the fungal HGT candidates. The programs SignalP, TargetP and TMHMM (Emanuelsson *et al.*, 2007) were used to predict secreted proteins among the candidates. Secretion was predicted for those proteins that had a signal peptide (SignalP), were predicted as secreted (TargetP) and had no transmembrane helices (TMHMM). The prediction has been added in Supplementary table 3 as part of the putative function annotation.

Results

Detection of inter-domain HGT events in fungi

Using 60 complete fungal genomes and >600 other genomes from prokaryotes and other eukaryotes, we designed a phylogenomic pipeline to detect cases of genes likely to have been acquired through HGT from prokaryotic donors (see the materials and methods). In brief, we adopted a conservative strategy that searched for genes present in few (<10) fungi and absent in other eukaryotes, but which could be found in a relatively high number of prokaryotic genomes (>30). This procedure detected 713 acquired genes in 53 (88%) of the analyzed genomes (Tables S3–6 in the online supplementary material). This number represents 0.12% of the genes analyzed, ranging from 0 (most Saccharomycetales) to 0.38% (*Aspergillus flavus*) of the genes in a given fungal genome. A prokaryote-to-fungi direction of all these transfers is suggested by a more taxonomically diverse distribution of blast hits in prokaryotic genomes compared with that in fungal genomes. Moreover, in most cases the number of hits in bacterial genomes is much higher than the number of hits in fungal genomes, even when differences in the composition of the initial dataset are taken into account (see Table S6 in online supplementary material). In total, 40% of the putatively transferred genes contained introns. Interestingly, the fraction of transferred genes with introns and their intron density was lower than the average values in the recipient genome, but a correlation was observed (see Figure S1 in the online supplementary material), suggesting that acquired genes are gaining introns in a lineage-specific fashion.

The adaptation to the recipient genomes is also suggested by similar values in terms of the codon adaptation index or GC content. Transferred genes could be grouped into 161 distinct families, on which we performed phylogenetic analyses to infer possible donors and functions and to assign the transfer to a specific lineage. Transferred genes were grouped into 235 monophyletic groups, suggesting at least that amount of transfer events, although more events could be inferred if a gene loss threshold is applied (see online supplementary material). Conversely, this estimate would overestimate transfers because some genes would have been transferred in a single event (discussed below). The species composition of the most supported sister branch in our phylogenetic analyses is provided for each transfer event in Table S3 (see online supplementary material). These putative donor groups are only indicative because uncertainty in the phylogenetic analyses and bias in the available genomic data prevent us from making unambiguous assignments.

Differential impact of HGT across the fungal tree of life

We mapped the 235 HGT events on the fungal tree of life (Marcet-Houben and Gabaldón, 2009), assuming the transfers affected the last common ancestor of all species in a monophyletic group. Our results (Figure 5.1) show that HGT affected most fungal lineages to various extents. Unexpectedly, nearly two-thirds (65%) of the events are mapped within Pezizomycotina, of which one-third appear at the base of this clade. More recent HGT events affect most

Pezizomycotina lineages, suggesting this is an ongoing process. By contrast, HGT events in Saccharomycotina seem to have been rather scarce with only 13 cases. Such large differences in the extent of HGT between these two Ascomycota groups are striking. The difference between the two groups was robust to our filtering criteria and is not the result of taxonomic bias in our dataset (see Figure S5 in online supplementary material). One main difference between the two groups is their genome size. However, when this is taken into account, different HGT rates remain apparent (Table S4 in the online supplementary material). Another peculiarity of Saccharomycotina genomes is that they tend to be packed, possibly complicating the insertion of foreign DNA in non-coding regions. Nevertheless, because high gene density is also common in prokaryotes, it cannot be considered by itself a strong barrier to HGT. Different lifestyles might explain the observed differences; however, the extreme diversity of lifestyles in Pezizomycotina (Spatafora *et al.*, 2006) makes it difficult to pinpoint a specific characteristic that might render them more amenable to HGT. Finally, a possible function of the transferred genes was inferred according to their closest prokaryotic homologs (see Table S3 in online supplementary material). This information can provide useful hints in assessing the possible role of the transferred gene in the recipient species and thereby explain potential evolutionary advantages. However, this task is not straightforward and might constitute a whole study in each case. Here, we highlight a few cases that we deemed of special relevance.

Restoration of the arsenate detoxification pathway

We identified two independent acquisitions of a bacterial arsenate reductase in *Yarrowia lipolytica* and *Rizopus oryzae*. The reduction of arsenate to arsenite is an important step in arsenic detoxification (Rosen, 1999), a pathway that is carried out in the yeast *Saccharomyces cerevisiae* by a different arsenate reductase (ARR2) in combination with membrane pumps that expel arsenite from the cytoplasm. Orthology assessment (Marcet-Houben and Gabaldón, 2009; Huerta-Cepas *et al.*, 2008) shows that *Y. lipolytica*, *R. oryzae* and their close relatives have no orthologs to the yeast arsenite reductase, whereas the remaining arsenic detoxification machinery is present. Thus, the acquisition of a bacterial arsenate reductase in these species might have provided them with the ability to detoxify arsenics. The fact that this transfer occurred twice independently is remarkable. Such independent origin is not only supported by phylogenetic analyses (see Figure S2 in online supplementary material), but also by a specific fusion of this enzyme to only one of the two recent duplicates of uracil phosphoribosyltransferase found in *R. oryzae*, suggesting that the transfer occurred after the recent whole genome duplication in this species (Ma *et al.*, 2009).

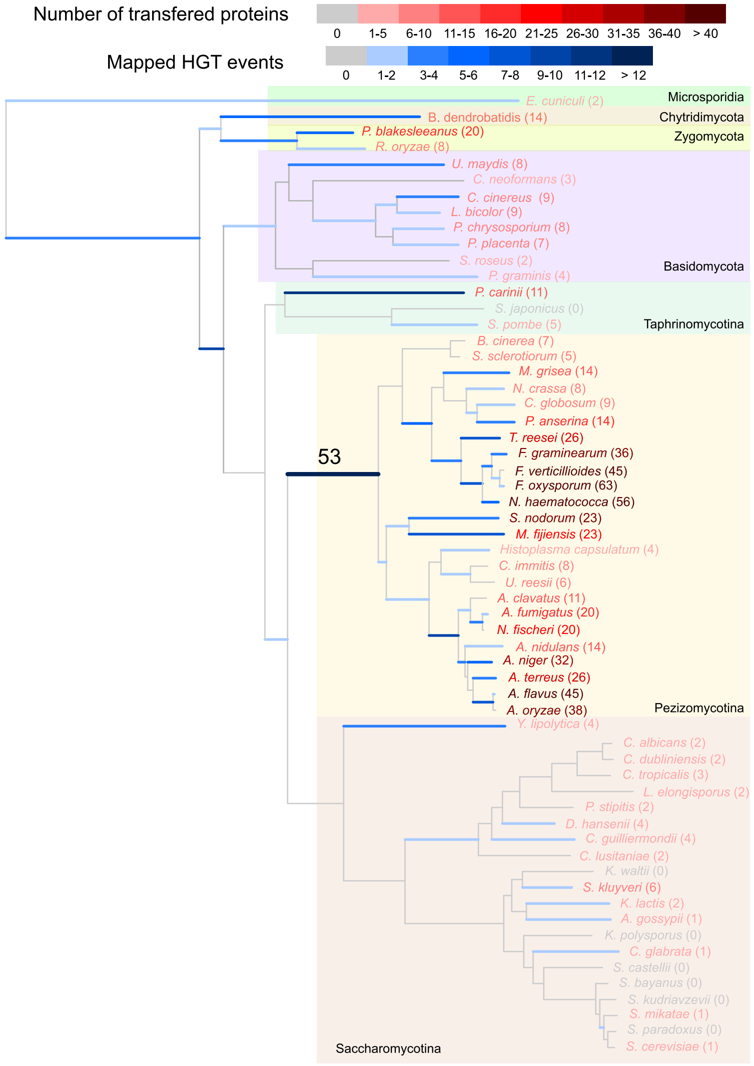

Figure 5.1: Fungal species tree showing the predicted transfer events. The species tree is based on a maximum likelihood analysis of a concatenation of 69 widespread proteins (Marcet-Houben and Gabaldón, 2009). The number of monophyletic transfer events at each lineage is marked in different shades of blue. Lineages without any mapped transferences are shown in gray. The number of transferred proteins in each species is indicated by a color scale in the species name. A particularly high ratio of putative horizontal transfers from bacteria is found in the Pezizomycotina group, with 53 events mapped to the last common ancestor of this clade. Later, lineage-specific transfers indicate that this is an ongoing process. This high rate of inter-domain HGT in Pezizomycotina contrasts with a low incidence in Saccharomycotina, another densely sampled fungal group.

Acquisition of multiple racemase genes

Racemases catalyze the inter-conversion of optical isomers (e.g. D- and L-amino acids). Recent studies describe putative cases of HGT involving various bacterial racemases in eukaryotic genomes. These include alanine racemase in *Schizosaccharomyces pombe* (Uo *et al.*, 2001) and *Adineta vaga* (Gladyshev *et al.*, 2008), and a proline racemase in *C. parapsilosis* (Fitzpatrick *et al.*, 2008). Interestingly, our survey detected seven additional HGT instances involving three different racemases: aspartate, hydantoin and mandelate racemases. Most such cases involve species from or around the Fusarium and Aspergillus groups, with the exception of an aspartate racemase in *Candida glabrata* (see Figure S3 in online supplementary material). This high number of transfer events of bacterial racemases is noteworthy. None of the transferred racemases have been characterized functionally so their specific physiological role in the receiving species is unclear. Possible roles might involve either the detoxification or assimilation of D-isomers. Indeed, besides forming part of bacterial cell walls, D-amino acids has been described in other natural tissues, and growth on D-amino acids has been reported for certain fungi (Pollegioni *et al.*, 2007).

Bacterial catalases in fungal pathogenic species

Through their action in decomposing reactive oxygen species, catalases can help pathogens overcome host defense mechanisms. The potential advantages conferred by the acquisition of foreign catalases have been extensively studied in pathogenic bacteria (Faguy and Doolittle, 2000; Klotz and Loewen, 2003). In fungi, a recent transfer of a bacterial catalase has been described in the microsporidian pathogen *Nosema locustae* (Fast *et al.*, 2003). In our study, we detected similar transfers to three important plant pathogens: the Dothideomycetes *Stagonospora nodorum* and *Mycosphaerella fijiensis* and the Leotiomycetes *Botrytis cinerea*. The large evolutionary distance between the first two species and *B. cinerea* and our phylogenetic analysis (see Figure S4 in online supplementary material) suggest that *B. cinerea* acquired the bacterial catalase from a Dothideomycetes. Interestingly, another plant pathogen *Pseudomonas syringae* is placed among the possible bacterial donors of this gene, which suggests the transfer of potential virulence factors among pathogens.

Transfer of a Mur operon

Multiple transfers of functionally-related enzymes are remarkable because they can provide the recipient species with a novel metabolic pathway. Our analyses found, in similar sets of Aspergillus species, the first three enzymes of the bacterial peptidoglycan biosynthesis pathway: MurA, MurB and MurC. Recipient species included *Aspergillus terreus*, *Aspergillus oryzae*, *Aspergillus* *flavus*, *Aspergillus fumigatus* and *Neosartorya fischeri*, although MurC was apparently lost in the latter two species. Interestingly, in *A. terreus*, MurB and MurC are fused into a single gene, with MurA located just besides. This suggests that all three genes might have been part of a bacterial operon at the time of the transfer. According to the STRING database (Jensen *et al.*, 2009), MurA, MurB and MurC have a conserved gene order only in some Firmicutes. However, our phylogenetic analyses only identify Firmicutes as the possible donor for MurA, whereas the other genes have distinct phylogenetic affiliations. Three concomitant transfers of related enzymes from different donors seems an unlikely scenario, and we are more inclined towards the possibility that an insufficient phylogenetic signal, absence from the genome databases of relatives of the true donor and HGT events among bacterial organisms preceding the transfer to fungi, have obscured the specific origin of the transferred gene. This view is consistent with the high level of heterogeneity in terms of bacterial groups found in phylogenetic proximity to the transferred genes in many of our analyses (see Table S3 in online supplementary material). Finally, the function for these three Mur enzymes in fungi is unclear, although a possible role in the degradation of bacterial cell walls can be speculated.

Concluding remarks

Our results reveal that inter-domain HGT is widespread in fungi and has played a role in the evolution of this eukaryotic group. We detected 713 transferred genes that, given the high stringency of our phylogenomic criteria, should be considered a minimal estimate. Additional studies will probably reveal further examples of transfers that have escaped our strict criteria (e.g. prokaryotic genes transferred to other eukaryotic groups). The ability of fungi to acquire alien prokaryotic genes suggests a more extensive exchange of genes between fungal species. Gene transfer between fungi has already been described (Khaldi *et al.*, 2008), and our study suggests at least an additional case. Further research will be necessary to address many open questions such as the impact of these transfers on the evolution of the recipient species and the intriguing differences in the extent of HGT found for different fungal clades.

Additional files can be found here:

http://www.cell.com/trends/genetics/supplemental/S0168-9525%2809%2900235-2

CHAPTER 6
PHYLOGENOMICS OF THE OXIDATIVE PHOSPHORYLATION IN FUNGI REVEALS EXTENSIVE GENE DUPLICATION FOLLOWED BY FUNCTIONAL DIVERGENCE

Marcet-Houben M, Marceddu G. and Gabaldón T. 2009. Phylogenomics of the oxidative phosphorylation in fungi reveals extensive gene duplication followed by functional divergence. BMC Evol Biol. Dec 21;9:295

Introduction

Oxidative phosphorylation (OXPHOS) is the primary energy-producing pathway in aerobic organisms (Saraste, 1999). It functions by coupling the energy obtained from the oxidation of certain metabolic substrates to the phosphorylation of adenosine biphosphate (ADP) to produce ATP. This is achieved by a process of electronic transference through an intricate assembly of more than 20 discrete carriers. These carriers are mainly grouped into four membrane-embedded protein complexes, named Complex I through Complex IV, which form the electron transport chain (ETC). Some of the complexes in this chain are able to use the energy liberated by the electron transfer to the pumping of protons across the membrane, thereby generating a proton gradient. Finally, the energy obtained from the dissipation of this gradient is used by a fifth protein complex, ATP-synthase or Complex V, to synthesize ATP.

In eukaryotes, the oxidative phosphorylation machinery resides in the inner membrane of the mitochondrion. Molecular phylogenies of eukaryotic OXPHOS components indicate that the core subunits of the complexes were inherited from the alpha-proteobacterial ancestor of mitochondria (Gabaldón and Huynen, 2003, 2004). In contrast, other subunits might have different origins and show complex phylogenetic distributions (Gabaldón, 2005). Besides providing important information on how complex systems evolve, knowledge about lineage-specific variations may serve to identify novel components or interactions. For instance, the evolutionary analysis of Complex I across a set of eighteen eukaryotes, led to the prediction that the so-far uncharacterised human protein B17L was involved in Complex I function (Gabaldón, 2005).

This protein was later found to be participating as a chaperone in Complex I assembly and a mutation in this gene was identified in patients showing severe encephalopathy (Ogilvie *et al.*, 2005).

Fungi is the group of eukaryotic organisms that is best sampled in terms of fully sequenced genomes (Galagan *et al.*, 2005; Marcet-Houben and Gabaldón, 2009). The adaptation of this kingdom to a diversity of environments is reflected in a high metabolic variability that also affects the respiratory pathway (Gabaldón and Huynen, 2004; Bullerwell and Lang, 2005). Indeed, the adaptation to oxygen-limited conditions or to high levels of oxidative stress during certain phases of their life cycle may have been crucial in the emergence of fermentative or pathogenic lifestyles. A recent comparative genomics study (Lavín *et al.*, 2008) has provided a comprehensive view of the patterns of presence and absence of OXPHOS components in 27 fungal species. Here we extend the analyses to 60 fully-sequenced fungal genomes and use a phylogenetics approach that enables us not only to obtain reliable orthology relationships but also to trace the history of duplications of OXPHOS components and related pathways during fungal evolution. In particular, we wanted to assess the role that gene duplication and functional divergence has played in the evolution of this pathway. A prediction of the gene-balance hypothesis is that independent duplications of protein complexes are likely to have deleterious effects (Papp *et al.*, 2003), thereby constraining this mode of evolution in a pathway that is mostly composed of large complexes. Moreover, we wanted to test whether some loss or duplications of OXPHOS components could be associated to specific phenotypes such as virulence or adaptation to anaerobic environments. Altogether, our results show a relatively high rate of duplication events that affect 76% of the protein families surveyed. Interestingly, some of these duplications have been directly followed by processes of functional divergence, sometimes involving the recruitment of one of the duplicates to other multi-protein complexes.

Methods

Sequence data

Proteins encoded in 60 fully-sequenced fungal genomes were downloaded from several databases. For consistency, we used in our analysis the species names as provided by the database source. Some of these species have been renamed and the corresponding new names and synonyms are listed in the additional file 1 (Additional table S1). Additionally, genomes from Homo sapiens and Arabidopsis thaliana were downloaded from ensembl http://www.ensembl.org. The final database comprises 626,834 unique protein sequences.

Reconstruction of the presence/absence matrix

Fungal proteins annotated as being part of the OXPHOS pathway were downloaded from the KEGG database (map 00190) (Kanehisa *et al.*, 2008). In addition, 6 proteins that were identified in the literature as belonging to complex I but were not present in the KEGG database were downloaded from UniProt and included in the analyses (NI9M, NURM, NUWM, NUXM and NUZM). The resulting 85 proteins were used to perform a blast search against a database of fungal proteins encoded in 60 fungal genomes (see table 3.1). Low complexity filters were used in the blast search. To detect homology, we used the same parameters that have been used previously in the same taxonomic range (Marcet-Houben and Gabaldón, 2009). In brief, only significant hits (E-val < 10-3) that aligned with a continuous region covering more than one third of the query sequence were selected. Note that the use of low complexity filters in the blast can reduce significantly the length of continuous regions of homology. Sets of homologous sequences were aligned and used to reconstruct a Maximum Likelihood tree from which orthology relationships were inferred (see below). These orthology relationships were used to build a presence/absence matrix in which for each OXPHOS component the species with a corresponding ortholog are indicated. Putative absences in the matrix were double-checked by tBlastN searches against the corresponding genome sequence and Blast searches from family members of more related species. These hits were checked manually and whenever they were considered orthologous to the already identified members they were added to the list.

Phylogenetic analyses

We used a similar pipeline to that described in (Huerta-Cepas *et al.*, 2008). Sets of homologous proteins were aligned using MUSCLE 3.6 (Edgar, 2004) with default parameters. Positions in the alignment with gaps in more than 10% of the sequences were trimmed with trimAl (Capella-Gutiérrez *et al.*, 2009). Finally, PhyML aLRT version (Guindon and Gascuel, 2003; Anisimova and Gascuel, 2006) was used to derive Maximum Likelihood (ML) trees. Four different evolutionary models were used for each seed sequence (JTT, WAG, Blosum62 and VT). In all cases, a discrete gamma-distribution model with four rate categories plus invariant positions was used, estimating the gamma parameter and the fraction of invariant positions from the data. The evolutionary model best fitting the data was determined by comparing the likelihood of the used models according to the AIC criterion (Akaike, 1973). Orthology and paralogy relationships among members of a family were inferred from the analysis of their corresponding phylogenetic trees, using a previously described algorithm that has been described before and has been shown to be accurate (Marcet-Houben and Gabaldón, 2009; Huerta-Cepas *et al.*, 2007). Phylogeny-based methods are considered to better reflect the actual complexity of orthology relationships than pair-wise methods such as best-bidirectional hits (Gabaldón, 2008). All phylogenetic trees are provided in figure S4 (see online supplementary material) as well as a list of proteins used as a seed in our analyses (see table S2 in online supplementary material). All duplications where manually checked to discard possible cases of spurious duplications. This was done by manually inspecting the alignments and the nucleotide sequences of the relevant duplicates. Moreover, the corresponding genome browsers or assembly data were searched to analyze the sequence context of the duplicates. Highly similar sequences in which one is only partially sequenced or in a small contig can be taken as possible source of errors. These cases were discarded. Total counts of duplications were also computed discarding the fraction of duplications that is expected to be more sensitive to error annotation: lineage-specific duplications with highly similar duplicates and duplications found in recently assembled genomes.

Results and Discussion

Phylogenomic profiling of the OXPHOS pathway

Sequences of fungal proteins annotated as OXPHOS components were retrieved from the KEGG database (Okuda *et al.*, 2008) and used as queries for blastp searches against the proteins encoded in 60 fully-sequenced fungal genomes (see table 3.1 and Methods section). A phylogenetic analysis was performed on each set of homologous proteins to derive a phylogenetic tree. This tree was used to establish orthology and paralogy relationships using a species-overlap algorithm that has been described earlier (Huerta-Cepas *et al.*, 2007). This phylogeny-based approach to orthology detection, approaches more closely the original definition of orthology and reflects more appropriately the complex evolutionary relationships within protein families (Fitch, 1970; Gabaldón, 2008). The presence of the different components of the respiratory pathway in the species surveyed is summarized in figures 6.1, 6.2 and 6.3.

Overall, our results agree with those reported by Lavin et. al in the 27 species that both surveys have in common (Lavín *et al.*, 2008). That the two approaches render so similar results, indicates that, despite using different approaches both methods have a similar stringency in the detection of OXPHOS components in this taxonomic range. In addition, our study extends the information on the distribution of components of the respiratory pathway to 33 additional species. The main advantage of our approach, however, is more qualitative than quantitative. By performing phylogenetic analyses on every protein family, we can readily obtain information on duplication events

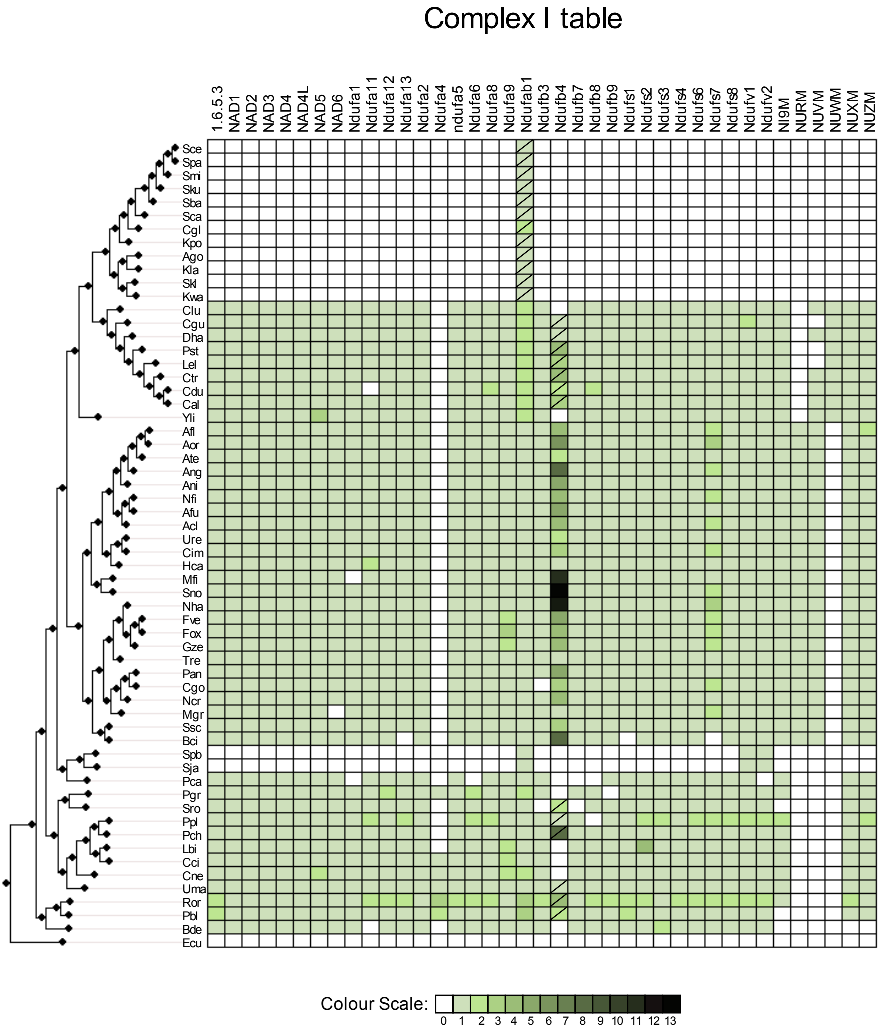

Figure 6.1: Phylogenetic distribution across 60 fungal species of Complex I subunits. Absences of a corresponding ortholog in a given species is indicated with a blank square or a crossed green square. Crossed green squares indicate that no ortholog was found but at least one paralog is present. Presence of orthologs is indicated with uncrossed green squares. The different colour intensities correspond to the number of homologs of the query protein found in that specific genome. The species are ordered according to their phylogenetic position in the fungal species tree Marcet-Houben and Gabaldón (2009).

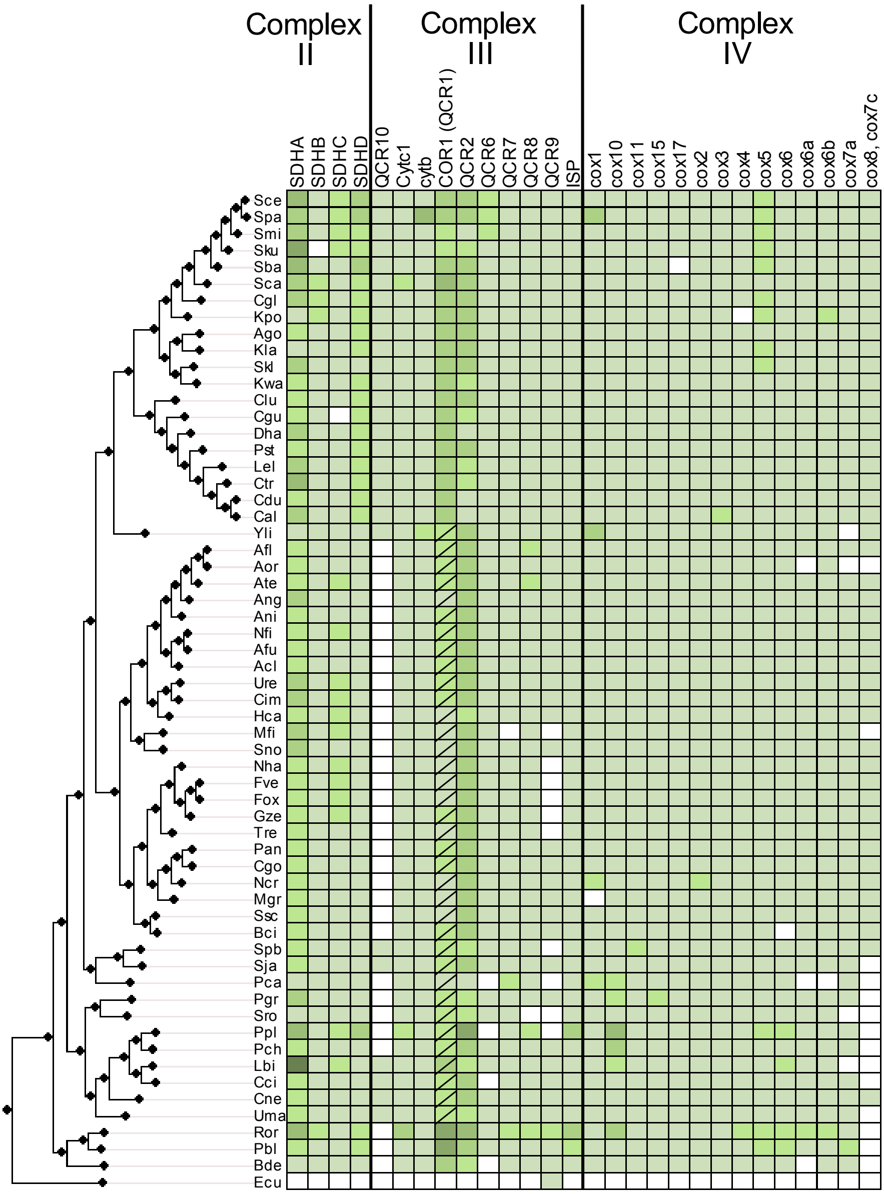

Figure 6.2: Phylogenetic distribution across 60 fungal species of subunits from Complexes II, III and IV subunits. Symbols and codes as in figure 6.1.

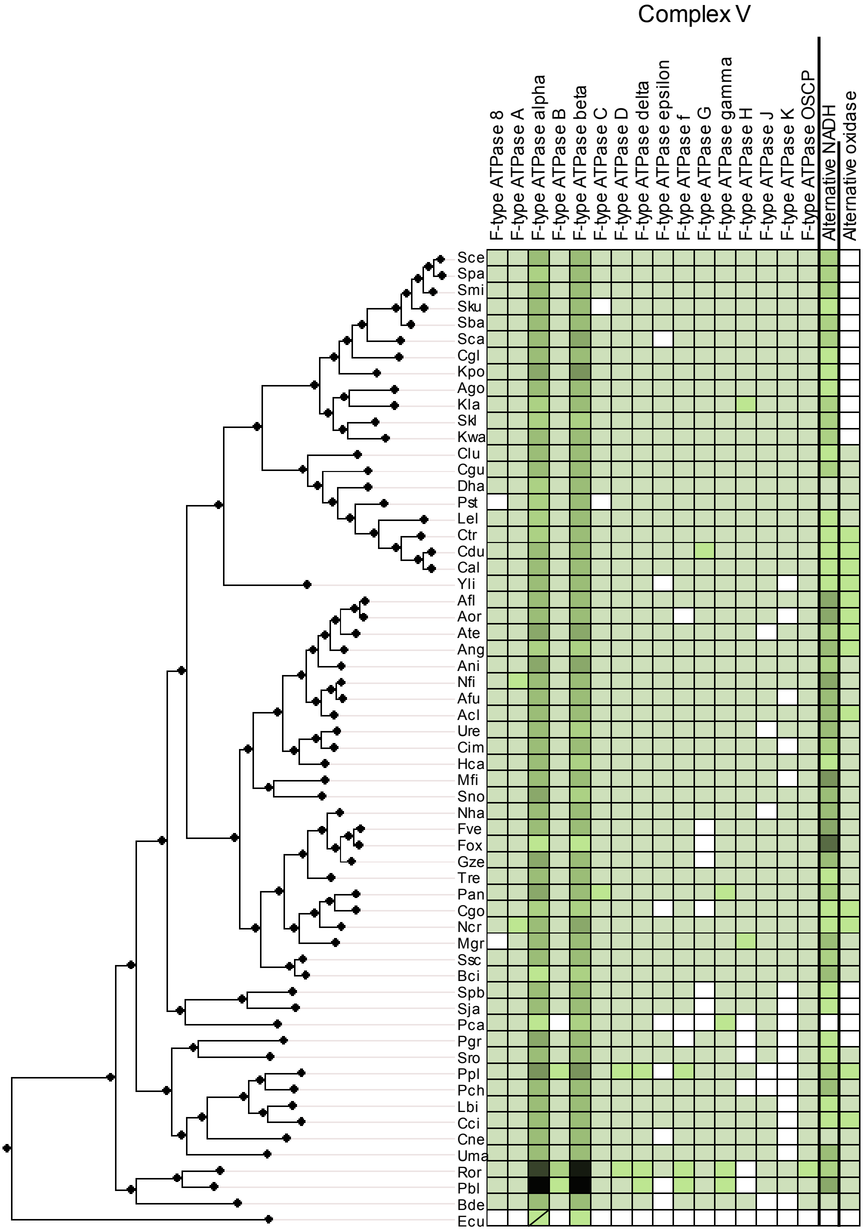

Figure 6.3: Phylogenetic distribution across 60 fungal species of Complex V and alternative oxidases and dehydrogenases. Symbols and codes as in figure 6.1.

affecting components of the respiratory pathway, an important evolutionary process that was ignored in the previous study. Recognizing gene duplications is important, since this process is considered one of the main processes that drive functional innovation (Ohno, 1970). Our study reveals that duplication events have affected the OXPHOS pathway extensively. Overall, we detect duplications in 76% of the families surveyed. These results were similar (duplications in 75% of the families surveyed), when more stringent cut-offs for homology detection were applied (see figures S1, S2 and S3 in the online supplementary material). Such high proportion of duplications is not the result of errors in the annotation or assembly of the genomes. We controlled for this by inspecting manually every duplication case to discard dubious cases. Moreover, even when species-specific duplications in which duplicates had more than 95% identity at the nucleotide level or all duplications from the recently assembled genomes *Postia placenta* and *Puccina graminis* were not taken into account, the fraction of OXPHOS families with a duplication event remained high (71% and 74%, respectively). Zygomycota, in particular, present the highest proportion of duplicated proteins in the OXPHOS pathway. For instance, we found duplications in 60% of the genes involved in *Rhizopus oryzae* OXPHOS pathway. A large percentage of these duplications (82%) can be mapped specifically to the R. oryzae lineage or to the lineage preceding the separation of *R. oryzae* and *Phycomyces blakesleeanus* and thus are specific of Zygomycota species. This large amount of lineage specific duplications seems to be general in *R. oryzae* and *P. blakesleeanus* (unpublished observation from our group). An interesting possibility is that the ancestors of these organisms underwent a Whole Genome Duplication (WGD) event, similar to that described for Saccharomyces (Kellis *et al.*, 2004). This possibility has recently been confirmed for *R. oryzae* (Ma *et al.*, 2009), in a comprehensive study that catalogues duplicated regions where the gene order is conserved. Consistently with our results, a duplication of nearly all subunits of the protein complexes associated with respiratory electron transport chains is detected, although our phylogeny-based approach detects additional, more ancestral, duplications that are not associated to the WGD event.

Complete loss of the OXPHOS pathway in microsporidia and two additional independent losses of Complex I coupled with alternative dehydrogenase expansions in Schizosaccharomyces and Saccharomycetales

Our results confirm earlier findings of a complete loss of the OXPHOS pathway in microsporidia (Katinka *et al.*, 2001) and the absence of most components of Complex I in Schizosaccharomyces and Saccharomycetales (Gabaldón, 2005). We are able to find most of the subunits of complex I in the Taphrinomycotina species Pneumocystis carinii, suggesting that the event of gene loss occurred after the diversification of Pneumocystis and Schizosaccharomyces lineages. The apparent multiple absences of Complex I subunits, and those of other complexes, in *P. carinii* is probably related to a low coverage of the genome sequence for this organism. Similarly, the presence of a complete repertoire of Complex I subunits in all species in the Candida cluster and the lack of this complex in all surveyed species from the Saccharomyces/Kluyveromyces clade, situates the loss of Complex I in the latter lineage. Remarkably, the two independent losses of Complex I in the Taphrinomycotina and Saccharomyces/Kluyveromyces clades are concomitant with independent expansions of their alternative NADH dehydrogenases repertoire by virtue of gene duplications. Alternative NADH dehydrogenases bypass Complex I electron transport, oxidizing NADH without pumping of protons. The duplication of alternative NADH dehydrogenases (Figure 6.4) might have provided a selective advantage for yeast species using predominantly fermentative metabolism, due to adaptation to anaerobic environments. Excess of NADH causes a problem under fermentative anaerobic growth, since it prevents further oxidation of substrates due to a lack of a sufficient NAD+ pool to accept electrons. Thus, the diversification of pathways to further oxidize NADH would have been beneficial in such conditions. The loss of Complex I in the same evolutionary periods might also be related to adaptation to fermentative growth. It is unclear which of the processes preceded the other or whether both processes were concomitant. A higher taxon sampling within the Saccharomycotina and Taphrinomycotina might help to solve this issue in the future. Also coupled with Complex I loss, and in line with adaptations to anaerobic environments in the abovementioned lineages, we observe the loss of alternative oxidases.

Duplications in Alternative Oxidases are not necessarily coupled with a pathogenic life-style

Alternative oxidases catalyze the cyanide-resistant alternative pathway of mitochondrial respiration in some fungi, plants and several protists. This pathway directly transfers electrons from the ubiquinone pool to oxygen, thereby bypassing complex III and cytochrome c oxidase (Veiga *et al.*, 2003). Alternative oxidases are common in yeasts but limited almost exclusively to non-fermentative and crabtree-negative yeasts. Alternative oxidases participate in energy production but also in antioxidant defense of cells. It has been shown that alternative oxidases represent an important factor for the survival of pathogenic fungi inside macrophages (Magnani *et al.*, 2007). Considering this, it could be postulated that the duplication of these enzymes might have played a role in the emergence of pathogenesis in several mammal fungal pathogens. In our survey we detect several copies of alternative oxidases in 13 species. Some of these duplications seem to have occurred quite recently in their respective lineages, such as the

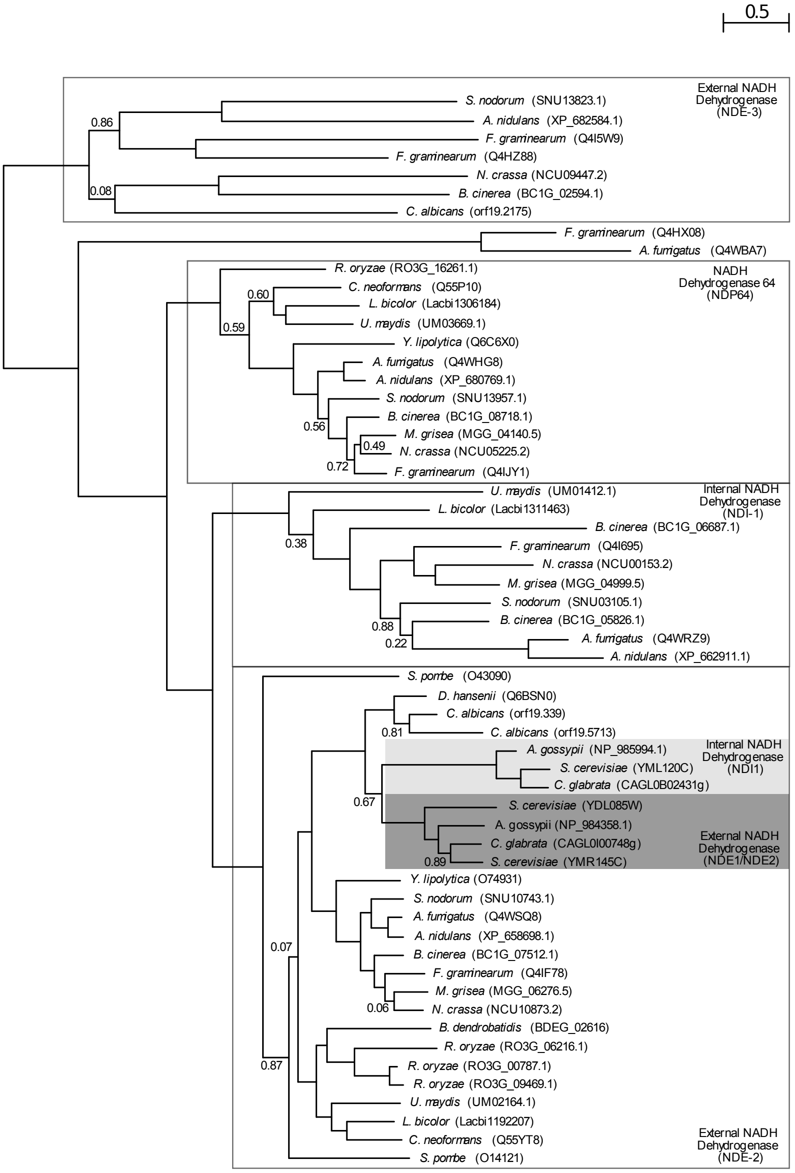

Figure 6.4: Phylogenetic tree representing the evolution of the alternative dehydrogenase protein family. The model used was WAG and approximate Likelihood (aLRT) support of the tree partitions is indicated if lower than 0.9. Duplications involving *S. cerevisiae* were marked with coloured boxes, while those involving *N. crassa* are indicated with white boxes. The species name is followed by the protein name according to the database from which the sequences where retrieved. Functional annotations were taken from Saccharomyces Genome Database (*S. cerevisiae*) (Christie *et al.*, 2004) and the Broad Institute (*N. crassa*). This tree represents a subset of the sequences used in the analysis, the tree with the full set of sequences can be accessed in the additional file 1.

duplication that lead to AOX1 and AOX2 (orf19.4774 and orf19.4773 in *C. albicans*) involving some Candida species, which can be mapped before the speciation of *C. tropicalis*, *C. dubliniensis* and *C. albicans*. Although many of these duplications do affect pathogenic genera such as Candida and Aspergillus, there are notable exceptions such as the intra-specific duplications found in the generally non-pathogenic species *Yarrowia lypolytica* or *Coprinus cinereus*. Conversely, we find pathogenic species such as *Histoplasma capsulatum* or *Cryptococcus neoformans* that have been shown to survive in macrophages (Johnson *et al.*, 2003) and nevertheless present a single alternative oxidase. Taken all together, our results suggest that a single copy of alternative oxidase gene is sufficient to protect fungal pathogens against macrophages and rather points to alternative selective advantages for the duplication of this gene. Conversely, alternative adaptations might be behind the emergence of the ability to survive inside macrophages in certain lineages. For instance, the presence of a polysaccharide capsule in Cryptococcus has been shown to confer resistance to oxidative stress (Zaragoza *et al.*, 2008).

Extensive duplication followed by functional divergence in the fungal OXPHOS pathway

According to the gene balance hypothesis (Papp *et al.*, 2003), the duplication of genes that encode for subunits of multi-protein complexes should have a higher chance of being deleterious due to dosage effects. As a result, one would expect to find few duplication events in the OXPHOS system, as this is mainly formed by intricate complexes. Contrary to that expectation, we find numerous cases of duplications in OXPHOS proteins, which overall affect 66 (76%) of the proteins surveyed. These have occurred at different moments in fungal evolution. At least for the genes duplicated during the Whole Genome Duplication event (WGD) occurred in the yeast lineage about 80 Myr ago, the rate of gene loss of duplicated OXPHOS genes is not higher than the overall rate for *Saccharomyces cerevisiae*. Indeed, our study finds six yeast proteins (all complex II subunits, Qcr6p and Cox5p), whose duplication is mapped to the WGD event. These represent 7% of the OXPHOS proteins, meaning that for 93% of the nuclear OXPHOS proteins supposedly duplicated in WGD were subsequently lost, a rate of gene loss that is roughly similar to the 88% estimated for the whole *S. cerevisiae* genome (Kellis *et al.*, 2004). It must be noted that one of the duplications affected the whole Complex II, meaning that the duplication was conservative in terms of stoichiometry of the different subunits. However, one of the duplicated subunits has been subsequently lost in the Saccharomyces sensu stricto species, suggesting the four duplicates do not form an alternative complex II. A possible reconciliation between the extensive rate of gene duplications and the gene balance hypothesis is that functional divergence directly followed the duplication event, thereby facilitating the retention of both duplicates (Lynch and Katju, 2004). Differences in the expression patterns of some of the WGD duplicates, point to a functional specialization of each duplicate. For instance, Cox5 (YNL052W) is expressed during aerobic growth whereas its paralog (YIL111W) is expressed under anaerobic growth (Hodge *et al.*, 1989). Similarly, the duplicate of the SDHA complex II subunit (YJL045W) is specifically expressed during the diauxic shift. Several other observations suggest that functional divergence processes have been common after duplication of OXPHOS protein families (see below).

Evolutionary cross-talk between the OXPHOS complexes and other multi-protein complexes

Several instances of paralogy relationships between complex I subunits and other mitochondrial multi-protein complexes have been previously reported (Gabaldón, 2005). This is the case for the NDUFA11 subunit, which is paralogous to the Tim17/22 family as well as that of NI8M (NDUFA2) and NUZM, which are paralogous to L43 and L2 subunits of the mitochondrial ribosome. It has been suggested that OXPHOS proteins with paralogs in other complexes would play a structural role rather than being involved in proton or electron transport, since ribosomes and the import machinery do not display those functions (Gabaldón, 2005). Similarly, we find several instances of paralogs of OXPHOS subunits that play a role in other complexes. Interestingly, another evolutionary connection between OXPHOS and the mitochondrial import machinery (MIM) is evidenced by the fact that the MIM subunit TIM18 (YOR297C) is a paralog of the Complex II subunit SDHD (YDR178W). Yet another paralog of the same Complex II subunit, which originated from a more recent duplication in the Saccharomycotina lineage (YLR164W) encodes for a mitochondrial inner membrane protein of yet unknown function. Paralogies to the protein import system in the mitochondrion extend to the two subunits of the Mitochondrial processing peptidase (MPP), an essential processing enzyme that cleaves the N-terminal targeting sequences from mitochondrially imported proteins (Gakh *et al.*, 2002). Indeed the large and small subunits MAS1 (YLR163C) and MAS2 (YHR024C) are homologous to QCR1 (YBL045C) and QCR2 (YPR191W) subunits of Complex III. Paralogy relationships to other multi-protein complexes extend beyond mitochondria. Indeed, several paralogs of Complex V subunits have been described as components of complexes from other cell compartments. For instance, the alpha and beta subunits of the F1 sector of the mitochondrial ATP synthase (YBL099W, YJR121W) are paralogous to the A and B subunits of the vacuolar ATP synthase (YDL185W, YBR127C). Vacuolar ATP synthases are found in the membranes of a large number of organelles which include endosomes, lysosomes and secretory vesicles. This duplication, however is not specific to fungi, since both paralogous groups have representatives in Arabidopsis thaliana and Homo sapiens (see phylogenetic trees additional file 1: figure S4), which indicates that the duplication preceded the diversification of plants and opisthokonts.

Yeast ACPM is possibly not a complex I remnant but a Saccharomycotina-specific paralog of complex I Acyl-carrier protein

Although previously identified as a remnant Complex I subunit in the Complex I devoid organism *S. cerevisiae* (Gabaldón, 2005), our current phylogenomic analysis suggest that this protein might actually be a paralog originated from an ancient duplication that is specifically conserved in Saccharomycotina (Figure 6.5). This paralogy relationship is supported by approximate Likelihood Ratio Tests (aLRT) analyses of the duplication node (0.79, shown in the figure) as well as by bootstrap analyses (74% bootstrap support, not shown in the figure). This finding clarifies apparent inconsistencies in the function of Complex I acyl-carrier protein and the isolated ACPM protein (YKL192C). Indeed *S. cerevisiae* ACPM protein has been found to participate in the synthesis of octanoic acid, a precursor of lipoic acid (Brody *et al.*, 1997), whereas the acyl carrier Complex I subunit in *Neurospora crassa* seems not to participate in this process (Schneider *et al.*, 1995). The fact that a mammalian homolog of this protein family has also been identified as a Complex I subunit in bovine mitochondria (Runswick *et al.*, 1991), suggests that association with Complex I is an ancestral feature of the family. Taken together, these results indicate that a process of functional divergence might have occurred after the duplication event diverting the new duplicate for specializing in the synthesis of octanoid acid. This specialization is presumably present in all Saccharomycotina species including the Candida group, which additionally possesses the true acyl carrier Complex I subunit.

Conclusions

Altogether our results shed light on how processes of gene loss, duplication and functional divergence have shaped the core of the respiratory pathway in fungi. Although most fungal organisms present a similar overall composition in terms of respiratory complexes, extensive differences in what particular units have been lost or duplicated in each complex, might help explaining differences found at the physiological level. This continuous evolution of OXPHOS components seems to be common in other groups of organisms (Gabaldón, 2005; Saccone *et al.*, 2006; De Grassi *et al.*, 2008), emphasizing the plasticity of this central energetic pathway.

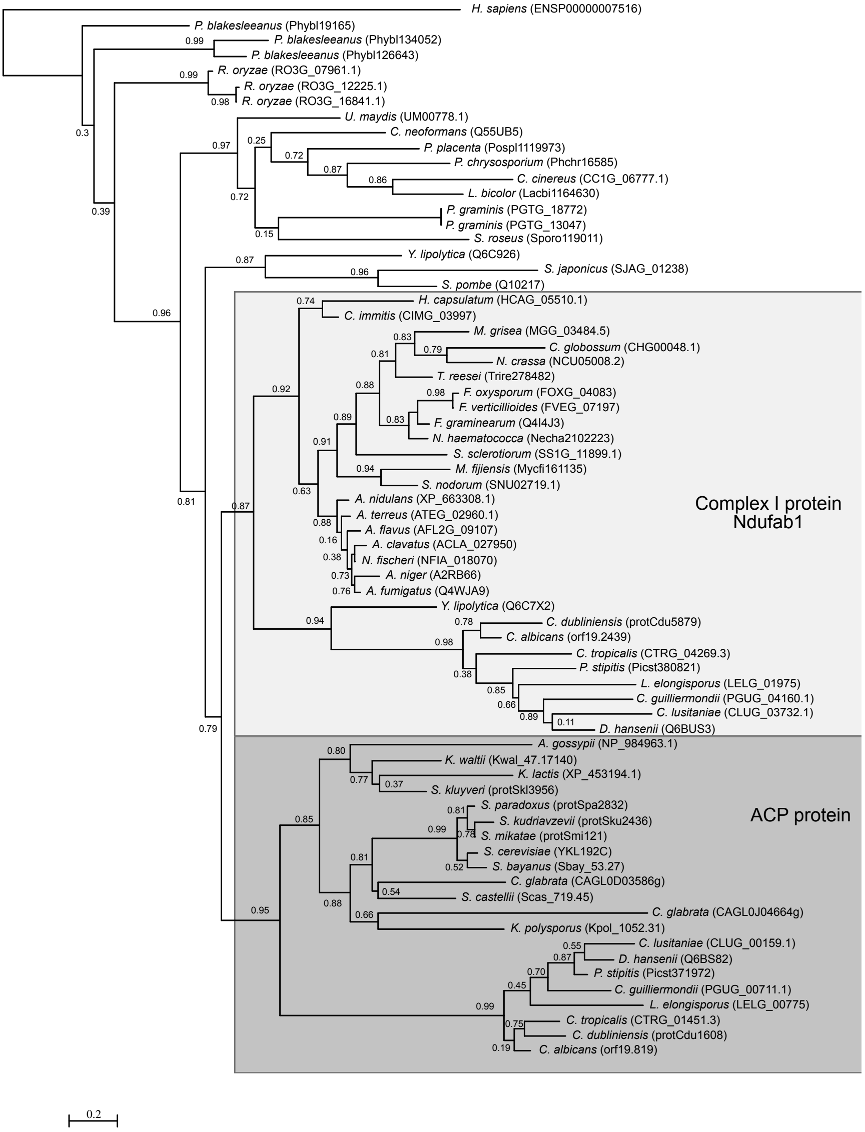
H. sapiens (ENSP00000007516)
P. blakesleeanus (Phybl19165)
0.99
P. blakesleeanus (Phybl134052)
P. blakesleeanus (Phybl126643)
0.99
R. oryzae (RO3G_07961.1)
R. oryzae (RO3G_12225.1)
0.98
R. oryzae (RO3G_16841.1)
U. maydis (UM00778.1)
0.3
C. neoformans (Q55UB5)
0.97
0.25
P. placenta (Pospl1119973)
0.72
P. chrysosporium (Phchr16585)
0.87
C. cinereus (CC1G_06777.1)
0.86
L. bicolor (Lacbi1164630)
0.72
0.39
P. graminis (PGTG_18772)
P. graminis (PGTG_13047)
0.15
S. roseus (Sporo119011)
Y. lipolytica (Q6C926)
0.87
S. japonicus (SJAG_01238)
0.96
S. pombe (Q10217)
0.96
H. capsulatum (HCAG_05510.1)
0.74
C. immitis (CIMG_03997)
M. grisea (MGG_03484.5)
0.83
C. globossum (CHG00048.1)
0.79
N. crassa (NCU05008.2)
0.81
T. reesei (Trire278482)
0.92
0.88
0.98
F. oxysporum (FOXG_04083)
F. verticillioides (FVEG_07197)
0.89
0.83
F. graminearum (Q4I4J3)
N. haematococca (Necha2102223)
0.91
S. sclerotiorum (SS1G_11899.1)
0.94
M. fijiensis (Mycfi161135)
S. nodorum (SNU02719.1)
0.63
A. nidulans (XP_663308.1)
0.81
0.87
Complex I protein
Ndufab1
A. terreus (ATEG_02960.1)
0.88
A. flavus (AFL2G_09107)
0.16
A. clavatus (ACLA_027950)
0.38
N. fischeri (NFIA_018070)
0.73
A. niger (A2RB66)
0.76
A. fumigatus (Q4WJA9)
Y. lipolytica (Q6C7X2)
0.94
0.78
C. dubliniensis (protCdu5879)
C. albicans (orf19.2439)
0.98
C. tropicalis (CTRG_04269.3)
0.38
P. stipitis (Picst380821)
0.85
L. elongisporus (LELG_01975)
0.66
C. guilliermondii (PGUG_04160.1)
0.89
C. lusitaniae (CLUG_03732.1)
0.11
D. hansenii (Q6BUS3)
0.79
A. gossypii (NP_984963.1)
0.80
K. waltii (Kwal_47.17140)
K. lactis (XP_453194.1)
0.77
0.37
S. kluyveri (protSkl3956)
S. paradoxus (protSpa2832)
0.81
S. kudriavzevii (protSku2436)
0.85
ACP protein
0.99
0.78
S. mikatae (protSmi121)
S. cerevisiae (YKL192C)
0.81
0.52
S. bayanus (Sbay_53.27)
C. glabrata (CAGL0D03586g)
0.54
S. castellii (Scas_719.45)
0.88
0.95
C. glabrata (CAGL0J04664g)
0.66
K. polysporus (Kpol_1052.31)
0.55
C. lusitaniae (CLUG_00159.1)
0.87
D. hansenii (Q6BS82)
0.70
P. stipitis (Picst371972)
0.45
C. guilliermondii (PGUG_00711.1)
L. elongisporus (LELG_00775)
0.99
C. tropicalis (CTRG_01451.3)
0.75
C. dubliniensis (protCdu1608)
0.19
C. albicans (orf19.819)
0.2

Figure 6.5: Phylogenetic tree representing the evolution of the ACPM protein family. The model used was WAG and approximate Likelihood (aLRT) support of the tree partitions is indicated. This tree represents a subset of the sequences used in the analysis, the tree with the full set of sequences can be accessed in the additional file 1.

Additional files can be found here:

http://www.biomedcentral.com/1471-2148/9/295/additional/

CHAPTER 7
FUNPATH CONSORTIUM

Introduction

Infections caused by fungal species consitute a growing medical problem. The appearance of AIDS, the evolution of immunosuppressive treatments used for organ transplants, improvements in health care, which have led to increased survival rate in critically ill patients, are all factors that have caused an increase in the number of reported cases of fungal infections.

One of the major fungal groups with the ability to infect humans are Candida species (such as *Candida albicans*, *Candida glabrata*, *Candida parapsilosis* and *Candida tropicalis*). Altogether, they represent the fourth leading cause of hospital-acquired diseases. The overall mortality of systemic Candida infection can be as high as 42%, exceeding the mortality of all gram-negative bacterial septicemia (Horn *et al.*, 2009; Méan *et al.*, 2008; Nace *et al.*, 2009; Pfaller and Diekema, 2007; Wenzel and Gennings, 2005). This high mortality rate is due, in part, to a lack of accurate and fast diagnostic tests, but also because current antifungal therapies are not always effective enough.

The major causative agent of candidiasis is *C. albicans*, which conforms ~54% of the isolated species, followed by *C. glabrata* and *C. parapsilosis* each with ~14% and *C. tropicalis* with 12% of all Candida isolates (Pfaller *et al.*, 2008). During the last 20 years, the distribution of the different Candida species found infecting humans has been changing gradually and varies across geographic regions. While *C. albicans* is the most prevalent species retrieved from clinical isolates, with a frequency of occurrence which varies throughout the world from of 37% in Latin America to 70% in Norway, *C. glabrata* is the second most frequent cause primarily in North America and the European Union. In contrast, *C. parapsilosis* and *C. tropicalis* are more prevalent in Asian and Pacific regions and account for almost 20% of all Candida infections in Latin America. The prevalence of *Candida krusei* is generally low but more common in Europe than in other regions (Pfaller *et al.*, 2008). The appearance of other Candida species remains low but the occurrence of these species is under constant observation (Shin *et al.*, 2007; Magill *et al.*, 2006; Panackal *et al.*, 2006).

The mechanisms that *C. glabrata* uses to infect hosts are largely different

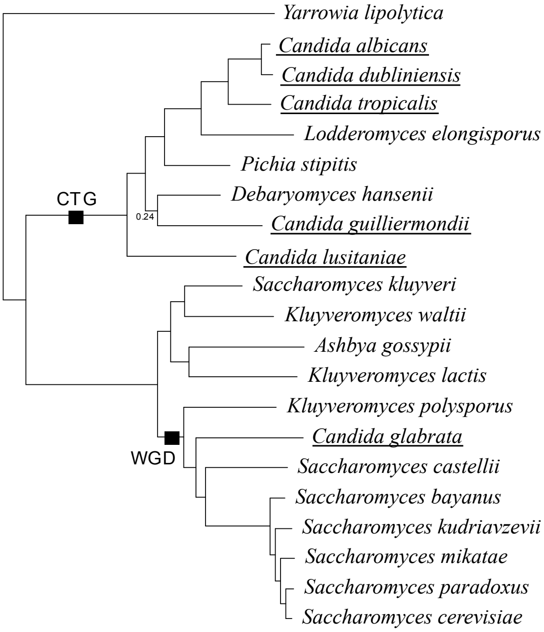

Figure 7.1: Phylogenetic tree depicting the evolutionary relationship of a group of completely sequenced species belonging to the Saccharomycotina subphylum (adapted from figure T21 in chapter 3). All nodes except one (indicated) have a high support (>0.99) based on an approximate likelihood ratio test analysis (aLRT). Underlined species names indicate human fungal pathogens. Squares mark the moment in which the CTG codon was reassigned to codify for a serine instead of leucine and the point in which the whole genome duplication occurred. According to this phylogenetic scenario, *C. glabrata* and other Candida species became human pathogens in independent evolutionary events.

from the known mechanisms in *C. albicans*. Virulence factors such as filamentation and secretion of large amounts of proteases are thought to be mostly absent in *C. glabrata*. This, coupled with an enhanced resistance to azole drugs leads to severe difficulties in treating infections caused by this species.

From an evolutionary standpoint, *C. glabrata* is more closely related to *Saccharomyces cerevisiae* than to other Candida species (see figure 7.1). This fact suggests that this species developed its pathogenesis independently from the other Candida species, which in turn explains why their mechanisms of infection differ. The assembly of the *C. glabrata* genome sequence (Dujon *et al.*, 2004) now provides the possibility to study the distinct pathogenicity of *C. glabrata* using genomic approaches (Schwarzmuller *et al.*, 2010). Additionally, its closeness to the well characterized, non-pathogenic budding yeast provides a good scenario for the use of comparative genomics.

The FunPath consortium

The FunPath consortium started in 2007 and has been funded through the ERA-NET Pathogenomics program of the European commission. The main objective of the consortium is to identify virulence factors used by *C. glabrata* to infect human hosts by using throughput genetic analysis. The consortium is formed by several experimental groups including the laboratories of Karl Kuchler and Cristoph Schüller in Austria, Christophe d'Enfert and Dominique Ferrandon from France and Steffen Rupp and Bernhard Hube from Germany. Our lab is the only bioinformatics component in the consortium and as such our complementary expertise has served to make predictions of potentially interesting genes, and to offer analysis support to the members of the consortium.

The project has been divided in three parts. The first part comprised the identification of genes that were likely to be involved in virulence. These genes have been the starting point for the construction of a deletion library that will one day contain deletion mutants for all genes of *C. glabrata*. The last part comprises the phenotypic characterization of the deletion mutants.

A total of 868 candidate genes were selected for a first round of knockout experiments. Selected genes included genes with homologs in *S. cerevisiae* that were known to be involved in cell wall synthesis, signaling cascades and transcriptional regulation. Candidates also included genes with no orthologous sequences in *S. cerevisiae*, as they could confer phenotypic differences, including the ability to infect humans. At the moment a total of 476 strains have been constructed (work done by Tobias Schwarzmüller and Karl Kuchler). The strain collection is now being tested for phenotypes under different stress conditions which try to emulate conditions that *C. glabrata* may find when infecting the host. They also are being used to infect animal models. Within the consortium, Dominique Ferrandon's lab uses fly models to test the virulence of mutants (Roetzer *et al.*, 2008) while the model of choice in Bernhard Hube's lab is the mouse (Jacobsen *et al.*, 2010). While the funded period of FunPath project has now reached its end the collaborations between the groups are still going on with the objective of obtaining the complete collection of fully characterized *C. glabrata* knockouts one day. All the data pertaining to the knockout collection has been stored in the newly created FunPath database (http://funpath.cdl.univie.ac.at/). The web page was constructed in a collaboration between our lab and Walter Glaser and Karl Kuchler, and is expected to host all the results derived from the FunPath consortium. At the moment, information about the knockout library can be found in a password protected section of the web page.

The *Candida glabrata* phylome: detection of potential virulence factors based on their evolutionary history.

S. cerevisiae is probably the fungal species with the most comprehensive genome annotation. Numerous groups have worked during more than a decade towards completely understanding the function of the genes in this species. The close relationship shared by *C. glabrata* and *S. cerevisiae* will enable us to safely transfer annotations from the baker's yeast to this pathogenic species. In order to do that an accurate orthology and paralogy prediction between these two species needs to be established. There are numerous ways to do this. While methods based on similarity such as best blast hit, best bidirectional hit, or inparanoid (Ostlund *et al.*, 2010) are often used, phylogenetic methods tend to be better suited, especially when multiple to one or multiple to multiple gene relationships need to be established (Gabaldón, 2008). The construction of a phylome (Sicheritz-Pontén and Andersson, 2001), which comprises the complete collection of gene phylogenies in a genome, enables the establishment of phylogenetic based orthology predictions at a genome-wide scale.

We reconstructed the *C. glabrata* phylome using the same dataset and in the same way as the yeast phylome (see chapter 3), but using *C. glabrata* as the seed proteome. All trees and alignments from the phylome are stored in phylomeDB (Huerta-Cepas *et al.*, 2008) and will be made publicly accessible soon. Orthology and paralogy relationships were inferred based on a species overlap algorithm (Huerta-Cepas *et al.*, 2007) using ETE (Huerta-Cepas *et al.*, 2010b). Annotations from *S. cerevisiae* were transferred to their orthologs in *C. glabrata* and were used to select the genes to be deleted. The availability of all evolutionary histories of *C. glabrata* genes allowed us to search for genes with interesting topologies. For instance, we searched for those phylogenies that contained *C. glabrata* genes with no homologs in *S. cerevisiae* or for those genes that had been specifically duplicated in *C. glabrata*.

Genes without orthologs in *Saccharomyces cerevisiae*

Differences in gene content between the pathogenic *C. glabrata* and the non-pathogenic *Saccharomyces cerevisiae* are excellent candidates of being virulence factors. Without taking into account those genes that have no tree, we found 699 genes that had no orthologs in the baker's yeast. Moreover, If we have a further look at how many genes do not have any homologous protein in yeast, we have 87 proteins. Of these 87 genes, only 4 do not have homologs in any other post whole genome duplication (post-WGD) species nor in the Kluyveromyces clade. Of these four genes, one was predicted as horizontally transferred from prokaryotes (see below and chapter 5). The remaining three have homologs in some other fungi, but they are not widespread, which may point to further horizontal gene transfer events among fungi.

Expansions of protein families in *Candida glabrata*

We looked for proteins that had specific duplications in *C. glabrata* but not in any other closely related species and identified 7 such families. One of these families contained a pfam domain related to adhesins. Efficient adherence to host tissues is often one of the key factors in pathogenic infection. In Candida species this adhesion is mediated by cell wall-located adhesins. So far in *C. glabrata* a large group of GPI-anchored cell wall proteins is predicted and many of them represent potential adhesins. The epithelial adhesin (EPA) gene family represents the largest group in *C. glabrata*. In vitro, epa1 mediates *C. glabrata's* adherence to epithelial tissue while other EPA genes are expressed at low levels. Even so the relation of Epa1 to virulence remains unclear because there are no differences found between colonization of wild-type and knockout strains. Besides the EPA genes, *C. glabrata* contains several other proteins that represent potential adhesins (e.g. Awp, Pwp families). The presence of these proteins in the cell wall has been shown to depend on the strains' genetic background and growth phase, and are thought to be controlled by external stimuli. The identification of new families of adhesins, that are specifically expanded in *C. glabrata*, can lead us to discovering new virulence factors, therefore they constitute perfect candidates for the construction of mutants to be subjected to further phenotypic studies.

Horizontal gene transfer in *Candida glabrata*

During the analysis performed in chapter 5, we found one horizontal gene transfer (HGT) event that involved *C. glabrata*. The predicted protein was identified as a putative Aspartate racemase, which is able to interconvert L-aspartic to D-Aspartic. *Candida albicans* and other Candida species contain amino acid oxidase genes that are used to detoxify D-amino acids, but they are absent in *C. glabrata* and *S. cerevisiae*. On the other hand, D-Aspartate can be found in the nervous and endocrine tissues in mammals (Schell *et al.*, 1997). As we know that *C. glabrata* is able to infect the brain (Srikantha *et al.*, 2008) and lacks the most common detoxification method for D-Aspartate, we hypothesized that *C. glabrata* needs an alternative method to deal with D-Aspartate and therefore the transference of an Aspartate racemase would have been beneficial. This hypothesis is currently under experimental testing in the labs of Bernhard Hube and Matthias Brock.

Prediction of transcription factor binding sites

Humans have numerous defensive mechanisms against fungal infection. Phagocytes play an essential role in this defense and successful pathogens have had to develop mechanisms to avoid their destruction by macrophages, neutrophils and others. One of the mechanisms of action of phagocytes, is the generation of reactive oxygen species which cause a situation of stress upon the pathogen. When cells are under oxidative stress, transcriptional remodeling occurs to ensure the proper response. In *S. cerevisiae* the response to oxidative stress is in part under the control of the well-studied transcription factors Yap1p, Skn7p, Msn2p, and Msn4p (Cuéllar-Cruz *et al.*, 2008). The lab of Christoph Schuller investigated the effect that deletion of Yap1, Skn7, and a double mutant of both, had on the transcriptome of *C. glabrata* under oxidative stress (unpublished results). Genes with changes in regulation under these conditions were divided into three groups depending on whether they were dependent on one of the two transcription factors (group 1), only on Yap1 (group3) or the case in which expression is repressed when either transcription factor is deleted, but the double mutant causes an overexpression, which lead us to believe in the existence of a third transcription factor (group2). We collaborated in this analysis by searching for transcription factor binding sites (TFBS) enriched in the promotors of each group of responsive genes.

TFBS search for Yap1 and Skn7.

For each gene in the abovementioned groups, we analyzed 2000 nucleotides upstream of the start codon. We retrieved TFBS motifs that can be found in yeastract (Teixeira *et al.*, 2006) for both Yap1 and Skn7. Then we used a pattern matching approach in order to identify sites that completely matched the described TFBS. Since a certain amount of variation can be expected, we repeated the search allowing for one change in any given position. The results correlate with the observed variations in expressions. Differences can be attributed to the fact that near-canonical Skn7 sites appear almost everywhere. We suspect that most of these variations will render the site non-functional, although we currently lack appropriate weight matrices to assess this. Moreover, the functionality of binding sites may also depend on other factors such as the genomic context or the distance to the start-site.

De novo predictions of TFBS for the identification of a third transcription factor

As explained above, Group 2 comprises genes that are repressed respect to the wild type when either Yap1 or Skn7 are deleted but in which the double mutant is once again overexpressed. A possible explanation for the observed expression pattern is that a third, unidentified transcription factor is involved in the regulation of these genes. We tried to predict patterns of nucleotides that were over-represented in the promotor regions of these groups of genes. There are numerous strategies to predict TFBS de novo (Tompa *et al.*, 2005). They can be divided in two different groups: algorithms based on word counting and methods based on probabilistic sequence models. We decided to combine three different programs: Olygo-analysis (van Helden *et al.*, 1998), BioProspector (Liu *et al.*, 2001) and MotifSampler (Thijs *et al.*, 2001, 2002). Olygo-analysis belongs to the first group and searches for over-represented oligonucleotides within a group of upstream regions. On the other hand, both BioProspector and MotifSampler use Gibbs sampling to predict statistical representative motifs.

Each program predicted several conserved TFBS in proteins from the second group with varying lengths (6-10 nucleotides). However, our results showed large inconsistencies among predictions for the various programs without a clear motif being consistently present in the promotor regions of group 2 and absent in the other groups.

Considering these results we changed the approach and tried to manually predict the motifs that were in group 2 but not in the other two groups. We devised a perl script that created all possible motifs of a certain length (6, 7 and 8) and then searched for their presence in the genes of each group. For this, we found that the results for motifs of 6 sites were not specific enough, but there were several motifs of 7 and 8 sites that were present in approximately one third of the promotors of the genes in group 2 and in none of the other groups. We then introduce single changes in those motifs in order to enlarge their scope in group 2 while still not allowing them to be present in the other two groups. The best result covers 60% of the gene promotors in group 2, but the motif is exclusively found in this group. In addition, if we take the motifs together, we are able to cover all the proteins in group 2. A graphical representation of the location of the motifs can be found in Figure 7.2.

Given the short length and the difficulty of assessing the statistical significance of a hit, the prediction of TFBS is still unreliable and often produces numerous conflicting results. Much work still needs to be done in order to improve and unify the predictions of such short stretches of DNA. In this kind of analysis the input of experimentalists is very valuable as they will be able to asses and direct the predictions so that the algorithms can be improved. Even so, computational predictions may be used to provide hints on where to start the experiments.

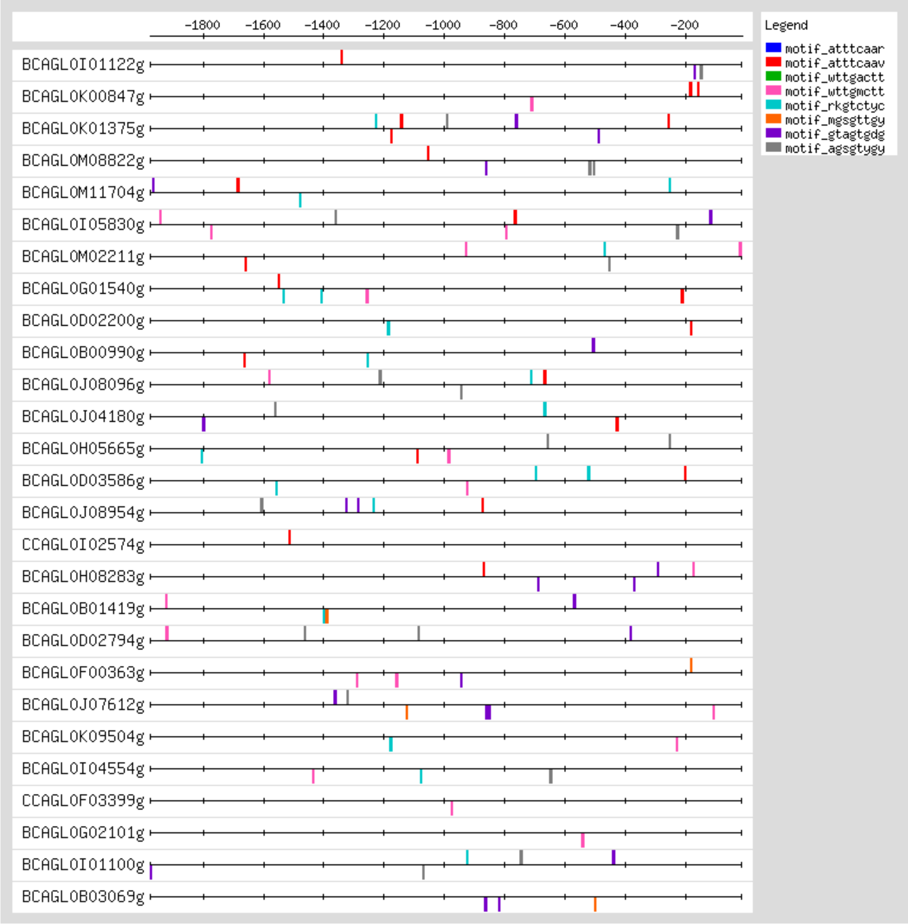

Figure 7.2: Prediction of TFBS in *Candida glabrata* proteins that are affected by the deletion of either Yap1 or Skn7, but are overexpressed in the double mutant. Lines in front of each gene represent the 2000 nucleotides upstream of each gene and the colored stripes represent the location in which each TFBS was predicted.

Tree reconstruction based on SNPs in *Candida albicans* strains

Within the FunPath consortium we initiated a collaboration with Christophe d'Enfert's lab. *C. albicans* is one of the main human fungal pathogens. More than 54% of the detected cases of candidiasis are caused by this fungi. Multi-locus sequence typing (MLST) measures the DNA sequence variations in a set of genes and is then used to characterize strains by their unique allelic profiles. This technique has been successfully applied in *C. albicans* isolates in order to identify different clades (Odds *et al.*, 2007; Tavanti *et al.*, 2005; Bougnoux *et al.*, 2004). Odds et al. (Odds *et al.*, 2007) defined 17 different clades from more than 1000 different isolates. Geographic location and ITS type (based on the presence or absence of an intron in DNA sequences encoding rRNA) were found to be the main properties that differentiated between clades. Different strains of *C. albicans* belonging to different clades were sequenced by Christophe d'Enfert's group. After assembling the genomes using SHORE (http://www.1001genomes.org/downloads/shore.html), single nucleotide polymorphisms (SNPs) were identified.

One of the objectives of the collaboration was to use these SNPs to reconstruct a phylogenetic tree and see whether the distribution of clades matched the results obtained by MLST data. The first tests were run with 6 *C. albicans* strains, three of them, including a re-sequenced reference strain, belonged to clade 1 while the other three belonged to clade 11. At later stages more strains were included to finally reach the point where 8 clades have at least one representative. In order to reconstruct the phylogenies we applied two different methodologies: parsimony and maximum likelihood.

Parsimonious tree reconstruction:

The idea was adapted from Ruderfer et al. (Ruderfer *et al.*, 2006). In their paper Ruderfer et al. used SNP predictions for three strains of *S. cerevisiae* and one of *Saccharomyces paradoxus*. Using *S. paradoxus* as an outgroup they were able to identify informative SNPs, in all such cases two baker's yeast strains contained the same SNP while the third strain shared the same nucleotide with *S. paradoxus*. Using this information they were able to determine the most likely evolution of the three *S. cerevisiae strains* as it would match the topology with the largest number of informative SNPs.

We performed a similar approach on the first six strains. For each SNP position we grouped the species by their allele. The groups of species that had more alleles in common were then used to reconstruct the phylogenetic tree. Groups of species that contradicted groups that had already been selected for the tree were discarded. The final tree (see figure 7.3) correctly grouped the

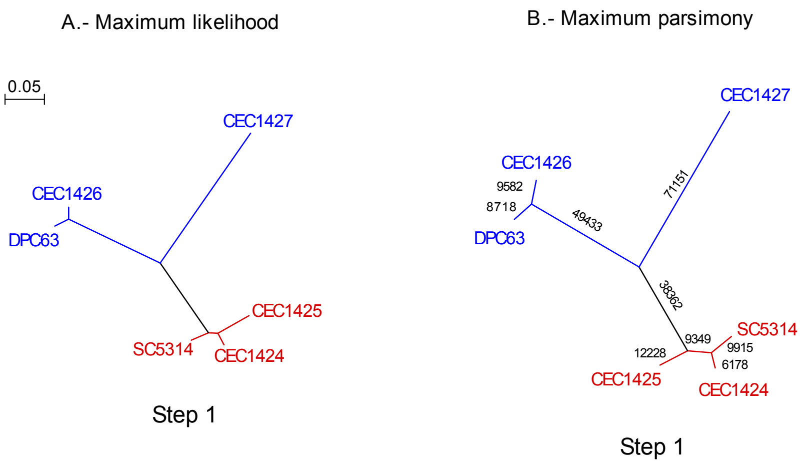

Figure 7.3: Phylogenetic trees based on SNP information of six *C. albicans* clades: 169,674 SNPs were used to reconstruct these two phylogenetic trees. Red leaves represent members of clade 1 while blue leaves represent members of clade 11. Numbers on the branches of the tree built using parsimony represent the number of SNPs that support each partition.

three clade 1 strains and the three clade 11 strains. It also clearly showed the difference between one of the clade 11 strains (CEC1427), which had already been identified as having a different genetic background than other clade 11 strains.

As most parsimony methods, this methodology can only be applied to reconstructions with a limited number of strains. When more strains are included, the number of possible groups grows and the number of SNPs that supports each group decreases, rendering this methodology unfeasible.

Tree reconstruction based on SNP concatenation

All positions in the six *C. albicans* strains that contained at least one SNPs were concatenated into a single multiple sequence alignment of 169,674 sites. Each SNP was represented by two positions in the alignment so as to take into account homozygous and heterozygous SNPs. The phylogenetic tree was reconstructed using GTR+I+gamma as implemented in PhyML3 (Guindon *et al.*, 2009). As seen in figure 7.3, the two trees presented almost the same topology. Only the inner distribution of the strains included in clade 1 (leaves in red) varied slightly between methods. This methodology was also applied to reconstruct larger trees, which included 20 newly sequenced strains belonging to 8 different clades. When compared to the MLST data it was shown that the relationship between the strains conformed to the clade grouping identified previously.

The methodology we applied during the collaboration with Christophe d'Enfert will be useful in our further work as we are now involved in a project to sequence eight different *C. glabrata* strains and we plan to extend similar analyzes to this group.

Final Remarks

Most virulence mechanisms that *C. glabrata* uses to infect human hosts are still unknown. The work of the FunPath consortium is only a first step in the quest to eradicate the threat that this fungal pathogen poses to human health. The creation of a deletion library has just started and will continue over the next years, with the expectation that it will constitute an important source of information regarding the investigation of the genes that influence *C. glabrata's* ability to infect humans. Bioinformatics has played an important role in the selection of the genes that should be deleted first and in analyzing the output data. In the future it will continue playing and important role when the time comes to analyse all the data related to high-throughput analysis.

Part III

Summarizing discussion

CHAPTER 8
SUMMARIZING DISCUSSION

This thesis has explored different aspects of the evolution of the fungal kingdom by using phylogenomics tools. In this chapter we will summarize the most important conclusions we have drawn from the analyzes performed and will provide a summarizing discussion.

The fungal species tree: node support versus tree distance

Often, basic knowledge on the species phylogeny is needed in order to provide an evolutionary framework to phylogenomic studies. This basic knowledge is not always readily available. Many parts of the species tree are still not well resolved and are the source of heated debate. While numerous fungal species trees have been reconstructed over the years (Marcet-Houben and Gabaldón, 2009; Fitzpatrick *et al.*, 2006; Wang *et al.*, 2009; James *et al.*, 2006; Lutzoni *et al.*, 2004), it is evident by their slight differences that there are some nodes that are not easily resolved (Marcet-Houben and Gabaldón, 2009). Species trees are used as an evolutionary framework in many studies and knowledge on their reliability is important when we use it as a starting point for further work. In concatenated alignments, the method of choice in most studies, including ours, it has been shown that measures usually applied to assess the robustness of tree nodes (i.e. bootstraps support (Felsenstein, 1985)) can be influenced by the length of the concatenated alignment (Gadagkar *et al.*, 2005), therefore they may fail to offer objective information about the robustness of tree nodes.

In addition, evolutionary events such as gene duplication with differential loss, horizontal gene transfer or lineage sorting, can result in differences in the topology of gene trees. Considering this, how can we be sure that the genes we are using to reconstruct the species tree are representative of the evolution of the different species and are not giving us information biased by the particular sample of genes used?

Ideally, in order to build a species tree, one would like to use all the gene families compressed in the organisms included in the tree, but this is still difficult with current methods. An alternative approximation is to identify problematic nodes independently from the genes used for the reconstruction. In chapters 3 and 4 we have addressed these issues by using a genome-wide collection of gene trees (a phylome) in order to assess the robustness and correctness of a species tree. In chapter 3 we used an algorithm that was able to provide support to each of the tree nodes based on the fraction of trees in the phylome that showed a congruent topology for that group of species. This comparison motivated our search for other distance measures that could go beyond the assessment of the existence of a given topology distribution. The lack of appropriate methods to assess the distance between two trees that contain duplications lead us to the development of treeKO (http://treeKO.cgenomics.org). In chapter 4, we describe this tree comparison program and use it to discriminate between different species tree topologies by comparing the distance distributions to a phylome.

Both measures are complementary and while treeKO provides a global vision of the similarity between each gene tree and the species tree, the algorithm presented in chapter 3 focuses more on the prevalence of specific sub-topologies in the phylome. In either case, the results of both methods show that there are some nodes in the species tree that we are not able to resolve satisfactorily, even when the information present in a whole phylome is used. This is true for those nodes in which gene trees largely disagree, providing similar support to various alternative topologies.

These results beg the question of how this issue can be solved. In chapter three we show that if a given node was dubious in a large tree, using less species centered only around the problematic region to sample more genes for the concatenated alignment did not solve the problem. It is our hope that an increase in the number of species at key points of the tree may one day provide enough information to unequivocally describe the fungal species tree. Increasing the species coverage at certain nodes is becoming more feasible now thanks to the advances in sequencing technologies. Alternatively, the possibility exists that standard sequence analysis cannot resolve certain topologies in which processes such as lineage sorting, hybridization or horizontal gene transfers have blurred the signal to the point in which not even using large amounts of genomic sequences allows the retrieval of a robust species tree. In such instances alternative data will need to be provided that is not directly related to sequences. The conservation of gene order for instance has been used in several instances to support specific species phylogenies. This is the case for the conflicting branching order of *Candida glabrata* and *Saccharomyces castellii*. Phylogenetic trees indicate almost unanimously that *C. glabrata* diverged before *S. castellii*, but the node has been identified as one of the poorly supported nodes (see chapter 3 and chapter 4). Gene order analyzes have provided the counter argument by showing that a basal placement of *S. castellii* is more parsimonious regarding the number of genomic rearrangements implied (Gordon *et al.*, 2009).

Implications of tree uncertainty for orthology prediction

As explained above, assessing unequivocally a species tree topology is not a trivial problem. Along the same lines, ensuring that we have a correct gene tree poses similar problems. Beyond the topological differences caused by evolutionary events, we showed in chapter 3 that tree topologies can also vary depending on the phylogenetic reconstruction method or the evolutionary model used during the reconstruction. While these changes may not affect the whole gene tree, they are still introducing uncertainty in the data. Phylogeny-based orthology predictions are more appropriate than predictions obtained by other methods (Gabaldón, 2008) as they are able to elucidate complex relationships between genes that may have undergone numerous gene duplication or loss events. Uncertainty in the data may greatly affect the prediction of orthologs and paralogs. Tree reconciliation methods (Page and Charleston, 1997) have often been used to establish evolutionary relationships between sequences, but their most basic premise is that both, the gene tree and the species tree have to be correct. This assumption is often not met and therefore the number of duplications may be over-estimated, which leads to erroneous predictions. Throughout this thesis we have used a more flexible orthology prediction algorithm. The species overlap algorithm (described in (Huerta-Cepas *et al.*, 2007)) is able to establish orthology and paralogy events based only on the information found in the gene tree, reducing the assumptions on a species phylogeny simply to that necessary to root the tree. This last requirement, however, can be avoided when alternative rooting methods such as midpoint rooting are applied. We show in chapter 3 that while orthology predictions of both methods had a similar predictive value, the species overlap algorithm had a higher sensitivity. This can be directly correlated to the flexibility shown by the algorithm, which will not be greatly affected by small changes in the gene tree topology and is completely independent from the species tree.

Horizontal gene transfer events in fungi are more abundant than previously thought

In chapter 5 we used an automatic pipeline to estimate the number of interdomain putative horizontal gene transfer events (HGT) between prokaryotes and fungi. Few events had been reported before, and most analyzes were centered around Saccharomycotina species. Previous work showed that Saccharomycotina species seem to be particularly reluctant to incorporate foreign DNA even though it has been experimentally demonstrated in vitro in *Saccharomyces cerevisiae*. Our results showed that while HGT events were scarce in Saccharomycotina, this was not the case in Pezizomycotina. While not reaching the high percentages described in bacterial species, we found that several Pezizomycotina species could have up to 0,38% of their genes transferred from bacteria, using a strict criteria. Species with the highest rates of predicted HGT events are Aspergillus and Fusarium species, which are well represented in terms of completely sequenced species. Therefore, the results can not be attributed to a poor sampling of close species since they are some of the best sampled groups in fungi. There is not a clear reason as to why this bias exists. One of the main differences between Pezizomycotina and Saccharomycotina is the size of their genomes. It has been hypothesized that the size of prokaryotic species plays an important role in the reasons as to why HGT is important in their evolution (Isambert and Stein, 2009), as they are constrained in terms of evolution by duplication-divergence. A similar reasoning, however, cannot be used for fungi. Despite the small size of the genomes of Saccharomycotina species, numerous duplication events have been found within their genomes, the largest of them spanning the whole genome duplication that occurred in the ancestor of *S. cerevisiae* and *C. glabrata*. Therefore it does not seem likely that fungi would need horizontal gene transfer events to be the main force behind their evolution. On the other hand, without this need, the small size and compact organization of the Saccharomycotina genomes, with fewer non-coding regions, my hamper the uptake of foreign DNA as any change in the conformation of their chromosomes may destabilize the whole organism. Pezizomycotina, which possess much larger genomes, would be more resistant to this effect, facilitating the existence of transfer events.

While in this thesis we have not treated transferences between fungal species, recent studies have shown that this event may be frequent. For instance, Novo et al. (Novo *et al.*, 2009) showed that there were several regions in a newly sequenced *S. cerevisiae* strain that were not part of the reference yeast genome. One of the three unique regions held a large degree of similarity to a region in *Zygosaccharomyces bailii*, which is a major yeast contaminant in wine and therefore can be found in close contact with this particular strain of S. cerevisiae, propitiating the exchange of DNA. *Z. bailii* tolerates common food preservatives, high concentrations of sugar and ethanol, and low pH and this may be characteristics that have conferred *S. cerevisiae* with a selective advantage during wine fermentation. A second example was provided by Ma et al. (Ma *et al.*, 2010). They showed that chromosomes can be interchanged between some Fusarium species. These events were related with the emergence of pathogenesis in previously non-pathogenic strains. This two examples show that transferences of genetic material are part of the evolution of fungi and it may be interesting to follow it through with a more in-depth analysis on how these events have shaped the evolution of fungal species.

Alternatively, fungal species also have been the donors in some transfers towards other Eukaryotes. Unlike most animals, the Pea Aphid (*Acyrthosiphon* *pisum*), whose genome was published recently and in which our group performed some phylogenomic analysis (Huerta-Cepas *et al.*, 2010b; Consortium, 2010), has been shown to have the machinery needed to be able to synthesize carotenes. Recently, Moran and Jarvik (Moran and Jarvik, 2010) suggested that some of the genes that are part of this pathway were of fungal origin and that they had been acquired in one single transfer event.

Several recent studies have shown that gene transfers in fungi, whether they act as donors or acceptors, is a reality and therefore we should take this into account. Our minimal estimates have shown that horizontal gene transfer events between prokaryotic and fungal species occur and while their rates are not close to the ones observed in prokaryotes, care should still be taken when applying methodologies that are sensible to HGT events. For instance, the concatenation of genes during species tree reconstruction has been shown to be sensitive to such events (Wolf *et al.*, 2001) and genes that may have undergone HGT should be detected and excluded from the reconstruction.

Gene duplication: shaping the oxidative phosphorilation pathway in fungi

In chapter 6 we studied the presence/absence patterns of the proteins that conform the oxidative phosphorilation (OXPHOS) pathway in fungal species. We expanded the work done before by Gabaldon et al. (Gabaldón, 2005) on the first complex of the OXPHOS pathway. Our results regarding this complex were congruent to the ones found before, though we were able to show that the ACPM protein, which had been dubbed as a complex I remnant in yeast species, is in fact a paralog. Indeed, Candida species still conserved both copies of this protein and the phylogeny shows that one of the copies is clustered with the yeast ACPM protein while the other is grouped with the Pezizomycotina proteins, which have been shown to be part of Complex I.

This is not the only case of duplications found in the five complexes that form the OXPHOS pathway. Indeed, we found that 76% of the protein families conforming this pathway have had at least one duplication in their evolutionary history. According to the gene dosage hypothesis, duplications in protein complexes are more likely to be deleterious due to dosage effects unless they affect the whole complex (Papp *et al.*, 2003). 90% of the duplications within the OXPHOS pathway can be traced back to the two existing points of whole genome duplication that occurred in fungal evolution (at the ancestor of the Saccharomyces group (Kellis *et al.*, 2004) and at the ancestor of the zygomycota *Rhizopus oryzae* (Ma *et al.*, 2009)). Both whole genome duplication events propitiated the complete duplication of the OXPHOS pathway. The retention rate of these duplicated protein complexes, however, does not seem different to that of other proteins in Saccharomyces. A 93% of the OXPHOS duplicates that occurred in this species during the whole genome duplication has already been lost. A similar value was found concerning the whole genome (88%) (Kellis *et al.*, 2004). On the other hand, in *Rhizopus oryzae* a 55% of gene loss after genome duplication was reported (Ma *et al.*, 2009). The loss rate in OXPHOS was slightly lower, with a 45% (20% reported in the paper describing the genome of *R. oryzae* (Ma *et al.*, 2009)). This would imply that in the first moment both duplicates of the complex are maintained, but that with time the organism is able to adapt and loose the subunits that are not longer needed for the new functionality that the complex adopts. Other duplicates have now been adapted towards other multi-protein complexes, indicating that duplicated subunits in the OXPHOS pathway can be adapted into a new function of another complex and thus bypass the gene balance hypothesis.

From the computer, to the lab, and back

While most of the work performed during this thesis has been mainly computational, we have participated in some collaborations with experimental groups. From the work done within the FunPath consortium, to other smaller projects with a specific aim in mind, we have worked to integrate bioinformatics in several experimental studies in the understanding that the union of both methodologies can often surpass the results that can be produced independently.

Within the FunPath consortium, we have used bioinformatics tools in order to select the putative virulence candidates in the *C. glabrata* deletion project, so that the deletions were focused towards genes of interest for the Consortium. Several predictions of evolutionary or functionally interesting genes were also performed as explained in chapter 7. But bioinformatics does not only perform guiding or supporting jobs. Indeed, often during large scale computational studies we have come across genes that seem interesting enough to request additional experimental validation. This was the case of the horizontally transferred aspartate racemase identified during the the work presented in chapter 6. The fact that several racemases have already been described as horizontal gene transfers in other eukaryotic species (Uo *et al.*, 2001; Fitzpatrick *et al.*, 2008; Gladyshev *et al.*, 2008) coupled with the fact that some racemases have been described as potential virulence factors (Goytia *et al.*, 2007) prompted us to seek a collaboration with an experimental lab who could validate some of our hypothesis. Experiments to characterize this protein are currently being undertaken in the labs of Bernhard Hube and Matthias Brock of the Hans-Knöll-Institute.

The importance of bioinformatics has grown during the last decade, specially in light of the large amount of experimental data produced with current technologies. Even so, bioinformatics provides a specific angle, and as such integrating it with other disciplines is one of the best ways to move forward in science.

Concluding remarks

The availability of an increasing amount of completely sequenced genomes coupled with the advances in the computational field have allowed for increasingly extensive large scale studies. Fungi is the eukaryotic group with a largest amount of completely sequenced species and therefore it is particularly well suited for comparative genomics analyzes. A species tree, rather than being a scientific curiosity, provides an evolutionary framework in which to interpret observations from various species. It provides ways to infer the directionality of changes and elucidate how current structures or phenotypes evolved. Species trees should represent the evolution of a group of species rather than that of a reduced set of genes. Concern about the reliability of current tree reconstruction methods led us to delve deeper into the congruence between gene trees and species trees and resulted in the design of several methods by which we could identify points in the species tree that were poorly supported by a whole phylome. We determined that the species tree was mostly well supported but some nodes showed large discrepancies to most genes. In this sense the bootstrap support was not a good predictor of genomic support.

These results could be attributed to the large topological diversity found within gene trees. Topological differences between a gene tree and a species tree can be caused by methodological artifacts or by evolutionary processes that result in changes in the topology of the gene tree. Our analyzes have shown that, contrary to previous assumptions, HGT may play an important role in fungal evolution. Gene duplications followed by differential loss have affected numerous fungal species and are often the cause of incongruences between species trees and gene trees. As described here, the OXPHOS pathway has been affected by this process at similar levels than the rest of the genome, despite being mainly formed by multi-protein complexes.

Future challenges will include finding new ways to exploit massive amounts of data to create more robust species trees and analyze large collections of phylogenetic trees. Whether we will one day contemplate a fully-resolved and robust Fungal Tree of Life remains an open question. There may be events that the pass of time have erased forever and thus we should be able to recognize what parts of the tree may not be reliable. What is certain is that data to come will reveal new surprises on the evolution of this highly diverse and versatile kingdom.

CONCLUSIONS

1. Phylomes can be used to assess the robustness of a species tree. We have designed different distance measures that have enabled us to detect nodes within the species trees that are not well represented among the topologies found in the complete collection of gene phylogenies of a genome.

2. We have designed treeKO, which is a tree comparison program that enables the user to compare any pair of trees. Unlike most existing programs it allows the comparison of trees with duplications in a biological comprehensive way.

3. Orthology detection based on phylogenies is able to elucidate the complex relationships between genes that have undergone multiple rounds of duplication. In this sense we have shown that the species overlap algorithm offers sensitivity as compared to tree reconciliation.

4. The number of horizontal gene transfer events between prokaryotes and fungi is more extensive than previously thought. We have detected 235 horizontal gene transfer events in which more than 700 genes were involved, in a survey of 60 completely sequenced genomes.

5. Additionally, not all fungal groups are affected equally by those events. Pezizomycotina in particular show a larger percentage of gene uptake from prokaryotes than other fungal groups. Aspergillus and Fusarium species show a higher tendency to retain transferred DNA.

6. Duplication and loss events have shaped the evolution of the oxidative phosphorylation pathway. 76% of the genes that comprise this pathway have experienced at least one duplication. This is primarily the result of whole genome duplications that occurred during the evolution of some fungal species.

7. Contrary to predictions of gene balance hypothesis, protein families that are part of an OXPHOS complex are affected by duplication to a similar extent. Functional divergence and recruitment to other complexes may explain this finding.

8. Single nucleotide polymorphisms can be used to reconstruct phylogenetic trees. The results obtained with this large amount of data are congruent with previous phylogenies based on multi locus sequence typing.

9. Comparison of closely related species and the analysis of a pathogens phylome can be exploited to select candidate genes likely to be involved in virulence.

Appendices

APPENDIX A
LIST OF PUBLICATIONS

1. **Marcet-Houben M** and Gabaldón T. 2010. *Acquisition of prokaryotic genes by fungal genomes*. Trends Genet. 26(1):5-8
2. **Marcet-Houben M** and Gabaldón T. 2009. *The tree versus the forest: the fungal tree of life and the topological diversity within the yeast phylome.* PLoS One. 4(2):e4357.
3. **Marcet-Houben M**, Marceddu G. and Gabaldón T. 2009. *Phylogenomics of the oxidative phosphorylation in fungi reveals extensive gene duplication followed by functional divergence*. BMC Evol Biol. Dec 21;9:295
4. **Marcet-Houben M** and Gabaldón T. *treeKO a duplication aware algorithm for the comparison of phylogenetic trees.* (submitted)
5. **Marcet-Houben M**, Puigbò P, Romeu A and Garcia-Vallve, S. 2007. *Towards reconstructing a metabolic tree of life.* Bioinformation, 2(4) 135-144
6. **Marcet-Houben M**, Cabré M, Paternáin JL, Romeu A. 2008. *Phylogenetic analysis of homologous fatty acid synthase and polyketide synthase involved in aflatoxin biosynthesis*. Bioinformation, 3(1):33-40.
7. International Aphid Genomics Consortium including **Marcet-Houben, M.**, *Genome sequence of the pea aphid Acyrthosiphon pisum.* PLoS Biol. 2010;8(2):e1000313.
8. Huerta-Cepas J, **Marcet-Houben M**, Pignatelli M, Moya A and Gabaldón T. 2010 *The pea aphid phylome: a complete catalogue of evolutionary histories and arthropod orthology and paralogy relationships for Acyrthosiphon pisum genes*, Insect Molecular Biology, 19:13-21.
9. Gabaldón T, **Marcet-Houben M**, and Huerta-Cepas J. 2008, *Reconstruction and analysis of large-scale phylogenetic data, challenges and opportunities.* Nova Sciences Publishers.
10. Schwarzmüller T, Gabaldón T, **Marcet-Houben M**, Glaser, W and Kuchler, K. *Molecular Basis and Genes Implicated in Pathogenicity and Drug Resistance of the Human Fungal Pathogen Candida glabrata*. (Submitted).
11. Roetzer A, Klopfa E, Gratzb N, **Marcet-Houben M**, Hillerd E, Rupp S, Gabaldón T, Kovarik P and Schüller C. *CgYap1/Skn7 and CgSod1 provide additive protection against oxidative stress responses in Candida glabrata.* (Submitted)

BIBLIOGRAPHY

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*, pages 267–281.

Akiba, T., Koyama, K., Ishiki, Y., Kimura, S., and Fukushima, T. (1960). On the mechanism of the development of multiple-drug-resistant clones of shigella. *Jpn J Microbiol*, **4**, 219–227.

Alberich, R., Cardona, G., Rossello, F., and Valiente, G. (2008). An algebraic metric for phylogenetic trees. *Applied Mathematics Letters*, **22**(9), 1320–1324.

Altekar, G., Dwarkadas, S., Huelsenbeck, J. P., and Ronquist, F. (2004). Parallel metropolis coupled markov chain monte carlo for bayesian phylogenetic inference. *Bioinformatics*, **20**(3), 407–415.

Altenhoff, A. M. and Dessimoz, C. (2009). Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol*, **5**(1), e1000262.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, **25**(17), 3389–3402.

Andersson, J. O. (2005). Lateral gene transfer in eukaryotes. *Cell Mol Life Sci*, **62**(11), 1182–1197.

Ané, C., Larget, B., Baum, D. A., Smith, S. D., and Rokas, A. (2007). Bayesian estimation of concordance among gene trees. *Mol Biol Evol*, **24**(2), 412–426.

Anisimova, M. and Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol*, **55**(4), 539–552.

Bakkeren, G., Kämper, J., and Schirawski, J. (2008). Sex in smut fungi: Structure, function and evolution of mating-type complexes. *Fungal Genet Biol*, **45 Suppl 1**, S15–S21.

Beiko, R. G. and Hamilton, N. (2006). Phylogenetic identification of lateral genetic transfer events. *BMC Evol Biol*, **6**, 15.

Berglund-Sonnhammer, A.-C., Steffansson, P., Betts, M. J., and Liberles, D. A. (2006). Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. *J Mol Evol*, **63**(2), 240–250.

Bille, P. (2005). A survey on tree edit distance and related problems. *Theor Comput Sci*, **337**, 217–239.

Bininda-Emonds, O. R. P. (2004). The evolution of supertrees. *Trends Ecol Evol*, **19**(6), 315–322.

Bininda-Emonds, O. R. P. (2005). Supertree construction in the genomic age. *Methods Enzymol*, **395**, 745–757.

Bougnoux, M.-E., Aanensen, D. M., Morand, S., Théraud, M., Spratt, B. G., and d'Enfert, C. (2004). Multilocus sequence typing of candida albicans: strategies, data exchange and applications. *Infect Genet Evol*, **4**(3), 243–252.

Brody, S., Oh, C., Hoja, U., and Schweizer, E. (1997). Mitochondrial acyl carrier protein is involved in lipoic acid synthesis in saccharomyces cerevisiae. *FEBS Lett*, **408**(2), 217–220.

Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E., and Stanhope, M. J. (2001). Universal trees based on large combined protein sequence data sets. *Nat Genet*, **28**(3), 281–285.

Bruns, T. D., Vilgalys, R., Barns, S. M., Gonzalez, D., Hibbett, D. S., Lane, D. J., Simon, L., Stickel, S., Szaro, T. M., and Weisburg, W. G. (1992). Evolutionary relationships within the fungi: analyses of nuclear small subunit rrna sequences. *Mol Phylogenet Evol*, **1**(3), 231–241.

Bull, J., Huelsenbeck, J. P., Cunningham, C. W., Swofford, D. L., and Waddell, P. J. (1993). Partitioning and combining data in phylogenetic analysis. *Systematic Biology, 42 (3):384-397*.

Bullerwell, C. E. and Lang, B. F. (2005). Fungal evolution: the case of the vanishing mitochondrion. *Curr Opin Microbiol*, **8**(4), 362–369.

Byrne, K. P. and Wolfe, K. H. (2006). Visualizing syntenic relationships among the hemiascomycetes with the yeast gene order browser. *Nucleic Acids Res*, **34**(Database issue), D452–D455.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**(15), 1972–1973.

Casaregola, S., Neuvéglise, C., Lépingle, A., Bon, E., Feynerol, C., Artiguenave, F., Wincker, P., and Gaillardin, C. (2000). Genomic exploration of the hemiascomycetous yeasts: 17. yarrowia lipolytica. *FEBS Lett*, **487**(1), 95–100.

Castresana, J. (2007). Topological variation in single-gene phylogenetic trees. *Genome Biol*, **8**(6), 216.

Castrillo, J. I. and Oliver, S. G. (2004). Yeast as a touchstone in post-genomic research: strategies for integrative analysis in functional genomics. *J Biochem Mol Biol*, **37**(1), 93–106.

Christie, K. R., Weng, S., Balakrishnan, R., Costanzo, M. C., Dolinski, K., Dwight, S. S., Engel, S. R., Feierbach, B., Fisk, D. G., Hirschman, J. E., Hong, E. L., Issel-Tarver, L., Nash, R., Sethuraman, A., Starr, B., Theesfeld, C. L., Andrada, R., Binkley, G., Dong, Q., Lane, C., Schroeder, M., Botstein, D., and Cherry, J. M. (2004). Saccharomyces genome database (sgd) provides tools to identify and analyze sequences from saccharomyces cerevisiae and related sequences from other organisms. *Nucleic Acids Res*, **32**, D311–D314.

Consortium, T. I. A. G. (2010). Genome sequence of the pea aphid acyrthosiphon pisum. *PLoS Biol*, **8**(2), e1000313.

Cornell, M. J., Alam, I., Soanes, D. M., Wong, H. M., Hedeler, C., Paton, N. W., Rattray, M., Hubbard, S. J., Talbot, N. J., and Oliver, S. G. (2007). Comparative genome analysis across a kingdom of eukaryotic organisms: specialization and diversification in the fungi. *Genome Res*, **17**(12), 1809–1822.

Cornman, R. S., Chen, Y. P., Schatz, M. C., Street, C., Zhao, Y., Desany, B., Egholm, M., Hutchison, S., Pettis, J. S., Lipkin, W. I., and Evans, J. D. (2009). Genomic analyses of the microsporidian nosema ceranae, an emergent pathogen of honey bees. *PLoS Pathog*, **5**(6), e1000466.

Corradi, N., Haag, K. L., Pombert, J.-F., Ebert, D., and Keeling, P. J. (2009). Draft genome sequence of the daphnia pathogen octosporea bayeri: insights into the gene content of a large microsporidian genome and a model for host-parasite interactions. *Genome Biol*, **10**(10), R106.

Cuéllar-Cruz, M., Briones-Martin-del Campo, M., Cañas-Villamar, I., Montalvo-Arredondo, J., Riego-Ruiz, L., Castaño, I., and De Las Peñas, A. (2008). High resistance to oxidative stress in the fungal pathogen candida glabrata is mediated by a single catalase, cta1p, and is controlled by the transcription factors yap1p, skn7p, msn2p, and msn4p. *Eukaryot Cell*, **7**(5), 814–825.

Dagan, T. and Martin, W. (2006). The tree of one percent. *Genome Biol*, **7**(10), 118.

Darwin, C. (1859). *The Origin of Species by Means of Natural Selection or The Preservation of Favoured Races in the Struggle for Life.* John Murray, Albemarle Street., first edition. edition.

De Grassi, A., Lanave, C., and Saccone, C. (2008). Genome duplication and gene-family evolution: the case of three oxphos gene families. *Gene*, **421**, 1–6.

de Vienne, D. M., Giraud, T., and Martin, O. C. (2007). A congruence index for testing topological similarity between trees. *Bioinformatics*, **23**(23), 3119–3124.

Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet*, **6**(5), 361–375.

Denning, D. W., Anderson, M. J., Turner, G., Latgé, J. P., and Bennett, J. W. (2002). Sequencing the aspergillus fumigatus genome. *Lancet Infect Dis*, **2**(4), 251–253.

Dufayard, J.-F., Duret, L., Penel, S., Gouy, M., Rechenmann, F., and Perrière, G. (2005). Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, **21**(11), 2596–2603.

Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuvéglise, C., Talla, E., Goffard, N., Frangeul, L., Aigle, M., Anthouard, V., Babour, A., Barbe, V., Barnay, S., Blanchin, S., Beckerich, J.-M., Beyne, E., Bleykasten, C., Boisramé, A., Boyer, J., Cattolico, L., Confanioleri, F., De Daruvar, A., Despons, L., Fabre, E., Fairhead, C., Ferry-Dumazet, H., Groppi, A., Hantraye, F., Hennequin, C., Jauniaux, N., Joyet, P., Kachouri, R., Kerrest, A., Koszul, R., Lemaire, M., Lesur, I., Ma, L., Muller, H., Nicaud, J.-M., Nikolski, M., Oztas, S., Ozier-Kalogeropoulos, O., Pellenz, S., Potier, S., Richard, G.-F., Straub, M.-L., Suleau, A., Swennen, D., Tekaia, F., Wésolowski-Louvel, M., Westhof, E., Wirth, B., Zeniou-Meyer, M., Zivanovic, I., Bolotin-Fukuhara, M., Thierry, A., Bouchier, C., Caudron, B., Scarpelli, C., Gaillardin, C., Weissenbach, J., Wincker, P., and Souciet, J.-L. (2004). Genome evolution in yeasts. *Nature*, **430**(6995), 35–44.

Dutilh, B. E., van Noort, V., van der Heijden, R. T. J. M., Boekhout, T., Snel, B., and Huynen, M. A. (2007). Assessment of phylogenomic and orthology approaches for phylogenetic inference. *Bioinformatics*, **23**(7), 815–824.

Edgar, R. C. (2004). Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

Edlind, T. D., Li, J., Visvesvara, G. S., Vodkin, M. H., McLaughlin, G. L., and Katiyar, S. K. (1996). Phylogenetic analysis of beta-tubulin sequences from amitochondrial protozoa. *Mol Phylogenet Evol*, **5**(2), 359–367.

Edwards, S. V., Liu, L., and Pearl, D. K. (2007). High-resolution species trees without concatenation. *Proc Natl Acad Sci U S A*, **104**(14), 5936–5941.

Eisen, J. A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res*, **8**(3), 163–167.

Eisen, J. A. and Fraser, C. M. (2003). Phylogenomics: intersection of evolution and genomics. *Science*, **300**(5626), 1706–1707.

Emanuelsson, O., Brunak, S., von Heijne, G., and Nielsen, H. (2007). Locating proteins in the cell using targetp, signalp and related tools. *Nat Protoc*, **2**(4), 953–971.

Esser, C., Ahmadinejad, N., Wiegand, C., Rotte, C., Sebastiani, F., Gelius-Dietrich, G., Henze, K., Kretschmann, E., Richly, E., Leister, D., Bryant, D., Steel, M. A., Lockhart, P. J., Penny, D., and Martin, W. (2004). A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol Biol Evol*, **21**(9), 1643–1660.

Estabrook, G. F., McMorris, F. R., and Meacham, C. A. (1985). Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Systematic Zoology 34(2):193-200*.

Faguy, D. M. and Doolittle, W. F. (2000). Horizontal transfer of catalase-peroxidase genes between archaea and pathogenic bacteria. *Trends Genet*, **16**(5), 196–197.

Fast, N. M., Law, J. S., Williams, B. A. P., and Keeling, P. J. (2003). Bacterial catalase in the microsporidian nosema locustae: implications for microsporidian metabolism and genome evolution. *Eukaryot Cell*, **2**(5), 1069–1075.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, **39**(4), 783–791.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst Zool*, **19**(2), 99–113.

Fitzpatrick, D. A., Logue, M. E., Stajich, J. E., and Butler, G. (2006). A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evol Biol*, **6**, 99.

Fitzpatrick, D. A., Logue, M. E., and Butler, G. (2008). Evidence of recent interkingdom horizontal gene transfer between bacteria and candida parapsilosis. *BMC Evol Biol*, **8**, 181.

Fowler, T. J., Mitton, M. F., Vaillancourt, L. J., and Raper, C. A. (2001). Changes in mate recognition through alterations of pheromones and receptors in the multisexual mushroom fungus schizophyllum commune. *Genetics*, **158**(4), 1491–1503.

Fraser, J. A. and Heitman, J. (2003). Fungal mating-type loci. *Curr Biol*, **13**(20), R792–R795.

Gabaldón, T. (2005). Evolution of proteins and proteomes: a phylogenetics approach. *Evol Bioinform Online*, **1**, 51–61.

Gabaldón, T. (2008). Large-scale assignment of orthology: back to phylogenetics? *Genome Biol*, **9**(10), 235.

Gabaldón, T. and Huynen, M. A. (2003). Reconstruction of the proto-mitochondrial metabolism. *Science*, **301**(5633), 609.

Gabaldón, T. and Huynen, M. A. (2004). Shaping the mitochondrial proteome. *Biochim Biophys Acta*, **1659**, 212–220.

Gabaldón, T. and Huynen, M. A. (2005). Lineage-specific gene loss following mitochondrial endosymbiosis and its potential for function prediction in eukaryotes. *Bioinformatics*, **21 Suppl 2**, ii144–ii150.

Gabaldón, T., Marcet-Houben, M., and Huerta-Cepas, J. (2008). *Reconstruction and analysis of large-scale phylogenetic data, challenges and opportunities*. Computational Biology: New Research. Nova Sciences Publishers, NY.

Gadagkar, S. R., Rosenberg, M. S., and Kumar, S. (2005). Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *J Exp Zool B Mol Dev Evol*, **304**(1), 64–74.

Gakh, O., Cavadini, P., and Isaya, G. (2002). Mitochondrial processing peptidases. *Biochim Biophys Acta*, **1592**(1), 63–77.

Galagan, J. E., Henn, M. R., Ma, L.-J., Cuomo, C. A., and Birren, B. (2005). Genomics of the fungal kingdom: insights into eukaryotic biology. *Genome Res*, **15**(12), 1620–1631.

Garcia-Vallvé, S., Romeu, A., and Palau, J. (2000). Horizontal gene transfer of glycosyl hydrolases of the rumen fungi. *Mol Biol Evol*, **17**(3), 352–361.

Gascuel, O. (1997). Bionj: an improved version of the nj algorithm based on a simple model of sequence data. *Mol Biol Evol*, **14**(7), 685–695.

Gladyshev, E. A., Meselson, M., and Arkhipova, I. R. (2008). Massive horizontal gene transfer in bdelloid rotifers. *Science*, **320**(5880), 1210–1213.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S. G. (1996). Life with 6000 genes. *Science*, **274**(5287), 546, 563–546, 567.

Gojković, Z., Knecht, W., Zameitat, E., Warneboldt, J., Coutelis, J.-B., Pynyaha, Y., Neuveglise, C., Møller, K., Löffler, M., and Piskur, J. (2004). Horizontal gene transfer promoted evolution of the ability to propagate under anaerobic conditions in yeasts. *Mol Genet Genomics*, **271**(4), 387–393.

Gordon, J. L., Byrne, K. P., and Wolfe, K. H. (2009). Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern saccharomyces cerevisiae genome. *PLoS Genet*, **5**(5), e1000485.

Gouret, P., Thompson, J. D., and Pontarotti, P. (2009). Phylopattern: regular expressions to identify complex patterns in phylogenetic trees. *BMC Bioinformatics*, **10**, 298.

Goytia, M., Chamond, N., Cosson, A., Coatnoan, N., Hermant, D., Berneman, A., and Minoprio, P. (2007). Molecular and structural discrimination of proline racemase and hydroxyproline-2-epimerase from nosocomial and bacterial pathogens. *PLoS One*, **2**(9), e885.

Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, **52**(5), 696–704.

Guindon, S., Delsuc, F., Dufayard, J.-F., and Gascuel, O. (2009). Estimating maximum likelihood phylogenies with phyml. *Methods Mol Biol*, **537**, 113–137.

Hall, C. and Dietrich, F. S. (2007). The reacquisition of biotin prototrophy in saccharomyces cerevisiae involved horizontal gene transfer, gene duplication and gene clustering. *Genetics*, **177**(4), 2293–2307.

Hall, C., Brachat, S., and Dietrich, F. S. (2005). Contribution of horizontal gene transfer to the evolution of saccharomyces cerevisiae. *Eukaryot Cell*, **4**(6), 1102–1115.

Hane, J. K., Lowe, R. G. T., Solomon, P. S., Tan, K.-C., Schoch, C. L., Spatafora, J. W., Crous, P. W., Kodira, C., Birren, B. W., Galagan, J. E., Torriani, S. F. F., McDonald, B. A., and Oliver, R. P. (2007). Dothideomycete plant interactions illuminated by genome sequencing and est analysis of the wheat pathogen stagonospora nodorum. *Plant Cell*, **19**(11), 3347–3368.

Hawksworth, D. (1991). The fungal dimension of biodiversity: magnitude, significance, and conservation. *Mycological Research, 95(6):641-655*.

Hegedus, D. D. and Rimmer, S. R. (2005). Sclerotinia sclerotiorum: when "to be or not to be" a pathogen? *FEMS Microbiol Lett*, **251**(2), 177–184.

Hodge, M. R., Kim, G., Singh, K., and Cumsky, M. G. (1989). Inverse regulation of the yeast cox5 genes by oxygen and heme. *Mol Cell Biol*, **9**(5), 1958–1964.

Horn, D. L., Neofytos, D., Anaissie, E. J., Fishman, J. A., Steinbach, W. J., Olyaei, A. J., Marr, K. A., Pfaller, M. A., Chang, C.-H., and Webster, K. M. (2009). Epidemiology and outcomes of candidemia in 2019 patients: data from the prospective antifungal therapy alliance registry. *Clin Infect Dis*, **48**(12), 1695–1703.

Hu, P., Janga, S. C., Babu, M., Díaz-Mejía, J. J., Butland, G., Yang, W., Pogoutse, O., Guo, X., Phanse, S., Wong, P., Chandran, S., Christopoulos, C., Nazarians-Armavil, A., Nasseri, N. K., Musso, G., Ali, M., Nazemof, N., Eroukova, V., Golshani, A., Paccanaro, A., Greenblatt, J. F., Moreno-Hagelsieb, G., and Emili, A. (2009). Global functional atlas of escherichia coli encompassing previously uncharacterized proteins. *PLoS Biol*, **7**(4), e96.

Hubbard, T. J. P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Herrero, J., Holland, R., Howe, K., Johnson, N., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Melsopp, C., Megy, K., Meidl, P., Ouverdin, B., Parker, A., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Severin, J., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wood, M., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Flicek, P., Kasprzyk, A., Proctor, G., Searle, S., Smith, J., Ureta-Vidal, A., and Birney, E. (2007). Ensembl 2007. *Nucleic Acids Res*, **35**(Database issue), D610–D617.

Huerta-Cepas, J., Dopazo, H., Dopazo, J., and Gabaldón, T. (2007). The human phylome. *Genome Biol*, **8**(6), R109.

Huerta-Cepas, J., Bueno, A., Dopazo, J., and Gabaldón, T. (2008). Phylomedb: a database for genome-wide collections of gene phylogenies. *Nucleic Acids Res*, **36**(Database issue), D491–D496.

Huerta-Cepas, J., Dopazo, J., and Gabaldón, T. (2010a). Ete: a python environment for tree exploration. *BMC Bioinformatics*, **11**, 24.

Huerta-Cepas, J., Marcet-Houben, M., Pignatelli, M., Moya, A., and Gabaldon, T. (2010b). The pea aphid phylome: a complete catalogue of evolutionary histories and arthropod orthology and paralogy relationships for acyrthosiphon pisum genes. *INSECT MOLECULAR BIOLOGY*, **19**, 13–21.

Idnurm, A., Rodríguez-Romero, J., Corrochano, L. M., Sanz, C., Iturriaga, E. A., Eslava, A. P., and Heitman, J. (2006). The phycomyces mada gene encodes a blue-light photoreceptor for phototropism and other light responses. *Proc Natl Acad Sci U S A*, **103**(12), 4546–4551.

Idnurm, A., Walton, F. J., Floyd, A., and Heitman, J. (2008). Identification of the sex genes in an early diverged fungus. *Nature*, **451**(7175), 193–196.

Isambert, H. and Stein, R. R. (2009). On the need for widespread horizontal gene transfers under genome size constraint. *Biol Direct*, **4**, 28.

Jacobsen, I. D., Brunke, S., Seider, K., Schwarzmüller, T., Firon, A., d'Enfért, C., Kuchler, K., and Hube, B. (2010). Candida glabrata persistence in mice does not depend on host immunosuppression and is unaffected by fungal amino acid auxotrophy. *Infect Immun*, **78**(3), 1066–1077.

James, T. Y., Kauff, F., Schoch, C. L., Matheny, P. B., Hofstetter, V., Cox, C. J., Celio, G., Gueidan, C., Fraker, E., Miadlikowska, J., Lumbsch, H. T., Rauhut, A., Reeb, V., Arnold, A. E., Amtoft, A., Stajich, J. E., Hosaka, K., Sung, G.-H., Johnson, D., O'Rourke, B., Crockett, M., Binder, M., Curtis, J. M., Slot, J. C., Wang, Z., Wilson, A. W., Schüssler, A., Longcore, J. E., O'Donnell, K., Mozley-Standridge, S., Porter, D., Letcher, P. M., Powell, M. J., Taylor, J. W., White, M. M., Griffith, G. W., Davies, D. R., Humber, R. A., Morton, J. B., Sugiyama, J., Rossman, A. Y., Rogers, J. D., Pfister, D. H., Hewitt, D., Hansen, K., Hambleton, S., Shoemaker, R. A., Kohlmeyer, J., Volkmann-Kohlmeyer, B., Spotts, R. A., Serdani, M., Crous, P. W., Hughes, K. W., Matsuura, K., Langer, E., Langer, G., Untereiner, W. A., Lücking, R., Büdel, B., Geiser, D. M., Aptroot, A., Diederich, P., Schmitt, I., Schultz, M., Yahr, R., Hibbett, D. S., Lutzoni, F., McLaughlin, D. J., Spatafora, J. W., and Vilgalys, R. (2006). Reconstructing the early evolution of fungi using a six-gene phylogeny. *Nature*, **443**(7113), 818–822.

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., and von Mering, C. (2009). String 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*, **37**(Database issue), D412–D416.

Johnson, C. H., Prigge, J. T., Warren, A. D., and McEwen, J. E. (2003). Characterization of an alternative oxidase activity of histoplasma capsulatum. *Yeast*, **20**(5), 381–388.

Kamada, T. (2002). Molecular genetics of sexual development in the mushroom coprinus cinereus. *Bioessays*, **24**(5), 449–459.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2008). Kegg for linking genomes to life and the environment. *Nucleic Acids Res*, **36**(Database issue), D480–D484.

Katinka, M. D., Duprat, S., Cornillot, E., Méténier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretaillade, E., Brottier, P., Wincker, P., Delbac, F., El Alaoui, H., Peyret, P., Saurin, W., Gouy, M., Weissenbach, J., and Vivarès, C. P. (2001). Genome sequence and gene compaction of the eukaryote parasite encephalitozoon cuniculi. *Nature*, **414**(6862), 450–453.

Keeling, P. J. and Doolittle, W. F. (1996). Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family. *Mol Biol Evol*, **13**(10), 1297–1305.

Keeling, P. J. and Palmer, J. D. (2008). Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet*, **9**(8), 605–618.

Keeling, P. J., Luker, M. A., and Palmer, J. D. (2000). Evidence from beta-tubulin phylogeny that microsporidia evolved from within the fungi. *Mol Biol Evol*, **17**(1), 23–31.

Keeling, P. J., Burger, G., Durnford, D. G., Lang, B. F., Lee, R. W., Pearlman, R. E., Roger, A. J., and Gray, M. W. (2005). The tree of eukaryotes. *Trends Ecol Evol*, **20**(12), 670–676.

Kellis, M., Birren, B. W., and Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, **428**(6983), 617–624.

Khaldi, N., Collemare, J., Lebrun, M.-H., and Wolfe, K. H. (2008). Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi. *Genome Biol*, **9**(1), R18.

Klotz, M. G. and Loewen, P. C. (2003). The molecular evolution of catalatic hydroperoxidases: evidence for multiple lateral transfer of genes between prokaryota and from bacteria into eukaryota. *Mol Biol Evol*, **20**(7), 1098–1112.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet*, **39**, 309–338.

Kües, U. (2000). Life history and developmental processes in the basidiomycete coprinus cinereus. *Microbiol Mol Biol Rev*, **64**(2), 316–353.

Kuramae, E. E., Robert, V., Snel, B., and Boekhout, T. (2006a). Conflicting phylogenetic position of schizosaccharomyces pombe. *Genomics*, **88**(4), 387–393.

Kuramae, E. E., Robert, V., Snel, B., Weiss, M., and Boekhout, T. (2006b). Phylogenomics reveal a robust fungal tree of life. *FEMS Yeast Res*, **6**(8), 1213–1220.

Kuramae, E. E., Robert, V., Echavarri-Erasun, C., and Boekhout, T. (2007). Cophenetic correlation analysis as a strategy to select phylogenetically informative proteins: an example from the fungal kingdom. *BMC Evol Biol*, **7**, 134.

Kurland, C. G. (2005). What tangled web: barriers to rampant horizontal gene transfer. *Bioessays*, **27**(7), 741–747.

Lang, B. F., O'Kelly, C., Nerad, T., Gray, M. W., and Burger, G. (2002). The closest unicellular relatives of animals. *Curr Biol*, **12**(20), 1773–1778.

Lavín, J. L., Oguiza, J. A., Ramírez, L., and Pisabarro, A. G. (2008). Comparative genomics of the oxidative phosphorylation system in fungi. *Fungal Genet Biol*, **45**(9), 1248–1256.

Lee, S. C., Weiss, L. M., and Heitman, J. (2009). Generation of genetic diversity in microsporidia via sexual reproduction and horizontal gene transfer. *Commun Integr Biol*, **2**(5), 414–417.

Liu, X., Brutlag, D. L., and Liu, J. S. (2001). Bioprospector: discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*, pages 127–138.

Liu, Y., Leigh, J. W., Brinkmann, H., Cushion, M. T., Rodriguez-Ezpeleta, N., Philippe, H., and Lang, B. F. (2009). Phylogenomic analyses support the monophyly of taphrinomycotina, including schizosaccharomyces fission yeasts. *Mol Biol Evol*, **26**(1), 27–34.

Lutzoni, F., Kauff, F., Cox, C. J., McLaughlin, D., Celio, G., Dentinger, B., Padamsee, M., Hibbett, D., James, T. Y., Baloch, E., Grube, M., Reeb, V., Hofstetter, V., Schoch, C., Arnold, A. E., Miadlikowska, J., Spatafora, J., Johnson, D., Hambleton, S., Crockett, M., Shoemaker, R., Sung, G.-H., Lucking, R., Lumbsch, T., O'Donnell, K., Binder, M., Diederich, P., Ertz, D., Gueidan, C., Hansen, K., Harris, R. C., Hosaka, K., Lim, Y.-W., Matheny, B., Nishida, H., Pfister, D., Rogers, J., Rossman, A., Schmitt, I., Sipman, H., Stone, J., Sugiyama, J., Yahr, R., and Vilgalys, R. (2004). Assembling the fungal tree of life: progress, classification, and evolution of subcellular traits. *Am. J. Bot.*, **91**(10), 1446–1480.

Lynch, M. and Katju, V. (2004). The altered evolutionary trajectories of gene duplicates. *Trends Genet*, **20**(11), 544–549.

Ma, L.-J., Ibrahim, A. S., Skory, C., Grabherr, M. G., Burger, G., Butler, M., Elias, M., Idnurm, A., Lang, B. F., Sone, T., Abe, A., Calvo, S. E., Corrochano, L. M., Engels, R., Fu, J., Hansberg, W., Kim, J.-M., Kodira, C. D., Koehrsen, M. J., Liu, B., Miranda-Saavedra, D., O'Leary, S., Ortiz-Castellanos, L., Poulter, R., Rodriguez-Romero, J., Ruiz-Herrera, J., Shen, Y.-Q., Zeng, Q., Galagan, J., Birren, B. W., Cuomo, C. A., and Wickes, B. L. (2009). Genomic analysis of the basal lineage fungus rhizopus oryzae reveals a whole-genome duplication. *PLoS Genet*, **5**(7), e1000549.

Ma, L.-J., van der Does, H. C., Borkovich, K. A., Coleman, J. J., Daboussi, M.-J., Di Pietro, A., Dufresne, M., Freitag, M., Grabherr, M., Henrissat, B., Houterman, P. M., Kang, S., Shim, W.-B., Woloshuk, C., Xie, X., Xu, J.-R., Antoniw, J., Baker, S. E., Bluhm, B. H., Breakspear, A., Brown, D. W., Butchko, R. A. E., Chapman, S., Coulson, R., Coutinho, P. M., Danchin, E. G. J., Diener, A., Gale, L. R., Gardiner, D. M., Goff, S., Hammond-Kosack, K. E., Hilburn, K., Hua-Van, A., Jonkers, W., Kazan, K., Kodira, C. D., Koehrsen, M., Kumar, L., Lee, Y.-H., Li, L., Manners, J. M., Miranda-Saavedra, D., Mukherjee, M., Park, G., Park, J., Park, S.-Y., Proctor,

R. H., Regev, A., Ruiz-Roldan, M. C., Sain, D., Sakthikumar, S., Sykes, S., Schwartz, D. C., Turgeon, B. G., Wapinski, I., Yoder, O., Young, S., Zeng, Q., Zhou, S., Galagan, J., Cuomo, C. A., Kistler, H. C., and Rep, M. (2010). Comparative genomics reveals mobile pathogenicity chromosomes in fusarium. *Nature*, **464**(7287), 367–373.

Magill, S. S., Shields, C., Sears, C. L., Choti, M., and Merz, W. G. (2006). Triazole cross-resistance among candida spp.: case report, occurrence among bloodstream isolates, and implications for antifungal therapy. *J Clin Microbiol*, **44**(2), 529–535.

Magnani, T., Soriani, F. M., Martins, V. P., Nascimento, A. M., Tudella, V. G., Curti, C., and Uyemura, S. A. (2007). Cloning and functional expression of the mitochondrial alternative oxidase of aspergillus fumigatus and its induction by oxidative stress. *FEMS Microbiol Lett*, **271**(2), 230–238.

Marcet-Houben, M. and Gabaldón, T. (2009). The tree versus the forest: the fungal tree of life and the topological diversity within the yeast phylome. *PLoS One*, **4**(2), e4357.

Méan, M., Marchetti, O., and Calandra, T. (2008). Bench-to-bedside review: Candida infections in the intensive care unit. *Crit Care*, **12**(1), 204.

Merten, D., Kothe, E., and BÃŒchel, G. (2004). Studies on microbial heavy metal retention from uranium mine drainage water with special emphasis on rare earth elements. *Mine Water and the Environment 23(1):34-43*.

Moran, N. A. and Jarvik, T. (2010). Lateral transfer of genes from fungi underlies carotenoid production in aphids. *Science*, **328**(5978), 624–627.

Mueller, G. M. and Schmit, J. P. (2007). Fungal biodiversity: what do we know? what can we predict? *Biodivers Conserv 16:1-5*.

Münsterkötter, M. and Steinberg, G. (2007). The fungus ustilago maydis and humans share disease-related proteins that are not found in saccharomyces cerevisiae. *BMC Genomics*, **8**, 473.

Nace, H. L., Horn, D., and Neofytos, D. (2009). Epidemiology and outcome of multiple-species candidemia at a tertiary care center between 2004 and 2007. *Diagn Microbiol Infect Dis*, **64**(3), 289–294.

Nishida, H. and Sugiyama, J. (1993). Phylogenetic relationships among taphrina, saitoella, and other higher fungi. *Mol Biol Evol*, **10**(2), 431–436.

Novo, M., Bigey, F., Beyne, E., Galeote, V., Gavory, F., Mallet, S., Cambon, B., Legras, J.-L., Wincker, P., Casaregola, S., and Dequin, S. (2009). Eukaryote-to-eukaryote gene transfer events revealed by the genome sequence of the wine yeast saccharomyces cerevisiae ec1118. *Proc Natl Acad Sci U S A*, **106**(38), 16333–16338.

Odds, F. C., Bougnoux, M.-E., Shaw, D. J., Bain, J. M., Davidson, A. D., Diogo, D., Jacobsen, M. D., Lecomte, M., Li, S.-Y., Tavanti, A., Maiden, M. C. J., Gow, N. A. R., and d'Enfert, C. (2007). Molecular phylogenetics of candida albicans. *Eukaryot Cell*, **6**(6), 1041–1052.

Ogilvie, I., Kennaway, N. G., and Shoubridge, E. A. (2005). A molecular chaperone for mitochondrial complex i assembly is mutated in a progressive encephalopathy. *J Clin Invest*, **115**(10), 2784–2792.

O'Gorman, C. M., Fuller, H. T., and Dyer, P. S. (2009). Discovery of a sexual cycle in the opportunistic fungal pathogen aspergillus fumigatus. *Nature*, **457**(7228), 471–474.

Ohno, S. (1970). *Evolution by gene duplication*.

Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., Goto, S., and Kanehisa, M. (2008). Kegg atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res*, **36**, W423–W426.

Ostlund, G., Schmitt, T., Forslund, K., Köstler, T., Messina, D. N., Roopra, S., Frings, O., and Sonnhammer, E. L. L. (2010). Inparanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res*, **38**(Database issue), D196–D203.

Page, R. D. and Charleston, M. A. (1997). From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol Phylogenet Evol*, **7**(2), 231–240.

Panackal, A. A., Gribskov, J. L., Staab, J. F., Kirby, K. A., Rinaldi, M., and Marr, K. A. (2006). Clinical significance of azole antifungal drug cross-resistance in candida glabrata. *J Clin Microbiol*, **44**(5), 1740–1743.

Papp, B., Pál, C., and Hurst, L. D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature*, **424**(6945), 194–197.

Peña-Castillo, L. and Hughes, T. R. (2007). Why are there still over 1000 uncharacterized yeast genes? *Genetics*, **176**(1), 7–14.

Pfaller, M. A. and Diekema, D. J. (2007). Epidemiology of invasive candidiasis: a persistent public health problem. *Clin Microbiol Rev*, **20**(1), 133–163.

Pfaller, M. A., Boyken, L., Hollis, R. J., Kroeger, J., Messer, S. A., Tendolkar, S., and Diekema, D. J. (2008). In vitro susceptibility of invasive isolates of candida spp. to anidulafungin, caspofungin, and micafungin: six years of global surveillance. *J Clin Microbiol*, **46**(1), 150–156.

Pisani, D., Cotton, J. A., and McInerney, J. O. (2007). Supertrees disentangle the chimerical origin of eukaryotic genomes. *Mol Biol Evol*, **24**(8), 1752–1760.

Pollegioni, L., Piubelli, L., Sacchi, S., Pilone, M. S., and Molla, G. (2007). Physiological functions of d-amino acid oxidases: from yeast to humans. *Cell Mol Life Sci*, **64**(11), 1373–1394.

Pounds, J. A., Bustamante, M. R., Coloma, L. A., Consuegra, J. A., Fogden, M. P. L., Foster, P. N., La Marca, E., Masters, K. L., Merino-Viteri, A., Puschendorf, R., Ron, S. R., Sánchez-Azofeifa, G. A., Still, C. J., and Young, B. E. (2006). Widespread amphibian extinctions from epidemic disease driven by global warming. *Nature*, **439**(7073), 161–167.

Puigbò, P., Garcia-Vallvé, S., and McInerney, J. O. (2007). Topd/fmts: a new software to compare phylogenetic trees. *Bioinformatics*, **23**(12), 1556–1558.

Rasmussen, M. D. and Kellis, M. (2007). Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. *Genome Res*, **17**(12), 1932–1942.

Redecker, D. and Raab, P. (2006). Phylogeny of the glomeromycota (arbuscular mycorrhizal fungi): recent developments and new gene markers. *Mycologia*, **98**(6), 885–895.

Retief, J. D. (2000). Phylogenetic analysis using phylip. *Methods Mol Biol*, **132**, 243–258.

Ricard, G., McEwan, N. R., Dutilh, B. E., Jouany, J.-P., Macheboeuf, D., Mitsumori, M., McIntosh, F. M., Michalowski, T., Nagamine, T., Nelson, N., Newbold, C. J., Nsabimana, E., Takenaka, A., Thomas, N. A., Ushida, K., Hackstein, J. H. P., and Huynen, M. A. (2006). Horizontal gene transfer from bacteria to rumen ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment. *BMC Genomics*, **7**, 22.

Robbertse, B., Reeves, J. B., Schoch, C. L., and Spatafora, J. W. (2006). A phylogenomic analysis of the ascomycota. *Fungal Genet Biol*, **43**(10), 715–725.

Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, **53**, 131–147.

Roden, M. M., Zaoutis, T. E., Buchanan, W. L., Knudsen, T. A., Sarkisova, T. A., Schaufele, R. L., Sein, M., Sein, T., Chiou, C. C., Chu, J. H., Kontoyiannis, D. P., and Walsh, T. J. (2005). Epidemiology and outcome of zygomycosis: a review of 929 reported cases. *Clin Infect Dis*, **41**(5), 634–653.

Roetzer, A., Gregori, C., Jennings, A. M., Quintin, J., Ferrandon, D., Butler, G., Kuchler, K., Ammerer, G., and Schüller, C. (2008). Candida glabrata environmental stress response involves saccharomyces cerevisiae msn2/4 orthologous transcription factors. *Mol Microbiol*, **69**(3), 603–620.

Rosen, B. P. (1999). Families of arsenic transporters. *Trends Microbiol*, **7**(5), 207–212.

Ruan, J., Li, H., Chen, Z., Coghlan, A., Coin, L. J. M., Guo, Y., Hériché, J.-K., Hu, Y., Kristiansen, K., Li, R., Liu, T., Moses, A., Qin, J., Vang, S., Vilella, A. J., Ureta-Vidal, A., Bolund, L., Wang, J., and Durbin, R. (2008). Treefam: 2008 update. *Nucleic Acids Res*, **36**(Database issue), D735–D740.

Ruderfer, D. M., Pratt, S. C., Seidel, H. S., and Kruglyak, L. (2006). Population genomic analysis of outcrossing and recombination in yeast. *Nat Genet*, **38**(9), 1077–1081.

Runswick, M. J., Fearnley, I. M., Skehel, J. M., and Walker, J. E. (1991). Presence of an acyl carrier protein in nadh:ubiquinone oxidoreductase from bovine heart mitochondria. *FEBS Lett*, **286**, 121–124.

Saccone, C., Lanave, C., and De Grassi, A. (2006). Metazoan oxphos gene families: evolutionary forces at the level of mitochondrial and nuclear genomes. *Biochim Biophys Acta*, **1757**, 1171–1178.

Saraste, M. (1999). Oxidative phosphorylation at the fin de siècle. *Science*, **283**(5407), 1488–1493.

Schell, M. J., Cooper, O. B., and Snyder, S. H. (1997). D-aspartate localizations imply neuronal and neuroendocrine roles. *Proc Natl Acad Sci U S A*, **94**(5), 2013–2018.

Schmidt, H. A., Strimmer, K., Vingron, M., and von Haeseler, A. (2002). Tree-puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, **18**(3), 502–504.

Schmit, J. and Mueller, G. (2007). An estimate of the lower limit of global fungal diversity. *Biodiversity and Conservation*, **16**(1), 99–111.

Schneider, R., Massow, M., Lisowsky, T., and Weiss, H. (1995). Different respiratory-defective phenotypes of neurospora crassa and saccharomyces cerevisiae after inactivation of the gene encoding the mitochondrial acyl carrier protein. *Curr Genet*, **29**(1), 10–17.

Schwarzmuller, T., Gabaldon, T., Marcet-Houben, M., Glaser, W., and Kuchler, K. (2010). Molecular basis and genes implicated in pathogenicity and drug resistance of the human fungal pathogen candida glabrata. *Submitted*.

Shin, J. H., Chae, M. J., Song, J. W., Jung, S.-I., Cho, D., Kee, S. J., Kim, S. H., Shin, M. G., Suh, S. P., and Ryang, D. W. (2007). Changes in karyotype and azole susceptibility of sequential bloodstream isolates from patients with candida glabrata candidemia. *J Clin Microbiol*, **45**(8), 2385–2391.

Sicheritz-Pontén, T. and Andersson, S. G. (2001). A phylogenomic approach to microbial evolution. *Nucleic Acids Res*, **29**(2), 545–552.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol*, **147**(1), 195–197.

Snel, B., Bork, P., and Huynen, M. A. (2002). Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res*, **12**(1), 17–25.

Snel, B., Huynen, M. A., and Dutilh, B. E. (2005). Genome trees and the nature of genome evolution. *Annu Rev Microbiol*, **59**, 191–209.

Soria-Carrasco, V., Talavera, G., Igea, J., and Castresana, J. (2007). The k tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. *Bioinformatics*, **23**(21), 2954–2956.

Spatafora, J. W., Sung, G.-H., Johnson, D., Hesse, C., O'Rourke, B., Serdani, M., Spotts, R., Lutzoni, F., Hofstetter, V., Miadlikowska, J., Reeb, V., Gueidan, C., Fraker, E., Lumbsch, T., Lücking, R., Schmitt, I., Hosaka, K., Aptroot, A., Roux, C., Miller, A. N., Geiser, D. M., Hafellner, J., Hestmark, G., Arnold, A. E., Büdel, B., Rauhut, A., Hewitt, D., Untereiner, W. A., Cole, M. S., Scheidegger, C., Schultz, M., Sipman, H., and Schoch, C. L. (2006). A five-gene phylogeny of pezizomycotina. *Mycologia*, **98**(6), 1018–1028.

Srikantha, T., Daniels, K. J., Wu, W., Lockhart, S. R., Yi, S., Sahni, N., Ma, N., and Soll, D. R. (2008). Dark brown is the more virulent of the switch phenotypes of candida glabrata. *Microbiology*, **154**(Pt 11), 3309–3318.

Steenkamp, E. T., Wright, J., and Baldauf, S. L. (2006). The protistan origins of animals and fungi. *Mol Biol Evol*, **23**(1), 93–106.

Steinberg, G. and Perez-Martin, J. (2008). Ustilago maydis, a new fungal model system for cell biology. *Trends Cell Biol*, **18**(2), 61–67.

Stoye, J., Evers, D., and Meyer, F. (1998). Rose: generating sequence families. *Bioinformatics*, **14**(2), 157–163.

Syvanen, M. (1985). Cross-species gene transfer; implications for a new theory of evolution. *J Theor Biol*, **112**(2), 333–343.

Talavera, G. and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol*, **56**(4), 564–577.

Tanabe, Y., Watanabe, M. M., and Sugiyama, J. (2005). Evolutionary relationships among basal fungi (chytridiomycota and zygomycota): Insights from molecular phylogenetics. *J Gen Appl Microbiol*, **51**(5), 267–276.

Tavanti, A., Davidson, A. D., Fordyce, M. J., Gow, N. A. R., Maiden, M. C. J., and Odds, F. C. (2005). Population structure and properties of candida albicans, as determined by multilocus sequence typing. *J Clin Microbiol*, **43**(11), 5601–5613.

Taylor, J. W. and Berbee, M. L. (2006). Dating divergences in the fungal tree of life: review and new analyses. *Mycologia*, **98**(6), 838–849.

Teixeira, M. C., Monteiro, P., Jain, P., Tenreiro, S., Fernandes, A. R., Mira, N. P., Alenquer, M., Freitas, A. T., Oliveira, A. L., and Sá-Correia, I. (2006). The yeastract database: a tool for the analysis of transcription regulatory associations in saccharomyces cerevisiae. *Nucleic Acids Res*, **34**(Database issue), D446–D451.

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouzé, P., and Moreau, Y. (2001). A higher-order background model improves the detection of promoter regulatory elements by gibbs sampling. *Bioinformatics*, **17**(12), 1113–1122.

Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouzé, P., and Moreau, Y. (2002). A gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J Comput Biol*, **9**(2), 447–464.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**(22), 4673–4680.

Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., Régnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, **23**(1), 137–144.

Uo, T., Yoshimura, T., Tanaka, N., Takegawa, K., and Esaki, N. (2001). Functional characterization of alanine racemase from schizosaccharomyces pombe: a eucaryotic counterpart to bacterial alanine racemase. *J Bacteriol*, **183**(7), 2226–2233.

Valiente, G. (2005). A fast algorithmic technique for comparing large phylogenetic trees. In *String Processing and Information Retrieval*, volume 3772, chapter String Processing and Information Retrieval, pages 370–375. Springer.

van der Heijden, R. T. J. M., Snel, B., van Noort, V., and Huynen, M. A. (2007). Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics*, **8**, 83.

van Helden, J., André, B., and Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol*, **281**(5), 827–842.

Veiga, A., Arrabaça, J. D., and Loureiro-Dias, M. C. (2003). Cyanide-resistant respiration, a very frequent metabolic pathway in yeasts. *FEMS Yeast Res*, **3**(3), 239–245.

Wainright, P. O., Hinkle, G., Sogin, M. L., and Stickel, S. K. (1993). Monophyletic origins of the metazoa: an evolutionary link with fungi. *Science*, **260**(5106), 340–342.

Wang, H., Xu, Z., Gao, L., and Hao, B. (2009). A fungal phylogeny based on 82 complete genomes using the composition vector method. *BMC Evol Biol*, **9**, 195.

Wang, Z., Binder, M., Schoch, C. L., Johnston, P. R., Spatafora, J. W., and Hibbett, D. S. (2006a). Evolution of helotialean fungi (leotiomycetes, pezizomycotina): a nuclear rdna phylogeny. *Mol Phylogenet Evol*, **41**(2), 295–312.

Wang, Z., Johnston, P. R., Takamatsu, S., Spatafora, J. W., and Hibbett, D. S. (2006b). Toward a phylogenetic classification of the leotiomycetes based on rdna data. *Mycologia*, **98**(6), 1065–1075.

Wenzel, R. P. and Gennings, C. (2005). Bloodstream infections due to candida species in the intensive care unit: identifying especially high-risk patients to determine prevention strategies. *Clin Infect Dis*, **41 Suppl 6**, S389–S393.

Wolf, Y. I., Rogozin, I. B., Grishin, N. V., Tatusov, R. L., and Koonin, E. V. (2001). Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol Biol*, **1**, 8.

Wolfe, K. H. and Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, **387**(6634), 708–713.

Wong, K. M., Suchard, M. A., and Huelsenbeck, J. P. (2008). Alignment uncertainty and genomic analysis. *Science*, **319**(5862), 473–476.

Wong, S., Fares, M. A., Zimmermann, W., Butler, G., and Wolfe, K. H. (2003). Evidence from comparative genomics for a complete sexual cycle in the 'asexual' pathogenic yeast candida glabrata. *Genome Biol*, **4**(2), R10.

Wu, Y. (2009). A practical method for exact computation of subtree prune and regraft distance. *Bioinformatics*, **25**(2), 190–196.

Zaragoza, O., Chrisman, C. J., Castelli, M. V., Frases, S., Cuenca-Estrella, M., Rodríguez-Tudela, J. L., and Casadevall, A. (2008). Capsule enlargement in cryptococcus neoformans confers resistance to oxidative stress suggesting a mechanism for intracellular survival. *Cell Microbiol*, **10**(10), 2043–2057.

Zmasek, C. M. and Eddy, S. R. (2001). A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics*, **17**(9), 821–828.

Zuckerkandl, E. and Pauling, L. (1965). Molecules as documents of evolutionary history. *J Theor Biol*, **8**(2), 357–366.